Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1504

Does Relative Performance Feedback Improve Academic Outcomes? Evidence from a Randomised Controlled Trial in a Spanish University

Cristian Macías
Universidad Rey Juan Carlos

Rosa Santero
London School of Economics

Ismael Sanz
Universidad Rey Juan Carlos

J. D. Tena
Università degli Studi di Sassari

The effects of relative performance feedback on student achievement remain contested in the economics of education literature. Some studies find that providing students with information about their standing relative to peers improves academic outcomes, while others show negative effects driven by reduced effort among students who learn they perform better than expected. This paper provides new experimental evidence on the effectiveness of relative performance feedback in higher education. We conduct a pre-registered randomised controlled trial involving 386 undergraduate students across four degree programmes at Universidad Rey Juan Carlos in Madrid, Spain. Following a midterm exam, students in the treatment group receive information about their percentile ranking within the class, while those in the control group receive only their absolute scores. We estimate intent-to-treat effects on final exam performance and find that the treatment increases final exam scores by 0.41 points on a 0–10 scale when controlling for baseline performance and student characteristics, an improvement equivalent to 17% of a standard deviation. The intervention is particularly effective for students with lower baseline performance, female students, and those who do not receive additional tutoring. Exploiting a measure of the gap between students' self-reported expected grade and their midterm performance, we further show that the treatment is concentrated among students who arrive at the intervention holding over-optimistic beliefs about their academic standing. Classifying students into pessimistic, accurate, and over-optimistic categories, over-optimistic students gain +0.67 points ($p < 0.01$), accurate students show an effect that is

VERSION: June 2026

Suggested citation: Macías Domínguez, Cristian, Rosa Santero Sánchez, Ismael Sanz Labrador, and J.D. Tena. (2026). Does Relative Performance Feedback Improve Academic Outcomes? Evidence from a Randomised Controlled Trial in a Spanish University. (EdWorkingPaper: 26-1504). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/za71-7x46

Does Relative Performance Feedback Improve Academic Outcomes?

Evidence from a Randomised Controlled Trial in a Spanish University

Cristian Macías[1], Rosa Santero[1], Ismael Sanz[1,2] and J. D. Tena[3,4]

[1] *Universidad Rey Juan Carlos, Department of Applied Economics, Madrid, Spain*
[2] *Department of Social Policy, London School of Economics, United Kingdom*
[3] *Department of Economics, University of Liverpool, Liverpool, United Kingdom*
[4] *Department of Economics, Università degli Studi di Sassari, Sassari, Italy*

June 2026

Abstract

The effects of relative performance feedback on student achievement remain contested in the economics of education literature. Some studies find that providing students with information about their standing relative to peers improves academic outcomes, while others show negative effects driven by reduced effort among students who learn they perform better than expected. This paper provides new experimental evidence on the effectiveness of relative performance feedback in higher education. We conduct a pre-registered randomised controlled trial involving 386 undergraduate students across four degree programmes at Universidad Rey Juan Carlos in Madrid, Spain. Following a midterm exam, students in the treatment group receive information about their percentile ranking within the class, while those in the control group receive only their absolute scores. We estimate intent-to-treat effects on final exam performance and find that the treatment increases final exam scores by 0.41 points on a 0–10 scale when controlling for baseline performance and student characteristics, an improvement equivalent to 17% of a standard deviation. The intervention is particularly effective for students with lower baseline performance, female students, and those who do not receive additional tutoring. Exploiting a measure of the gap between students' self-reported expected grade and their midterm performance, we further show that the treatment is concentrated among students who arrive at the intervention holding over-optimistic beliefs about their academic standing. Classifying students into pessimistic, accurate, and over-optimistic categories, over-optimistic students gain +0.67 points ($p < 0.01$), accurate students show an effect that is statistically indistinguishable from zero (−0.14), and pessimistic students display a negative point estimate (−0.43). These results show that relative performance feedback improves academic outcomes when delivered to students whose prior beliefs are misaligned with their realised performance, and that it can be implemented at low cost across higher education institutions.

Keywords: relative performance feedback; randomised controlled trial; higher education; student expectations; reference-dependent preferences.

JEL codes: I21, I23, D83, C93.

1. Introduction

The provision of relative performance feedback to students generates mixed results in the economics of education literature. Some studies find positive effects on academic achievement (Azmat & Iriberri, 2010; Megalokonomou & Zhang, 2024), while others document negative impacts driven by complacency among students who receive better-than-expected news about their standing (Azmat et al., 2019). This heterogeneity reflects fundamental differences in how students respond to information about their position relative to peers. Reference-dependent models of effort, in which students evaluate outcomes against an internal expectation rather than in absolute terms, suggest that the direction of the effect depends on students' prior beliefs and the scope for behavioural adjustment (Kőszegi & Rabin, 2006). Understanding when and for whom relative performance feedback improves outcomes is therefore essential for designing effective educational interventions. This paper addresses the following research question: does providing university students with personalised information about their relative academic performance causally affect subsequent academic achievement, and how does this effect vary with students' prior beliefs and baseline characteristics?

The effectiveness of relative performance feedback can be understood through behavioural models of reference-dependent preferences and social comparison. Students form expectations about their academic standing based on limited information, and these beliefs influence their effort allocation (Kőszegi & Rabin, 2006). When students receive information about their relative position within a class, this updates their reference point and can trigger behavioural responses. Students performing below their expected rank may increase effort to close the gap, while those performing above expectations may experience either increased motivation or reduced effort due to complacency (Bursztyn & Jensen, 2015). Azmat et al. (2019) show that in a Spanish university context the majority of students underestimate their relative position; when these students receive feedback revealing that they perform better than expected, the information constitutes "good news" that can reduce effort and harm performance. Aucejo and Wong (2025) extend this logic by showing, in a U.S. university setting, that the effectiveness of personalised feedback is concentrated among students who arrive at the intervention holding overly optimistic beliefs about their own performance, beliefs that feedback then helps to recalibrate. This asymmetry in responses to information helps explain why average treatment effects in the literature are often small or negative, while heterogeneous effects by baseline performance and prior beliefs can be substantial.

Building on this literature, the objectives of this paper are threefold: (i) to estimate the causal effect of relative performance feedback on final exam outcomes in higher education; (ii) to examine heterogeneity in treatment effects by baseline academic performance, gender, and access to alternative academic support; and (iii) to test directly whether treatment effects are mediated by the misalignment between students' prior beliefs and their realised academic performance.

This study is conducted at Universidad Rey Juan Carlos (URJC), one of the public universities in the Madrid region. The intervention targets undergraduate students enrolled in applied economics and business-related degrees, including Business Administration, Accounting and Finance, Marketing, and Architecture. The Spanish higher education system shares several features with those examined in previous experiments, such as a strong emphasis on final examinations. Courses at URJC follow a two-exam structure, with a midterm assessment followed by a final examination. The midterm is primarily practical and typically takes place around mid-semester. The final examination, scheduled during the official examination period, includes both theoretical and practical components. Final grades are determined by a weighted average of 40% from the midterm and 60% from the final examination. Comparative feedback is rare in this setting, which makes it well suited to identifying the causal impact of relative performance information delivered as a single informational shock between assessments.

Our analysis relies on data from a pre-registered randomised controlled trial (AEA RCT Registry ID: AEARCTR-0014016) conducted between January 2022 and December 2024. The analytical sample comprises 386 undergraduate students across four degree programmes who were randomly assigned to treatment (191) or control (195) conditions. The primary outcome variable is the final exam score, measured on a 0–10 scale. Among these students, the average midterm score is 4.91 and the average final exam score is 5.10. The sample shows balanced representation by gender, with an average age of 21 years. Approximately 33% of students hold a scholarship, and 28% report working while studying. Covariate information is collected through baseline surveys at the beginning of the course, capturing demographic characteristics, academic background, expectations regarding performance, and access to resources such as internet, personal computers, and quiet study spaces.

We adopt an intent-to-treat (ITT) framework to estimate the causal effect of relative performance feedback on academic outcomes. The ITT approach measures the effect of being offered percentile feedback, regardless of whether the student actually read or attended to the message, and is therefore the policy-relevant parameter for an intervention that an institution can roll out at scale. Students are randomly assigned to treatment and control groups, with stratification by course and academic year to ensure balance. The treatment group receives the absolute midterm score alongside a personalised message indicating the student's percentile rank within the course cohort. The control group receives only the absolute midterm grade. Feedback is delivered electronically through the university's internal messaging system, accompanied by a short explanation of what the percentile rank represents. Our main specification regresses final exam scores on a treatment indicator, controlling for midterm performance, demographic characteristics, and programme fixed effects, with heteroskedasticity-robust standard errors.

The estimated treatment effect is positive and statistically significant. In the baseline specification controlling only for midterm performance and programme fixed effects, the treatment coefficient is 0.17 points and not significant. When the full set of controls is included, the treatment effect increases to a significant 0.41 points on a 0–10 scale, an improvement equivalent to

approximately 17% of a standard deviation in final exam scores. Across both specifications, midterm performance is a strong predictor of final exam outcomes, with a coefficient of approximately 0.5, confirming the importance of controlling for baseline academic ability.

The heterogeneity analysis reveals that the impact of informational feedback is not uniform across student characteristics. The treatment is more effective for students with lower baseline academic performance, with the estimated effect for students in the lowest academic percentile being positive and statistically significant. Female students exhibit a positive and statistically significant treatment effect, while the effect for male students is smaller and not significant. Students who do not attend extra classes experience a stronger and statistically significant treatment effect, suggesting that informational feedback is more valuable when alternative sources of academic support are absent.

Beyond these dimensions, we show that the effectiveness of the intervention depends on the gap between students' prior beliefs and their realised performance. Drawing on the framework of Aucejo and Wong (2025), we construct a measure of the "expectation shock" as the difference between each student's self-reported expected grade, elicited in the baseline survey before the experiment, and the midterm grade observed immediately before treatment delivery. We find that 79% of students in our sample are over-optimistic under the binary definition, in the sense that their expected grade exceeds their midterm performance, and that 55% are strongly over-optimistic (by more than two grade points on the 0–10 scale). When we interact the treatment with this measure, the effect of the intervention is concentrated among over-optimistic students: the average marginal effect of the treatment is a significant 0.63 grade points for students who overestimate their own performance, compared with −0.37 for those who do not. Students who learn that they are doing worse than they had expected respond by raising their effort, whereas students whose expectations were already aligned with their performance have little reason to adjust. The interaction coefficient is large (0.99) and statistically significant at the 5% level.

This pattern fits the reference-dependent preferences framework and the evidence in Aucejo and Wong (2025) that feedback helps most when students hold inaccurate expectations about their own performance.

The internal validity of the experiment rests on successful randomisation. Balance tests confirm that baseline covariates are statistically indistinguishable between treatment and control groups for all continuous variables. For example, the average midterm score is 4.88 in the control group and 4.94 in the treatment group ($p = 0.83$). The expectation shock itself is balanced across treatment arms (mean 2.55 in control vs. 2.34 in treatment, $p = 0.46$), which is expected given that both expectations and midterm grades are measured before treatment delivery. We conduct robustness checks excluding outliers in midterm or final exam scores and re-running regressions on balanced panels with complete outcome and covariate data, and we use an alternative mapping of the expected-grade scale (which yields a marginally significant interaction of 0.75, $p < 0.10$), obtaining consistent results throughout.

This study contributes to the literature by providing experimental evidence that helps reconcile apparently contradictory findings on relative performance feedback. Azmat and Iriberri (2010) find positive effects in a Spanish secondary school when students receive information about their standing relative to the class average. Megalokonomou and Zhang (2024) show that achievement rank has positive effects on subsequent performance in Chinese middle schools. In contrast, Azmat et al. (2019) implement a field experiment in a Spanish university and find that feedback has no average effect on performance, with the negative effect driven by students who underestimate their position and reduce effort upon learning that they perform better than expected. Our results complement this evidence in two ways. First, the positive average effect we estimate suggests that a single-shot feedback delivered between two assessments within the same course allows students to adjust effort before the final examination, rather than inducing complacency over repeated semesters. Second, by documenting that the effect is driven by over-optimistic students, we provide direct empirical support for the prior-beliefs mechanism highlighted by Aucejo and Wong (2025) and predicted by reference-dependent models, in which behaviour is governed by the gap between an outcome and a prior expectation that serves as a reference point (Kőszegi & Rabin, 2006). In our setting, over-optimism is the mirror image of the under-estimation documented by Azmat et al. (2019), and feedback recalibrates beliefs in the productive direction.

We also document heterogeneity in treatment effects that helps target the intervention. Like several previous studies, we examine heterogeneous effects of feedback by baseline performance and gender; our richer baseline survey additionally lets us explore how the treatment interacts with socioeconomic background, participation in extra classes, and a direct measure of the belief–reality gap. We are not the first to study mis-calibrated beliefs, since the literature on over- and under-estimation of academic standing is well developed (Azmat et al., 2019; Aucejo & Wong, 2025), but combining a pre-treatment measure of the expectation shock with a clean single-shock design lets us test the prior-beliefs mechanism directly. By showing that feedback works best for students whose prior expectations are misaligned with their actual performance, our findings suggest how feedback interventions can be targeted at low cost to the students who stand to gain most.

The remainder of the paper is structured as follows. Section 2 reviews the related literature. Section 3 describes the context, data, and experimental design. Section 4 outlines the empirical strategy. Section 5 presents the main results and heterogeneity analysis, including the new expectation-shock evidence. Section 6 concludes.

2. Related Literature

A growing literature in economics and education examines how students respond to information about their relative academic performance. These studies show that relative performance feedback can affect students' effort, expectations, and achievement, although the direction and magnitude of the effects depend on context, prior beliefs, and the design of the intervention. We review this evidence, identify the gap our study addresses, and explain how our design allows us to contribute to understanding the mechanism behind these effects.

In educational settings, several studies show positive effects of relative performance feedback. Azmat and Iriberri (2010) exploit a natural experiment in a Spanish secondary school and find that providing students with information about their standing relative to the class average leads to improved grades. In follow-up work, Azmat and Iriberri (2016) use a laboratory experiment to show that feedback enhances performance under piece-rate incentives but not under flat pay, and that this effect occurs regardless of whether participants learn that they are above or below average; the mere provision of relative information drives effort when it has monetary consequences. This indicates that the incentive structure moderates the effectiveness of informational interventions. Megalokonomou and Zhang (2024) use randomised classroom assignment in Chinese middle schools to estimate the effects of salient ordinal ranking and find that a one-standard-deviation increase in relative position improves subsequent academic achievement by 0.06 standard deviations. These effects are partly mediated by parental expectations and student confidence, which together account for approximately 48% of the total effect, and rank effects are stronger for male students. Dobrescu et al. (2021) implement real-time relative performance feedback in a university setting and show positive effects of 0.21 standard deviations on average grades, operating through increased peer interactions and social learning.

The literature also identifies contexts in which relative performance feedback has negative or null effects. Azmat et al. (2019) implement a field experiment in a Spanish university in which students receive periodic updates on their academic rank over multiple semesters. While the intervention has no average effect on performance, and if anything a negative one, it improves student satisfaction, especially among those who had underestimated their standing. The authors show that the majority of students underestimate their relative position: when these students receive feedback revealing that they perform better than expected, the information constitutes "good news" that reduces effort and harms performance. This highlights the crucial role of prior beliefs in determining how students respond to feedback. In line with the evidence that average effects may be null in higher education, preliminary pilot studies conducted in the same institutional context as the present study also found no significant average effects of relative performance feedback on academic outcomes (Macías, 2023; Macías & Santero, 2025). Collins and Lundstedt (2024) study the effects of more informative grading in Swedish secondary schools, exploiting a reform that increased the number of passing grades from three to five. They report negative effects on high-school graduation rates of about 0.13 standard deviations (3.3 percentage points) and larger negative effects on STEM-track completion, with discouragement as the likely mechanism. Together, these mixed findings suggest that the response to relative performance information is not straightforward and depends on how students incorporate feedback into their beliefs and effort decisions.

Recent work has advanced our understanding of the mechanisms through which feedback affects performance and the student characteristics that moderate these effects. Aucejo and Wong (2025) study personalised feedback that combines relative performance information with encouragement at Arizona State University. They find that first-generation students in

synchronous classes benefit significantly, with an effect of 0.19 standard deviations on course grades, while the effect is absent in asynchronous classes and among continuing-generation students. They also show that first-generation students are more likely to be overconfident about their academic performance, with expected grades that exceed actual grades by 0.14 letter points more than those of continuing-generation students, and that overconfident first-generation students in synchronous classes are the ones who primarily benefit from feedback. Feedback thus appears to help most when students hold inaccurate expectations about their own performance. Our analysis builds directly on this framework: we construct an analogous measure of the gap between expected and realised grades in our Spanish sample and test whether the treatment effect of relative performance feedback concentrates among over-optimistic students.

A related strand documents heterogeneity by baseline performance and shows that feedback design can mitigate discouragement. Chen et al. (2024) design a league-based feedback system that dynamically groups students with similar scores to reduce demoralisation among low performers. They find that low-performing students improve by 0.27 standard deviations while high-performing students decline by 0.25 standard deviations, suggesting that the framing of relative information matters as much as its content. This pattern of larger benefits for lower-performing students aligns with the finding in Azmat et al. (2019) that students who overestimate their position respond positively to corrective feedback, while those who underestimate reduce effort.

Beyond rank-based interventions, Zhou et al. (2025) study teacher-to-student feedback in rural Chinese primary schools. They find that frequent personalised feedback using scorecards of schoolwork and behaviour improves mathematics scores by 0.16 to 0.20 standard deviations and language scores by 0.09 standard deviations, with effects particularly strong for Grade 3 students. Communicating these assessments to parents produces additional benefits of approximately 0.30 standard deviations for left-behind children whose parents work in distant cities. The effectiveness of feedback thus depends not only on student characteristics but also on the involvement of parents and complementary support structures. Gender differences have also received attention. Megalokonomou and Zhang (2024) find stronger effects for male students in their Chinese sample, while evidence in other contexts suggests that female students may respond more positively to affirming feedback, particularly in subjects where gender stereotypes are prevalent, indicating that the moderating role of gender varies across settings.

At the university level, Brade et al. (2026) study ongoing feedback on accumulated course credits in a German university using two natural field experiments with six years of data. They find that feedback increases graduation within one year of the scheduled duration by 3.7 percentage points (an 8% increase), accelerates graduation by 0.15 semesters, and raises grades by 0.063 standard deviations. Students with medium ex-ante graduation probabilities are most responsive, and above-average feedback improves outcomes for middle performers while below-average feedback worsens them.

Despite this extensive evidence, the literature offers limited experimental evidence on relative performance feedback delivered as a single informational shock between two assessments within the same course. Existing university-level experiments either provide feedback repeatedly over multiple semesters (Azmat et al., 2019; Brade et al., 2026) or update rank information continuously through classroom interactions (Dobrescu et al., 2021). We do not claim that one piece of information is inherently superior to several; rather, the repeated nature of feedback in these designs makes it difficult to isolate how students respond to a discrete intervention delivered at a strategic moment, when they can still adjust effort before a final assessment. By contrast, cleaner before-after designs are typically conducted in primary or secondary schools (Azmat & Iriberri, 2010; Megalokonomou & Zhang, 2024; Zhou et al., 2025), where institutional features, student populations, and assessment stakes differ substantially from higher education.

While recent work documents heterogeneous effects by baseline performance and socioeconomic background, less is known about whether the effectiveness of feedback depends on the alignment between students' prior beliefs and their actual performance. This dimension matters because reference-dependent models predict that the response to feedback should depend on the gap between the reference point (expected performance) and the realised outcome. Several features of our setting allow us to address it. The two-exam structure of courses at URJC creates a natural before-after design that isolates the effect of a single informational intervention: the midterm provides a baseline measure of performance, feedback is delivered shortly after midterm results are released, and the final exam, administered several weeks later, measures outcomes, avoiding the confounding present in designs with continuous or repeated feedback. Our baseline survey lets us examine heterogeneous effects across several dimensions at once, including a self-reported expected grade elicited before the intervention; combining this measure with the midterm score yields a direct proxy for the expectation shock that feedback is expected to recalibrate. Finally, the Spanish university context allows direct comparison with Azmat et al. (2019), who study a similar institutional environment with a different feedback design, and with Aucejo and Wong (2025), who explicitly study the role of prior beliefs in a U.S. setting, while also complementing evidence from secondary schools in Spain (Azmat & Iriberri, 2010) and university settings elsewhere (Chen et al., 2024; Dobrescu et al., 2021; Brade et al., 2026).

By providing causal evidence that relative performance feedback improves outcomes for students who arrive with over-optimistic expectations, we show that informational interventions can benefit precisely those students whose beliefs most need recalibration. This finding aligns with Aucejo and Wong (2025) and helps reconcile the mixed results in the literature: average treatment effects can mask sizeable heterogeneity along a belief-based dimension that has received little empirical attention, even though theory places it at the centre of how feedback works.

3. Context, Experimental Design and Data

This study was conducted at Universidad Rey Juan Carlos (URJC), one of the public universities in the Madrid region. The intervention targeted undergraduate students enrolled in

applied economics and business-related degrees, including Business Administration, Accounting and Finance, Marketing, and Architecture. The Spanish institutional environment shares several features with those examined in previous higher-education experiments, such as centralised curricula and a strong emphasis on final examinations. There is limited use of ranking or comparative feedback in economics and finance subjects, which makes this a suitable setting to test the effects of relative performance feedback.

3.1 Institutional and Academic Context

Access to higher education in Spain is available through several routes: the standard pathway via the national university entrance examination (Evaluación para el Acceso a la Universidad, EVAU), vocational training programmes (Formación Profesional, FP), and other non-traditional entry channels. The EVAU is by far the most common route in our sample, accounting for roughly three-quarters of enrolled students. Within this system, there is limited use of ranking or comparative feedback in course management, making it a suitable context in which to test the impact of relative performance feedback.

At URJC, students are assessed in each subject through both a midterm and a final examination. The midterm is primarily practical and typically takes place around half to two-thirds of the way through the semester. The final examination, scheduled during the official examination period, includes both theoretical and practical components; our primary outcome is the theoretical component of the final exam. Final grades are determined by a weighted average of 40% from the midterm and 60% from the final examination, with marks awarded on a 0–10 numerical scale.

3.2 Experimental Design and Implementation

The intervention was implemented as an individual-level randomised controlled trial and is registered with the AEA RCT Registry (ID: AEARCTR-0014016). The trial began in January 2022 and concluded in December 2024. Students were randomly assigned to treatment and control groups using Stata's random-number generator, with stratification by course and academic year to ensure balance within each programme cohort.

The treatment group received the absolute midterm score alongside a personalised message indicating the student's percentile rank within the course cohort. The control group received only the absolute midterm grade. Feedback was delivered electronically through the university's internal messaging system, accompanied by a short explanation of what the percentile rank represented, in order to reduce misinterpretation. The intervention was implemented between the midterm and final exams, enabling clean identification of its effect on subsequent academic effort.

Covariate information was collected through baseline surveys at the beginning of the course. These captured demographic characteristics, academic background, expectations regarding performance, and access to resources such as internet, personal computers, and quiet study spaces. The baseline survey elicited each student's self-reported expected academic performance on a four-point scale (very low, low, medium, high), which we use to construct the expectation-shock

measure described in Section 4.3. At the end of the course, a follow-up questionnaire assessed students' perceptions of the feedback, along with self-reported motivation and effort.

3.3 Sample Description and Timeline

The analytical sample consists of 386 students with valid treatment assignments (195 in the control group and 191 in the treatment group). The main regression models use between 331 and 381 of these observations depending on covariate availability, as some students have missing values for specific control variables collected in the baseline survey. The experiment covered four undergraduate degree programmes: Valuation and Acquisition of Companies (Business Administration, Year 4); Financial Planning and Business Valuation (Accounting and Finance, Years 3 and 4); Economic-Financial Analysis II (Marketing, Year 2); and Real Estate Management (Architecture, Year 3).

Each semester followed a standardised structure: the midterm was held in weeks 8–10, feedback was delivered in week 12, and the final exam took place in weeks 15–16. The feedback intervention was low-cost, required minimal administrative input, and is readily scalable across similar university settings. The primary outcome is the final exam score (measured on a 0–10 scale). We also analyse the change in score between the final and midterm exams as an indicator of academic progress. Additional covariates include parental education, scholarship status, and course enrolment load, which allow us to examine heterogeneous treatment effects across student subgroups.

Table 1. Descriptive statistics

Variable	Mean	SD	Min	Max	N	Mean	Mean
	(All)					*(Control)*	*(Treatment)*
Enrolments	1.19	0.55	1	6	386	1.17	1.21
Previous exam attempts	1.32	1.01	1	11	386	1.29	1.35
Midterm score	4.91	2.69	0	10	383	4.88	4.94
Final exam score	5.10	2.41	0	10	384	5.05	5.14
Age	21.20	2.54	19	48	379	21.07	21.34
Gender (1 = male, 2 = female)	1.55	0.50	1	2	384	1.53	1.57
Birth month	6.67	3.48	1	12	357	6.81	6.53
Siblings	1.17	0.67	0	3	383	1.16	1.19
Courses enrolled	2.36	0.69	1	3	382	2.40	2.32
Scholarship	0.33	0.47	0	1	382	0.31	0.35
Driving licence	0.68	0.47	0	1	385	0.69	0.66
Own car	0.34	0.47	0	1	384	0.38	0.30
Own computer	0.99	0.10	0	1	385	0.99	0.99
Internet access	1	0	1	1	385	1	1
Employed	0.28	0.45	0	1	383	0.28	0.27
Extra classes	0.15	0.36	0	1	379	0.12	0.18
Father education	3	0.92	1	4	381	3.06	2.94
Mother education	3.09	0.88	1	4	386	3.20	2.97

Expected performance	3.18	0.60	1	4	386	3.22	3.14
N (Total)					**386**	**195**	**191**

Notes: The analytical sample includes 386 students with valid treatment assignments (195 control, 191 treatment). Some variables have fewer observations due to item non-response in the baseline survey. Gender is coded 1 = male, 2 = female. Parental education and expected performance are measured on four-point ordinal scales.

Table 1 presents summary statistics for the main variables used in the analysis. On average, students enrolled in just over one course and attempted the exam around 1.3 times, with a mean midterm score of 4.91 and a final exam score of 5.10 on a 0–10 scale. The average age of students is just over 21 years, with balanced representation by gender. Access to study-related infrastructure is high, with nearly all students reporting ownership of a personal computer and internet access. Only around 33% hold a scholarship, and less than 28% report working while studying. Parental education is skewed towards medium and high levels, and most students report relatively high expected academic performance. Overall, the table highlights considerable variation in baseline characteristics.

Table 2. Balance test between treatment and control groups

Variable	Mean Control	Mean Treatment	Difference	t / χ^2 stat.	p-value	Test
Enrolments	1.169	1.215	0.045	−0.804	0.422	t-test
Previous exam attempts	1.287	1.351	0.064	−0.618	0.537	t-test
Midterm score	4.877	4.936	0.059	−0.213	0.832	t-test
Final exam score	5.050	5.143	0.093	−0.378	0.706	t-test
Age	21.068	21.335	0.267	−1.020	0.309	t-test
Gender	1.529	1.571	0.042	−0.829	0.408	t-test
Birth month	6.814	6.528	−0.286	0.776	0.438	t-test
Siblings	1.162	1.189	0.027	−0.395	0.693	t-test
Courses enrolled	2.402	2.325	−0.078	1.096	0.274	t-test
Scholarship	—	—	—	0.884	0.347	χ^2
Extra classes	—	—	—	2.285	0.131	χ^2
Father education	—	—	—	1.893	0.595	χ^2
Mother education	—	—	—	12.454	0.006	χ^2
Expected performance	—	—	—	9.570	0.023	χ^2
Entry route	—	—	—	1.530	0.465	χ^2
Residence area	—	—	—	0.549	0.760	χ^2

Notes: For continuous and ordinal variables, mean differences are tested with unequal-variance t-tests; categorical variables are tested with Pearson χ^2 tests, for which group means are not reported. $***p < 0.01$, $**p < 0.05$, $*p < 0.10$.

Table 2 presents the balance test between the treatment and control groups across baseline covariates. For continuous and ordinal variables, mean differences are small and none are statistically significant at conventional levels. For example, the average age is 21.07 in the control group and 21.34 in the treatment group (p = 0.309), while the midterm score averages 4.88 and 4.94, respectively (p = 0.832). Most categorical variables also show no evidence of distributional

imbalance. The χ^2 tests do, however, indicate distributional differences for mother's education ($p = 0.006$) and expected academic performance ($p = 0.023$). Importantly, the direction of these imbalances favours the control group: control students have mothers with higher average education (3.20 vs. 2.97) and report higher expected academic performance (3.22 vs. 3.14). Since both variables are positively correlated with academic outcomes, this imbalance works against finding a positive treatment effect. We control for both variables in the full specification, so that to the extent residual imbalance persists, our treatment-effect estimates should be interpreted as a conservative lower bound of the true causal effect. We return to the imbalance in expected performance in Section 5.2, where it directly motivates the expectation-shock analysis. Overall, these results support the internal validity of the randomised controlled trial design.

4. Empirical Strategy

Our empirical strategy follows the intent-to-treat (ITT) framework commonly employed in the education literature (Azmat et al., 2019; Megalokonomou & Zhang, 2024; Macías, 2026). This approach estimates the causal effect of treatment assignment on academic outcomes, regardless of whether students actively engaged with the feedback. The ITT framework is appropriate for our setting because it captures the policy-relevant effect of offering percentile feedback to students and avoids the selection bias that would arise from conditioning on actual feedback receipt or attention. Because assignment is random, the ITT estimand also recovers the average effect of being offered feedback for the full population of assigned students; we use the term "average marginal effect of the treatment" throughout, rather than "average treatment effect," to make explicit that our estimates are derived from the treatment-assignment indicator within this ITT design.

We complement the main analysis with heterogeneity analyses designed to capture subgroup-specific effects. As recent work on relative performance feedback emphasises, average effects may obscure meaningful differences across student subgroups (Azmat et al., 2019; Chen et al., 2024; Aucejo & Wong, 2025). Students with different baseline beliefs, academic positions, and access to alternative support may respond differently to the same informational intervention. In this section we outline the main specifications, discuss our identification strategy, and explain how we address key methodological challenges.

4.1 Baseline Specification

Our primary outcome of interest is the theoretical component of the final examination score of student i in programme c, denoted Y_{ic}, measured on a 0–10 scale. This outcome is observed several weeks after treatment, allowing sufficient time for students to adjust their effort in response to the feedback. The key explanatory variable is a treatment indicator T_i, equal to 1 for students randomly assigned to the treatment group, who received percentile feedback alongside their absolute midterm score, and 0 for students in the control group, who received only their absolute midterm score.

To increase precision and control for pre-existing academic differences, we include the midterm exam score M_i as a covariate. Conditioning on a pre-treatment measure of the outcome follows the ANCOVA approach standard in randomised education trials (McKenzie, 2012; Azmat & Iriberri, 2010; Megalokonomou & Zhang, 2024) and serves two purposes: it increases statistical power by reducing residual variance, and it allows us to interpret the treatment effect as the impact of feedback conditional on prior academic ability.

We also control for a vector of individual-level pre-treatment characteristics X_i, including gender, age, parental education, scholarship status, employment status, and access to resources such as a personal computer and a quiet study space. Since students are drawn from different courses and academic years, we include programme fixed effects μ_c to absorb systematic differences in grading standards, course difficulty, and cohort composition. Our baseline estimation model is:

$$Y_{ic} = \alpha + \beta T_i + \gamma M_i + X_i'\delta + \mu_c + \varepsilon_{ic} \quad (1)$$

where Y_{ic} is the final exam score for student i in programme c; T_i is the treatment indicator; M_i is the midterm score; X_i is the vector of individual controls; μ_c denotes programme fixed effects; and ε_{ic} is the error term. We report heteroskedasticity-robust standard errors, since the small number of courses makes cluster-robust inference at the course level unreliable. The coefficient β captures the average effect of being offered percentile feedback on final exam performance, conditional on baseline characteristics and fixed effects.

4.2 Heterogeneity Analysis

We also examine whether the impact of the intervention differs across student types. The reference-dependent preferences framework (Kőszegi & Rabin, 2006) predicts that students respond differently to relative performance feedback depending on their prior beliefs about their academic standing: students who discover that they are performing worse than expected may increase effort, while those who learn that they are performing better than expected may reduce effort due to complacency (Azmat et al., 2019). Recent evidence further suggests that feedback may be more valuable for students whose expectations are misaligned with their realised performance (Aucejo & Wong, 2025) and for those who lack alternative sources of academic support (Zhou et al., 2025).

To investigate these patterns, we estimate heterogeneous treatment effects by interacting the treatment indicator T_i with indicators for the following pre-determined subgroups: baseline academic performance (students with midterm grades below 5); socioeconomic status proxied by scholarship receipt; parental education (students whose mothers have low education); access to alternative academic support (participation in extra classes); expected academic performance (students reporting the lowest categories in the baseline survey); and the expectation shock constructed as the gap between the expected and the realised midterm grade. Each interaction is introduced separately into the extended specification. For a generic subgroup indicator Z_i, the regression takes the form:

$$Y_{ic} = \alpha + \beta_1 T_i + \beta_2 Z_i + \beta_3 (T_i \times Z_i) + \gamma M_i + X_i' \delta + \mu_c + \varepsilon_{ic} \quad (2)$$

where Z_i is the subgroup indicator (for example, an indicator for students with midterm grades below 5). The coefficient β_3 captures the differential effect of the intervention for the subgroup relative to its complement. This design follows closely the approach used by Azmat et al. (2019), Chen et al. (2024), and Megalokonomou and Zhang (2024) to uncover behavioural responses to informational interventions.

Rather than focusing on raw interaction coefficients, we report average marginal effects of the treatment evaluated at the relevant values of the interacting variable. This approach matters because the interaction coefficient alone does not provide a complete test of a conditional hypothesis. As emphasised by Kingsley et al. (2017), the statistical significance of an interaction coefficient indicates only whether marginal effects differ across values of the moderating variable, not whether those marginal effects are themselves statistically different from zero. Relying solely on the interaction coefficient may therefore overstate or understate empirical support for a hypothesis. Evaluating marginal effects across the range of the moderating variable allows for a more accurate characterisation of treatment heterogeneity and a transparent comparison across subgroups.

4.3 The Expectation Shock

Building on Aucejo and Wong (2025), we construct a direct measure of the gap between students' prior beliefs about their own performance and their realised academic outcomes. We are not the only study with information on student expectations, since the literature on over- and under-estimation is well established, but our pre-treatment elicitation, combined with the single-shock design, allows us to use this gap as a clean conditioning variable. At the beginning of the course, each student reported her expected academic performance on a four-point ordinal scale: very low, low, medium, or high. To make this variable comparable with the 0–10 grade scale used in the Spanish higher-education system, we map these categories onto grade-point equivalents consistent with the standard qualitative labels: very low → 2.5 (clear fail), low → 5.0 (borderline pass), medium → 7.0 (good), and high → 9.0 (excellent). The resulting variable, E_i, is a self-reported expected grade.

We then define the expectation shock as the difference between the expected grade and the midterm grade:

$$S_i = E_i - M_i \quad (3)$$

A positive value of S_i indicates that the student was over-optimistic at the beginning of the course, since her ex-ante expected grade exceeded her first observed performance, while a negative value indicates that she was pessimistic. Both components of S_i are measured before treatment delivery, which ensures that S_i itself is not affected by the experiment and is a valid pre-treatment conditioning variable. Using S_i, we build three complementary indicators: a continuous measure

of the shock; a binary indicator for over-optimism ($S_i > 0$); and a three-category classification into pessimistic ($S_i < -0.5$), accurate ($|S_i| \leq 0.5$), and over-optimistic ($S_i > 0.5$) students.

The heterogeneity specification corresponds to equation (2) with the expectation-shock indicator in place of the generic subgroup Z_i. We also report results with the continuous shock and with a strong-over-optimism indicator ($S_i > 2$ grade points) as robustness checks, and an alternative mapping of the expected-grade scale (3.0 / 5.0 / 6.5 / 8.0) in Section 5.4.

4.4 Alternative Outcome Specification

As a secondary outcome, we analyse the change in performance between the midterm and final exam, $\Delta_i = Y_i - M_i$. This difference captures academic progress over the semester and provides a measure of improvement that is independent of students' initial level. We estimate:

$$\Delta_{ic} = \alpha + \beta T_i + X_i'\delta + \mu_c + \varepsilon_{ic} \qquad (4)$$

Since the midterm exam is conducted before treatment and the final exam after the intervention, this outcome provides a proxy for academic progress attributable to the information shock, complementing the levels specification in equation (1).

4.5 Identification and Internal Validity

The causal interpretation of our estimates rests on the validity of random assignment. Randomisation was implemented at the individual level using Stata's random-number generator, stratified by course and academic year to ensure balance within each course-year cohort. We verified that observable characteristics were balanced across treatment and control groups using t-tests for numerical variables and χ^2 tests for categorical variables. As reported in Table 2, no statistically significant imbalances were detected for continuous covariates. The χ^2 tests reveal distributional differences in mother's education and expected academic performance, but these imbalances favour the control group and therefore bias against finding a positive treatment effect; both variables are included as controls in the full specification.

The expectation shock S_i introduced in Section 4.3 is itself balanced across treatment arms: the mean shock is 2.55 in the control group and 2.34 in the treatment group ($p = 0.457$), and the three-category shock distribution does not differ significantly across arms ($\chi^2(2) = 2.96$, $p = 0.228$). This balance is expected given that both the expected and the midterm grades are measured before treatment delivery, and it provides an additional validity check on the randomisation.

To further support internal validity, we conduct several robustness checks. First, we exclude outliers in midterm or final exam scores to ensure that our results are not driven by extreme values. Second, we re-run regressions on balanced panels with complete outcome and covariate data to verify robustness to sample composition. Third, we examine the stability of the midterm coefficient across specifications; its consistency provides reassurance about the robustness of the estimates. Fourth, we show that the heterogeneity-by-expectation-shock result is robust to an alternative numerical mapping of the expected-grade scale.

5. Empirical Results

This section presents the main findings from the randomised evaluation of the percentile feedback intervention. We first report the average effects of treatment assignment on students' final exam scores and their improvement relative to the midterm (Section 5.1). We then explore heterogeneous treatment effects across subgroups defined by academic ability, socioeconomic background, and self-reported expectations (Section 5.2). Section 5.3 introduces our new evidence on the role of the expectation shock, the central new contribution of this version of the paper. Section 5.4 presents robustness checks.

5.1 Average Treatment Effects

We estimate the causal effect of the treatment on students' final theoretical exam performance using specification (1). In all models the dependent variable is the final theoretical exam score, measured on a 0–10 scale, and all specifications include programme fixed effects and heteroskedasticity-robust standard errors.

Table 3. Effect of treatment on final exam score

	(1) Baseline	**(2) Full controls**
Treatment	0.168	0.406∗∗
	(0.175)	(0.203)
Midterm score	0.530∗∗∗	0.505∗∗∗
	(0.036)	(0.042)
Female		0.573∗∗∗
		(0.213)
Scholarship		−0.060
		(0.219)
Driving licence		−0.118
		(0.265)
Own car		0.504∗
		(0.265)
Own computer		−2.992∗∗∗
		(0.565)
Extra classes		−0.409
		(0.318)
Residence: Madrid region		−0.103
		(0.225)
Residence: outside Madrid		0.245
		(0.367)
Entry route: FP		−0.316
		(0.445)

Entry route: other		−0.235
		(0.789)
Constant	2.423***	5.266***
	(0.236)	(1.105)
Enrolment indicators	No	Yes
Month-of-birth indicators	No	Yes
Siblings indicators	No	Yes
Courses-enrolled indicators	No	Yes
Mother education indicators	No	Yes
Expected performance indicators	No	Yes
Programme fixed effects	Yes	Yes
Observations	381	331
R^2	0.503	0.550

*Notes: Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. The dependent variable is the final theoretical exam score (0–10). Column (1) controls only for the midterm score and programme fixed effects. Column (2) adds the full set of controls described in Section 4.1. Coefficients are reported for the continuous and binary controls; categorical controls entered as sets of indicators (number of enrolments, month of birth, number of siblings, number of courses enrolled, mother's education, and expected performance) are included but summarised as "Yes" to conserve space, with the omitted category serving as the reference in each case. Both columns include programme fixed effects.*

Table 3 reports the estimated effect of treatment assignment on students' theoretical exam scores. Column (1) presents a baseline specification including the treatment indicator, midterm performance (the practical exam score), and programme fixed effects. Here the estimated treatment effect is positive but small (0.168 points) and not statistically significant, suggesting limited evidence of an average effect when only minimal controls are included.

Column (2) augments the baseline model with the full set of controls capturing enrolment characteristics, demographic background, family circumstances, and expected academic performance. Once these controls are included, the estimated treatment effect increases to 0.406 points and becomes statistically significant at the 5% level. This indicates that, conditional on observable characteristics and programme fixed effects, the treatment raises the theoretical exam score by an amount equivalent to approximately 17% of a standard deviation in final exam scores. Moreover, as discussed in Section 3.3, the distributional imbalances detected in the balance test favour the control group on dimensions positively correlated with academic outcomes, so this estimate may represent a conservative lower bound of the true treatment effect.

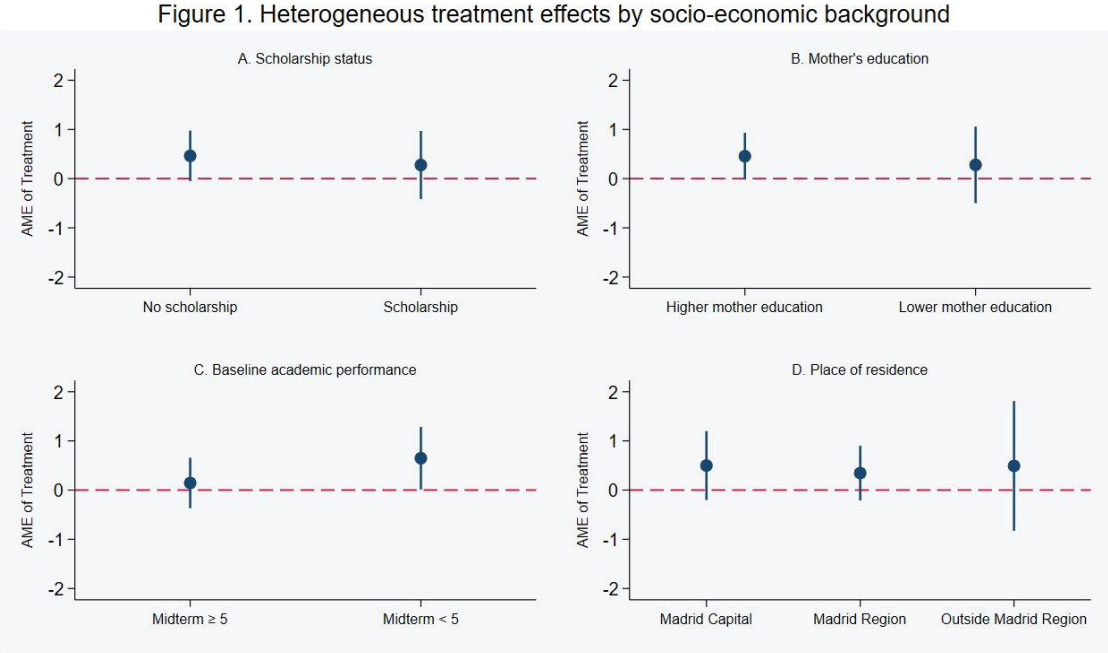
Across both specifications, performance in the practical (midterm) exam is a strong and highly significant predictor of theoretical exam outcomes, with a coefficient close to 0.5. This confirms the importance of controlling for baseline academic ability. Among the additional controls, female students score about 0.57 points higher than male students conditional on midterm performance ($p < 0.01$), and owning a private car is associated with a small positive coefficient. The negative coefficient on owning a personal computer should not be read causally: almost all students in the sample own one (see Table 1), so the indicator identifies a very small and atypical group. The

inclusion of additional controls reduces the sample size due to missing covariate information but improves explanatory power, as reflected in the higher R^2 in Column (2). The stability of the midterm coefficient across specifications provides further reassurance about the robustness of the estimates.

5.2 Heterogeneity by Student Characteristics

To investigate whether the effect of treatment varies across observable student characteristics, we estimate the interaction specification in equation (2) for each pre-determined subgroup. As explained in Section 4.2, and following Kingsley et al. (2017), we report average marginal effects of the treatment evaluated at the relevant values of the interacting variable rather than the raw interaction coefficients, since this conveys both whether marginal effects differ across groups and whether each marginal effect is itself distinguishable from zero. Figures 1–4 present coefficient plots of these average marginal effects across a wide range of observable student characteristics; all regressions use the full set of controls and programme fixed effects (N = 331).

Figure 1. Average marginal effects of treatment by socio-economic background



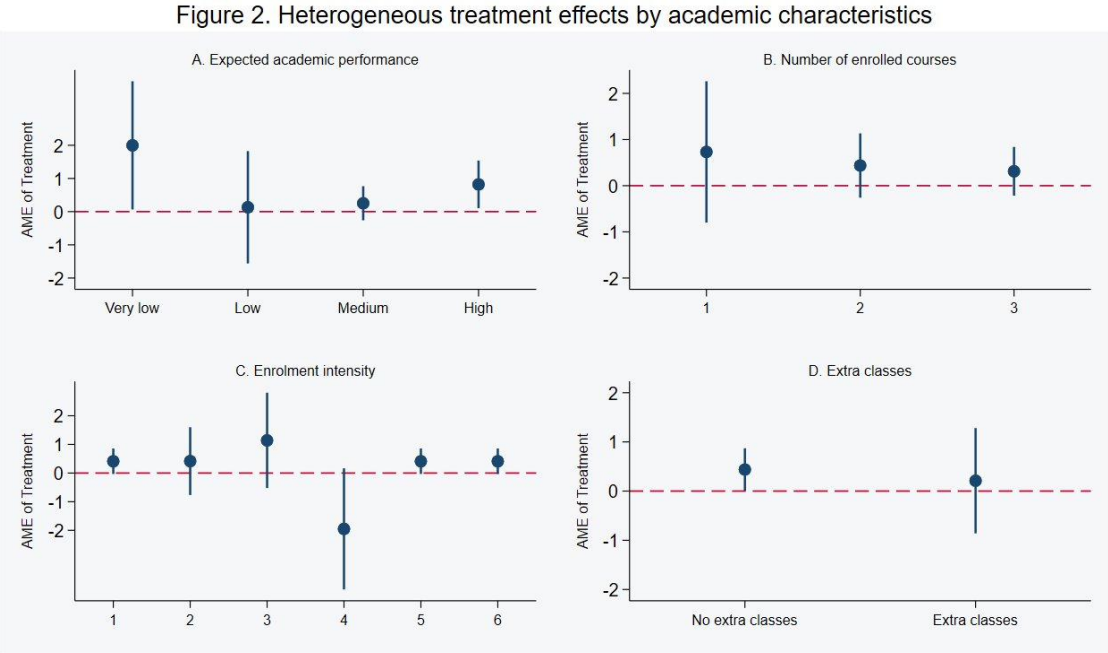
Notes: Each panel reports the average marginal effect of the treatment on the final theoretical grade, computed from the interaction specification in equation (2). Point estimates are shown with 95% confidence intervals based on heteroskedasticity-robust standard errors. All regressions include programme fixed effects and the full set of controls described in Section 4 (N = 331).

Figure 1 examines heterogeneity by socio-economic background: scholarship status, mother's education, baseline academic position, and place of residence. Across most dimensions the estimated effects are similar in magnitude and the confidence intervals overlap substantially. Students who are not on a scholarship show a positive marginal effect of about +0.45 grade points,

marginally significant, while those receiving a scholarship show a smaller, statistically insignificant positive effect. No systematic gradient appears with respect to mother's education: students with higher- and lower-educated mothers show similar marginal effects of around +0.30 to +0.45, both with confidence intervals that include zero.

An interesting pattern emerges by baseline academic position. The estimated treatment effect for students in the lowest academic percentile (midterm grade below 5) is positive (about +0.65 grade points) and larger than for higher-performing students, consistent with the theoretical prediction that students who receive information suggesting room for improvement increase effort to close the gap (Azmat et al., 2019; Chen et al., 2024) and with the pattern reported by Megalokonomou and Zhang (2024). Differences across place of residence are small and statistically insignificant, suggesting that the intervention is equally effective regardless of where students live.

Figure 2. Average marginal effects of treatment by academic characteristics



Notes: See the note to Figure 1.

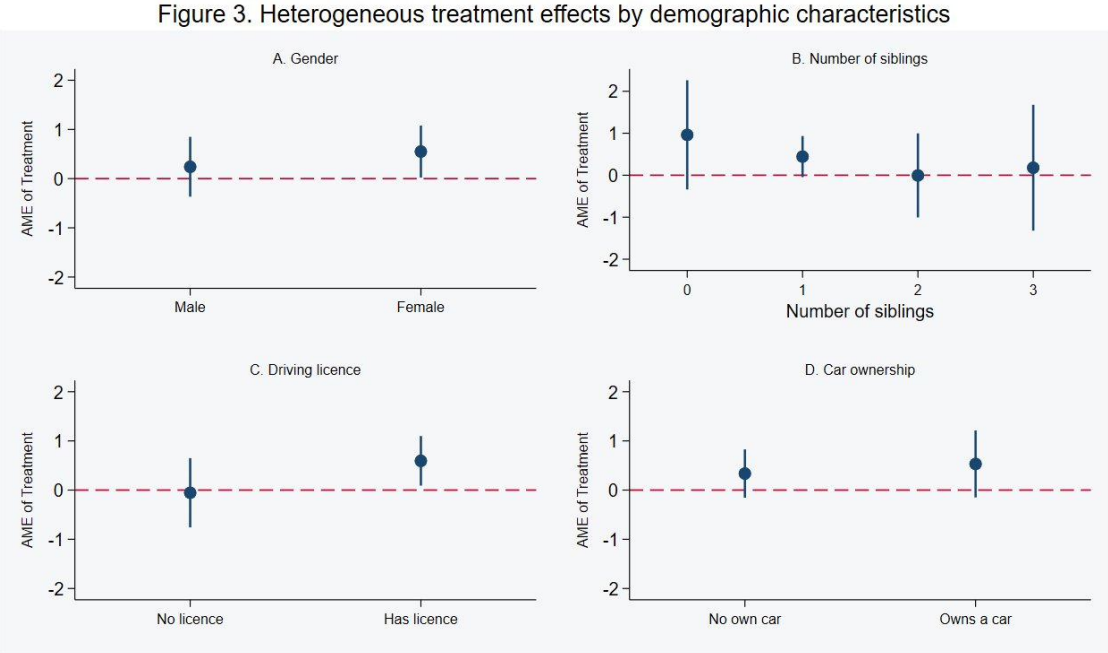
Figure 2 focuses on heterogeneity by academic characteristics: expected academic performance, number of enrolled subjects, number of enrolments, and participation in extra classes. The most striking pattern appears in panel A: the treatment effect for students who report "very low" expected performance is large and positive (around +2 grade points), with a confidence interval that excludes zero. Students with "high" expectations also display a positive marginal effect of around +0.8, while students with "low" and "medium" expectations show effects close to zero. This non-monotonic pattern suggests that informational feedback is especially beneficial for students with extreme prior beliefs about their own performance, either very low or very high,

plausibly because the information helps correct misaligned beliefs or recalibrate effort. This pattern is the visual counterpart of the expectation-shock mechanism that we examine systematically in Section 5.3.

Differences across the number of enrolled subjects are small and statistically insignificant, indicating that course load does not meaningfully moderate the treatment effect. Heterogeneity by the number of enrolments (which captures whether the student is taking the course for the first time or has re-enrolled) reveals positive marginal effects for first-time enrolment and lower enrolment counts, with one outlier point estimate for the highest enrolment count that has a wide confidence interval and reflects a small subgroup. The pattern is broadly consistent with informational feedback being more useful for first-time enrolees, who lack prior course-specific information.

Regarding participation in extra classes, students who do not attend extra classes experience a stronger and marginally significant treatment effect (marginal effect ≈ +0.45), whereas students receiving additional instruction show a smaller and statistically insignificant point estimate. This is suggestive of informational feedback being more valuable when alternative sources of academic support are absent, consistent with information and external tutoring acting as partial substitutes. The finding has clear policy implications: informational interventions may be most cost-effective when targeted at students who lack access to private tutoring or supplementary instruction.

Figure 3. Average marginal effects of treatment by demographic and household characteristics

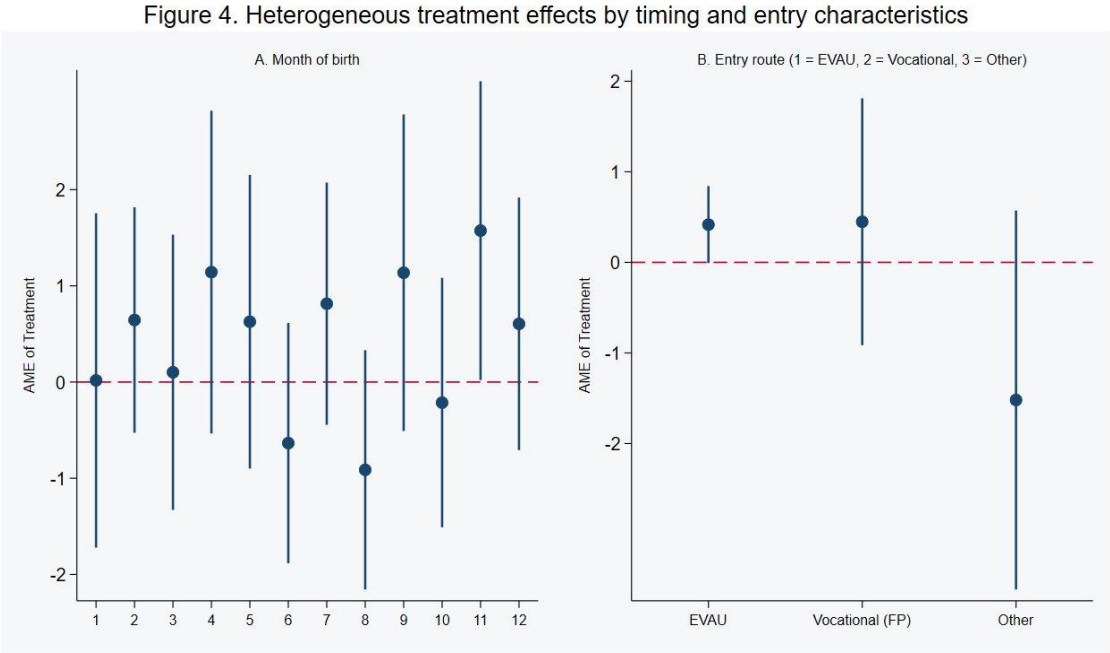


Notes: See the note to Figure 1.

Figure 3 reports heterogeneity by gender, number of siblings, possession of a driving licence, and access to a private car. A clear gender pattern emerges in panel A: the estimated treatment effect is positive and marginally significant for female students (marginal effect ≈ +0.55, with the lower bound of the 95% confidence interval close to zero), while it is smaller and not statistically significant for male students (marginal effect ≈ +0.25). This indicates that informational feedback is more effective for female students, consistent with evidence that affirming or informative feedback can narrow gender gaps in confidence, particularly in quantitative and male-dominated fields (Azmat & Iriberri, 2016; Aucejo & Wong, 2025).

Heterogeneity by number of siblings shows a slightly declining pattern in the point estimate from zero to two or more siblings, but the confidence intervals are wide and overlap substantially. Students with a driving licence show a positive marginal effect (≈ +0.55) that is marginally significant, while the estimate for those without a licence is not statistically significant. Differences by car ownership are small and statistically insignificant. Overall, household composition and mobility constraints do not appear to systematically moderate the treatment effect.

Figure 4. Average marginal effects of treatment by timing and entry characteristics



Notes: See the note to Figure 1. EVAU denotes the national university entrance examination; FP denotes vocational training (Formación Profesional).

Figure 4 explores heterogeneity by month of birth and entry route. The estimated effects across months of birth fluctuate around zero without any systematic seasonal pattern, and most confidence intervals are wide given the small number of observations within each month, suggesting that relative-age effects, well documented in primary and secondary education, do not meaningfully moderate the treatment effect in our university sample. Turning to entry route, the

treatment effect is positive and statistically significant only for students admitted through the EVAU, the national entrance examination, which is the majority route in our sample (roughly three-quarters of students). The effect for students entering through vocational training (FP) is smaller and not significant, and the residual category is imprecisely estimated given its small size. The intervention is thus most clearly effective for the predominant group of students, who enter university through the standard examination pathway.

5.3 Heterogeneity by Expectation Shock

Sections 5.1 and 5.2 show a positive average effect of treatment assignment and reveal that this effect is particularly large for students at the extremes of the self-reported expected-performance distribution. This pattern is consistent with the reference-dependent preferences framework of Kőszegi and Rabin (2006) and with the evidence in Aucejo and Wong (2025) that the effectiveness of informational feedback depends on the gap between students' prior beliefs and their realised performance. In this section we exploit the direct measure of this gap introduced in Section 4.3, the expectation shock, to test the mechanism explicitly.

Before turning to the interaction analysis, we describe the distribution of the shock in our sample. On the expected-grade scale described in Section 4.3, the mean expected grade is 7.35 out of 10, while the mean midterm grade is 4.91. The resulting shock distribution is strongly right-skewed: the mean of the shock is 2.45 grade points (standard deviation 2.77) and the median is 2.50, indicating that, on average, students expect to score about 2.5 grade points higher than they actually achieve in the midterm. Overall, 79% of students in the analytical sample are over-optimistic under the binary definition that expected performance exceeds realised midterm performance (shock > 0), and 55% are strongly over-optimistic (shock > 2). Under the mutually exclusive three-category classification used in Figure 5, 74% of students are over-optimistic (shock > 0.5), 11% hold accurate expectations ($|\text{shock}| \leq 0.5$), and 15% are pessimistic (shock < -0.5). For comparison, Aucejo and Wong (2025) report that 61% of students at Arizona State University anticipated a grade higher than their final course grade, with an average overestimation of 0.9 letter-grade points. The larger prevalence of over-optimism in our setting may reflect differences in the elicitation format (a qualitative four-point scale rather than a direct numeric forecast) and in the institutional context, but the qualitative pattern, with most students over-optimistic about their own performance, is strikingly similar.

Before interacting the shock with the treatment, we replicate the cross-sectional result of Aucejo and Wong (2025, Table 2) by regressing the shock on pre-determined student characteristics. Table 4 reports the results. In the parsimonious specification (column 1), female students display a significantly smaller shock than male students (-0.556, $p < 0.05$), consistent with the evidence that women tend to be less overconfident about their academic abilities (Azmat & Iriberri, 2016; Aucejo & Wong, 2025). As additional controls are added (columns 2 and 3), the gender gap attenuates, but the midterm grade emerges as a strong negative predictor of the shock: a one-point increase in the midterm grade is associated with a 0.46-point smaller expectation–

reality gap ($p < 0.01$). In other words, students who actually perform better at the midterm are closer to their own expectations, while weaker students overestimate their performance the most. Scholarship status and mother's education have no significant independent association with the shock after conditioning on performance.

Table 4. Correlates of the expectation shock (expected − midterm grade)

	(1) Baseline	(2) Academic controls	(3) Full specification
Female	−0.556∗∗	−0.439∗	−0.262
	(0.259)	(0.229)	(0.237)
Scholarship	−0.099	0.077	0.055
	(0.277)	(0.245)	(0.244)
Low mother education	−0.242	−0.056	0.050
	(0.300)	(0.254)	(0.258)
Midterm score		−0.461∗∗∗	−0.458∗∗∗
		(0.040)	(0.047)
Enrolments		−0.297	0.549
		(1.112)	(1.107)
Previous exam attempts		0.152	−0.313
		(0.583)	(0.585)
Extra classes			0.084
			(0.331)
Age			−0.081
			(0.063)
Courses enrolled			0.139
			(0.198)
Constant	2.660∗∗∗	4.896∗∗∗	5.251∗∗∗
	(0.220)	(0.658)	(1.554)
Programme fixed effects	No	No	Yes
Observations	378	375	358
R^2	0.014	0.250	0.366

Notes: OLS regressions with robust standard errors in parentheses. ∗∗∗ $p < 0.01$, ∗∗ $p < 0.05$, ∗ $p < 0.10$. The dependent variable is the expectation shock (expected minus midterm grade). "Female" is an indicator for female students; "Low mother education" for students whose mothers did not complete upper-secondary education. Column (3) includes programme fixed effects (one large positive coefficient corresponds to the Architecture programme, where over-optimism is most pronounced).

We then estimate equation (2), replacing the subgroup indicator in turn with the continuous shock, the binary over-optimism dummy (shock > 0), a strong-over-optimism dummy (shock > 2), and the mutually exclusive three-category shock classification. Table 5 reports the key coefficients. The continuous-shock specification (column 1) yields a positive interaction coefficient of 0.131 (SE = 0.075, p = 0.080): each additional point of over-optimism is associated

with an approximately 0.13-point larger treatment effect. The over-optimism-dummy specification (column 2) is more informative: the interaction coefficient is 0.991 (SE = 0.450, p = 0.029), meaning that over-optimistic students respond to the feedback by about one full grade point more than non-over-optimistic students. The strong-over-optimism specification (column 3) shows a similar pattern with an interaction coefficient of 0.641 (SE = 0.408, p = 0.118), and the three-category specification (column 4) yields a particularly large and significant interaction with the over-optimistic group (Treatment × Over-optimistic = 1.094, SE = 0.475, p = 0.022).

Table 5. Heterogeneous treatment effects by expectation shock

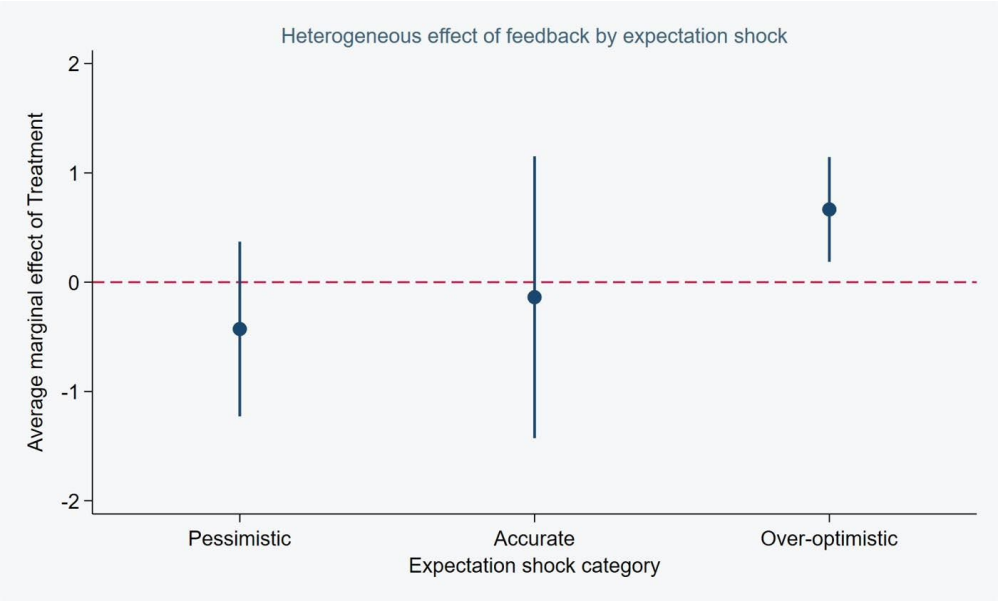
	(1) Continuous	(2) Over-optimism	(3) Strong over-optimism	(4) Three categories
Treatment	0.079	−0.366	0.054	−0.429
	(0.248)	(0.377)	(0.280)	(0.406)
Shock (continuous)	−0.056			
	(0.106)			
Treatment × Shock	0.131∗			
	(0.075)			
Treatment × Over-optimistic		0.991∗∗		
		(0.450)		
Treatment × Strong over-optimistic			0.641	
			(0.408)	
Treatment × Over-optimistic (3-cat.)				1.094∗∗
				(0.475)
Midterm score	0.520∗∗∗	0.525∗∗∗	0.536∗∗∗	0.575∗∗∗
	(0.097)	(0.060)	(0.073)	(0.065)
Full controls	Yes	Yes	Yes	Yes
Programme fixed effects	Yes	Yes	Yes	Yes
Observations	331	331	331	331

Notes: OLS regressions with robust standard errors in parentheses. ∗∗∗ p < 0.01, ∗∗ p < 0.05, ∗ p < 0.10. The dependent variable is the final theoretical exam score (0–10). All columns include the full set of controls used in column (2) of Table 3 and programme fixed effects (N = 331). "Over-optimistic" is an indicator for shock > 0; "Strong over-optimistic" for shock > 2; in column (4) the omitted category is the pessimistic group.

The magnitudes are clearer when we compute the average marginal effect of the treatment at each level of the shock variable. In the over-optimism-dummy specification, the estimated marginal effect is −0.37 (SE = 0.38, p = 0.333) for students who are not over-optimistic and +0.63 (SE = 0.24, p = 0.009) for those who are; the latter is statistically significant at the 1% level and represents an effect size of about 26% of a standard deviation in final exam scores. In the three-category specification, the marginal effects are −0.43 (SE = 0.41, p = 0.292) for pessimistic students, −0.14 (SE = 0.66, p = 0.834) for accurate students, and +0.67 (SE = 0.24, p = 0.007) for over-optimistic students. Figure 5 displays these estimates. The contrast between pessimistic and over-optimistic students is economically large, a difference of about 1.1 grade points on the 0–10

scale, and statistically meaningful: only the over-optimistic group shows a positive marginal effect whose confidence interval excludes zero, while the pessimistic group displays a negative point estimate, consistent with the prior-beliefs mechanism.

Figure 5. Average marginal effect of treatment by expectation-shock category



Notes: Point estimates and 95% confidence intervals for the average marginal effect of treatment (percentile feedback) on the final exam score, by expectation-shock category. Pessimistic: expected grade below midterm grade by more than 0.5 points. Accurate: |expected − midterm| ≤ 0.5. Over-optimistic: expected grade exceeds midterm grade by more than 0.5 points (N = 331). All controls of column (2) of Table 3 and programme fixed effects are included.

This pattern is consistent with the prior-beliefs mechanism emphasised by Aucejo and Wong (2025) and, more generally, with reference-dependent models of educational effort. Students who arrive at the intervention with expectations that overshoot their actual midterm performance receive the percentile feedback as a disconfirming signal that recalibrates their beliefs downward; if they respond to this negative update by increasing effort, the feedback generates a positive treatment effect. By contrast, students who were already pessimistic learn that they are, if anything, doing better than they thought, and the resulting upward belief update leaves little room for a positive effort response. The point estimate we recover for this group is even negative, though it should be interpreted with caution given the relatively small share of pessimistic students in the sample.

A key feature of our identification strategy is that both components of the expectation shock are measured before treatment delivery: the expected-performance variable is elicited in the beginning-of-course survey, and the midterm grade is observed before the feedback intervention is rolled out. The shock is therefore a valid pre-treatment conditioning variable, and its distribution is balanced across treatment arms (see Section 4.5). This rules out the main concerns that would arise if the shock were measured after treatment or could be affected by anticipation.

5.4 Robustness

We assess the robustness of the expectation-shock result along several dimensions. First, we replace the expected-grade mapping used in the main specification (2.5 / 5.0 / 7.0 / 9.0) with a narrower mapping that spreads the four categories across a smaller range of the 0–10 scale (3.0 / 5.0 / 6.5 / 8.0). The interaction between the over-optimism indicator and the treatment remains positive at 0.750 (SE = 0.414, p = 0.071), marginally significant under this more compressed mapping; the slight reduction in magnitude is consistent with the narrower mapping mechanically reducing the variation in the shock. The full estimates are reported in Table A1 in the appendix.

Second, we verify that the result is not driven by the small set of students with extreme shock values: restricting the sample to students with $|\text{shock}| \leq 6$ leaves the point estimates essentially unchanged. Third, we check that the three-category classification is not sensitive to the cutoff used to define the "accurate" group: widening the threshold from ±0.5 to ±1.0 grade points preserves the qualitative pattern of a larger, positive effect among over-optimistic students. Fourth, we re-run the balance checks of Section 4.5 separately within each shock category and find no evidence of differential selection into treatment, confirming that the random assignment holds within each belief group. Finally, the imbalances detected in the overall balance test (Section 3.3) favour the control group on dimensions positively correlated with academic outcomes, suggesting that the positive marginal effect among over-optimistic students is, if anything, underestimated.

Summary of Heterogeneity Findings

The heterogeneity analysis shows that the impact of informational feedback is not uniform. While the treatment effect is broadly similar across many socio-economic, demographic, and institutional dimensions, systematic and statistically significant differences emerge for specific subgroups: the treatment is more effective for students with lower baseline academic performance, for students who do not receive additional instruction through extra classes, for female students, and for students who arrive at the intervention with over-optimistic expectations about their own performance. The evidence in Section 5.3 suggests that this last dimension is in many respects the unifying mechanism behind the others: lower-performing students are, on average, more over-optimistic, and belief recalibration is the most plausible channel through which the percentile-rank signal induces the observed effort response.

These patterns suggest that informational feedback is most useful when students have greater scope for improvement, face higher uncertainty about their own performance, or lack alternative sources of academic support. The finding that feedback benefits students who do not attend extra classes is of practical interest, since it suggests that informational interventions can partly substitute for more resource-intensive forms of academic support at a much lower cost.

6. Conclusion

This paper provides experimental evidence on the effectiveness of relative performance feedback in higher education. We conduct a pre-registered randomised controlled trial involving 386 undergraduate students across four degree programmes at Universidad Rey Juan Carlos in Madrid, Spain. Following a midterm exam, students in the treatment group receive information about their percentile ranking within the class, while those in the control group receive only their absolute scores. This design isolates the causal effect of a single informational intervention delivered at a strategic moment, when students can still adjust their effort before the final examination.

Our main finding is that relative performance feedback has a positive and statistically significant effect on final exam scores when we control for baseline performance and student characteristics. The treatment increases final exam scores by 0.41 points on a 0–10 scale ($p < 0.05$), an improvement equivalent to approximately 17% of a standard deviation. This estimate is likely a conservative lower bound, since the distributional imbalances in mother's education and expected academic performance favour the control group. Even a simple, low-cost informational intervention can therefore meaningfully improve academic outcomes in higher education.

The heterogeneity analysis shows that the impact of feedback is not uniform. The treatment is most effective for four groups: students with lower baseline academic performance, who have greater scope for improvement; students who do not attend extra classes, suggesting that feedback and tutoring act as partial substitutes; female students; and, above all, students whose prior expectations exceed their realised midterm performance. For students who are over-optimistic under the binary definition (shock > 0), which represents 79% of our sample, we estimate an average marginal effect of the treatment of +0.63 grade points ($p < 0.01$), compared with −0.37 for students who were not over-optimistic under this binary definition. When students are classified into pessimistic, accurate, and over-optimistic categories, only over-optimistic students show a positive marginal effect whose 95% confidence interval excludes zero (+0.67, $p < 0.01$), while pessimistic students display a negative point estimate (−0.43) and accurate students show an effect that is statistically indistinguishable from zero (−0.14). In this three-category classification, the over-optimistic group represents 74% of the sample.

This provides direct empirical support for the reference-dependent mechanism through which relative performance feedback is hypothesised to operate.

Our results suggest that the mixed evidence in the literature reflects not a contradiction but differences in students' prior beliefs and in the timing and frequency of feedback. While Azmat et al. (2019) find null or negative average effects in a Spanish university with repeated feedback over multiple semesters, we find positive effects with single-shot feedback delivered between assessments. Such feedback allows students to recalibrate their beliefs and adjust effort while improvement is still feasible, rather than inducing complacency. By showing that the treatment effect is driven by over-optimistic students, we provide direct support for the prior-beliefs mechanism highlighted by Aucejo and Wong (2025) in a U.S. setting and for the prediction of

reference-dependent preferences (Kőszegi & Rabin, 2006). In the Spanish higher-education context, where over-optimism appears even more widespread than in the U.S., feedback has the greatest scope to improve outcomes precisely because so many students hold expectations that need recalibration. Relative to Azmat et al. (2019) and Aucejo and Wong (2025), what is new here is the combination of a clean single-shock design with a pre-treatment measure of the belief–reality gap, which lets us identify the prior-beliefs channel directly rather than infer it.

The intervention is also easy to implement. Relative performance feedback requires only the calculation of percentile rankings from midterm scores and the delivery of personalised messages through standard communication channels, so universities can roll it out at low cost using existing infrastructure. Those seeking to improve academic outcomes may consider incorporating percentile feedback into their assessment practices and, where possible, directing it at students whose self-reported expectations deviate from their observed performance.

Our study has several limitations. The sample is large enough to detect meaningful average effects, but estimates for some subgroups remain imprecise and should be read with caution. The setting is also specific, namely applied economics and business courses at a Spanish public university, so the results may not carry over to settings with different grading practices, student populations, or norms around academic competition. We observe only short-term effects on final exam performance within the same semester, and whether these effects persist over longer horizons remains an open question. Finally, our expected-grade measure is ordinal and self-reported, and its mapping onto the 0–10 scale involves judgement; the qualitative pattern survives alternative mappings in our robustness checks, but finer-grained numerical expectations would allow more precise estimates of the belief–reality gap.

Future work could follow students over time to measure the long-term effects of feedback on later course performance, graduation, and labour-market outcomes; vary its content and framing to learn which designs work best; combine numerical expectations elicited at several points during the course with rank feedback to trace how beliefs evolve; and replicate the design in other institutional contexts to test how far the results travel.

References

Aucejo, E. M., & Wong, K. (2025). The effect of feedback on student performance. *Journal of Public Economics, 241*, 105274. https://doi.org/10.1016/j.jpubeco.2024.105274

Azmat, G., Bagues, M., Cabrales, A., & Iriberri, N. (2019). What you don't know… can't hurt you? A natural field experiment on relative performance feedback in higher education. *Management Science, 65*(8), 3714–3736. https://doi.org/10.1287/mnsc.2018.3131

Azmat, G., & Iriberri, N. (2016). The provision of relative performance feedback: An analysis of performance and satisfaction. *Journal of Economics & Management Strategy, 25*(1), 77–110. https://doi.org/10.1111/jems.12151

Azmat, G., & Iriberri, N. (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics, 94*(7–8), 435–452. https://doi.org/10.1016/j.jpubeco.2010.04.001

Brade, R., Himmler, O., & Jäckle, R. (2026). No student left behind? Relative feedback and university completion. *Journal of Economic Behavior & Organization, 241*, 107383. https://doi.org/10.1016/j.jebo.2025.107383

Bursztyn, L., & Jensen, R. (2015). How does peer pressure affect educational investments? *The Quarterly Journal of Economics, 130*(3), 1329–1367. https://doi.org/10.1093/qje/qjv021

Chen, J., Dobrescu, L. I., Foster, G., & Motta, A. (2024). Can leagues mitigate the demoralization effect of rank feedback? A randomized controlled trial. *Labour Economics, 90*, 102602. https://doi.org/10.1016/j.labeco.2024.102602

Collins, M., & Lundstedt, J. (2024). The effects of more informative grading on student outcomes. *Journal of Economic Behavior & Organization, 218*, 514–549. https://doi.org/10.1016/j.jebo.2023.12.001

Dobrescu, L. I., Faravelli, M., Megalokonomou, R., & Motta, A. (2021). Relative performance feedback in education: Evidence from a randomized controlled trial. *The Economic Journal, 131*(640), 3145–3181. https://doi.org/10.1093/ej/ueab043

Kingsley, A. F., Noordewier, T. G., & Vanden Bergh, R. G. (2017). Overstating and understating interaction results in international business research. *Journal of World Business, 52*(2), 286–295. https://doi.org/10.1016/j.jwb.2016.12.010

Kőszegi, B., & Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics, 121*(4), 1133–1165. https://doi.org/10.1093/qje/121.4.1133

Macías, C. (2026). Returning data is not enough: Feedback, managerial beliefs and student outcomes. *International Journal of Educational Development, 123*, 103572. https://doi.org/10.1016/j.ijedudev.2026.103572

Macías, C., & Santero, R. (2025). Evaluación con retroalimentación relativa: Un experimento aleatorio en la Universidad Rey Juan Carlos. *Revista de Innovación Docente en el Aula Universitaria (RIDAU), 1*, 39–61. https://doi.org/10.33732/ridau.41

Macías, C. (2023). Random experiment on relative performance feedback in higher education at URJC. In J. Sainz & I. Sanz (Eds.), *Addressing inequities in modern educational assessment* (pp. 127–138). Springer. https://doi.org/10.1007/978-3-031-45802-6_8

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics, 99*(2), 210–221. https://doi.org/10.1016/j.jdeveco.2012.01.002

Megalokonomou, R., & Zhang, Y. (2024). How good am I? Effects and mechanisms behind salient rank. *European Economic Review, 170*, 104870. https://doi.org/10.1016/j.euroecorev.2024.104870

Zhou, X., Wong, H. L., Wei, X., & Siebert, W. S. (2025). Improving the teacher feedback process in primary education: Evidence from randomized controlled trials in schools in rural China. *Education Economics*. Advance online publication. https://doi.org/10.1080/09645292.2024.2412680

Appendix

Table A1. Robustness — expectation-shock interaction under an alternative grade mapping (3.0 / 5.0 / 6.5 / 8.0)

	Final exam score
Treatment	−0.132
	(0.337)
Over-optimistic (alt. mapping)	−0.004
	(0.401)
Treatment × Over-optimistic (alt. mapping)	0.750∗
	(0.414)
Midterm score	0.555∗∗∗
	(0.064)
Female	0.563∗∗∗
	(0.214)
Own car	0.508∗
	(0.264)
Own computer	−2.842∗∗∗
	(0.586)
2nd enrolment	−0.003
	(0.369)
5th enrolment	2.228∗∗∗
	(0.548)
6th enrolment	−3.048∗∗∗
	(0.750)
Constant	5.201∗∗∗
	(1.126)
Other controls	Yes
Programme fixed effects	Yes
Observations	331
R^2	0.556

Notes: OLS regression with robust standard errors in parentheses. ∗∗∗ $p < 0.01$, ∗∗ $p < 0.05$, ∗ $p < 0.10$. The dependent variable is the final theoretical exam score (0–10). The model includes the full set of controls used in column (2) of Table 3 (enrolment indicators, gender, month of birth, siblings, courses enrolled, entry route, scholarship, parental education, access to resources) and programme fixed effects; selected coefficients are shown. "Over-optimistic (alt. mapping)" classifies students using the compressed expected-grade mapping 3.0 / 5.0 / 6.5 / 8.0 described in Section 5.4.