Revista Española de Educación Médica

UNIVERSIDAD DE MURCIA

Post-Class Formative Retrieval Quizzes and Academic Performance in First-Year Higher Education: Pre–Post Changes and Heterogeneous Student Profiles.

Cuestionarios Formativos de Recuperación y Rendimiento Académico en Estudiantes Universitarios de Primer Año: Cambios Pre-Post y Perfiles Heterogéneos.

Miguel A. Pérez[1]* and Gabriela Urrejola-Contreras[2].

[1]Facultad de Ciencias de la Vida, Universidad Viña del Mar, Viña del Mar, Chile. https://orcid.org/0000-0002-7257-1713.
[2]Unidad de Ciencias Aplicadas, Facultad de Ciencias de la Vida, Universidad Viña del Mar, Chile. https://orcid.org/0000-0002-8370-4550.
*Correspondencia: miguel.perez@uvm.cl

Recibido: 21/9/26; Aceptado: 26/10/26; Publicado: 28/10/26

Abstract: Introduction: Formative assessment may support learning through retrieval practice, feedback, and monitoring of academic progress, although students may respond differently to these strategies. This exploratory study examined temporal changes in academic performance following the implementation of post-class formative retrieval quizzes. **Materials and methods**: A retrospective single-cohort pre–post study was conducted at a Chilean university, including 46 students in the final analytical sample. Four formative retrieval quizzes were administered immediately after class, followed by immediate feedback. Pre- and post-implementation performance was calculated as the mean of Midterms 1–2 and Midterms 3–4, respectively. Analyses included a paired-samples t-test, Pearson correlation, simple linear regression, and exploratory K-means clustering. **Results**: Mean midterm performance increased from 3.79 (SD = 1.28) before implementation to 4.24 (SD = 1.44) afterward, corresponding to a mean difference of 0.45 points ($p = 0.004$, Cohen's $d = 0.45$). The increase was statistically significant among female students (+0.56, $p = 0.015$), but not among male students. Formative quiz performance was positively associated with final course grade ($r = 0.60$, $p < 0.001$; $R^2 = 0.365$). Exploratory cluster analysis identified three academic performance profiles: Higher, Intermediate, and Lower Performance. The Intermediate Performance group showed the largest descriptive increase (+0.8 points), whereas the Lower Performance group showed virtually no change. **Conclusions**: Academic performance was higher after the implementation of post-class formative retrieval quizzes and immediate feedback, and exploratory analyses revealed heterogeneous academic trajectories. Given the absence of a control group, these temporal changes cannot be causally attributed to the intervention. The findings provide preliminary evidence to inform further research on formative retrieval assessment and differences in students' academic performance trajectories.

Keywords: formative assessment; retrieval practice; testing effect; academic performance; higher education; cluster analysis

Resumen: Introducción: La evaluación formativa puede favorecer el aprendizaje mediante la práctica de recuperación, la retroalimentación y el seguimiento del progreso académico, aunque los estudiantes pueden responder de manera diferente a estas estrategias. Este estudio exploratorio examinó los cambios temporales en el rendimiento académico tras la implementación de cuestionarios formativos de recuperación posteriores a las clases. **Materiales y métodos**: Se realizó un estudio retrospectivo de cohorte única con diseño pre-post en una universidad chilena. La

muestra analítica final estuvo compuesta por 46 estudiantes. Se administraron cuatro cuestionarios formativos de recuperación inmediatamente después de las clases, seguidos de retroalimentación inmediata. El rendimiento previo y posterior a la implementación se calculó como el promedio de las pruebas parciales 1–2 y 3–4, respectivamente. Se realizaron una prueba t de muestras relacionadas, una correlación de Pearson, una regresión lineal simple y un análisis exploratorio de conglomerados mediante K-means. **Resultados**: El rendimiento medio en las pruebas aumentó de 3,79 (DE = 1,28) antes de la implementación a 4,24 (DE = 1,44) después, con una diferencia media de 0,45 puntos (p = 0,004; d de Cohen = 0,45). El aumento fue estadísticamente significativo entre las mujeres (+0,56; p = 0,015), pero no entre los hombres. El rendimiento en los cuestionarios formativos se asoció positivamente con la calificación final del curso (r = 0,60; p < 0,001; R^2 = 0,365). El análisis exploratorio de conglomerados identificó tres perfiles de rendimiento académico: alto, intermedio y bajo. El grupo de rendimiento intermedio presentó el mayor aumento descriptivo (+0,8 puntos), mientras que el grupo de rendimiento bajo mostró cambios prácticamente nulos. **Conclusiones**: El rendimiento académico fue mayor después de la implementación de los cuestionarios formativos de recuperación y la retroalimentación inmediata, y los análisis exploratorios revelaron trayectorias académicas heterogéneas. Debido a la ausencia de un grupo de control, estos cambios temporales no pueden atribuirse causalmente a la intervención. Los hallazgos aportan evidencia preliminar para orientar futuras investigaciones sobre la evaluación formativa basada en la recuperación y la heterogeneidad de las trayectorias académicas.

Palabras clave: evaluación formativa; práctica de recuperación; efecto prueba; rendimiento académico; educación superior; análisis de conglomerados

1. Introduction

Formative assessment supports learning by providing information that students and instructors can use to identify learning difficulties, monitor progress, and guide subsequent instruction (1, 2). Evidence suggests that formative assessment and feedback can improve academic performance, although their effectiveness depends on how assessment information is generated and used (3). Recent reviews highlight the importance of immediate, specific feedback in supporting self-regulation and learning in higher education (4).

Retrieval practice is one mechanism through which formative assessment may promote learning. By requiring learners to reconstruct previously encountered information rather than simply restudy it, retrieval practice can improve long-term retention (5, 6). Evidence from classroom studies and systematic reviews supports its application in authentic educational settings, although outcomes vary according to implementation characteristics (7, 8). Post-class retrieval quizzes provide a practical way to integrate this approach into routine teaching (9), enabling students to retrieve recently learned information while helping instructors identify concepts requiring reinforcement (10). Immediate feedback may strengthen this process by helping students recognize misconceptions and adjust their study strategies (11). However, feedback quality and question format also matter: structured feedback training has been associated with improved formative assessment performance (12), while short-answer questions may yield stronger retrieval benefits than multiple-choice questions when feedback is absent (13). Accordingly, retrieval quizzes combine active retrieval and feedback as complementary learning mechanisms (14, 15) and reflect important pedagogical competencies for educational professionals (16).

Nevertheless, formative assessment does not produce uniform benefits across instructional contexts. Its effectiveness may depend on assessment frequency, stakes, timing, and how students use the information provided. Evidence from higher education indicates that increasing assessment frequency does not necessarily improve academic performance consistently (17), emphasizing the

importance of implementation and student engagement. These considerations are particularly relevant in the first year of higher education, when students adapt to new academic expectations and learning strategies (18). Students with different levels of baseline achievement may respond differently to the same instructional approach. Consequently, average changes in academic performance may conceal heterogeneous student trajectories. Because most formative assessment studies focus on average outcomes (19, 20), exploratory cluster analysis may complement conventional analyses by identifying groups with similar performance patterns and examining differences in their pre–post trajectories.

This study examined academic performance before and after the implementation of post-class formative retrieval quizzes in a first-year undergraduate biology course during the first semester of 2026. Four low-stakes quizzes with feedback were administered after class sessions corresponding to the instructional units assessed in Midterms 3 and 4. The objectives were to (1) compare mean midterm performance before and after implementation; (2) examine the association between formative quiz performance and final course grade; and (3) explore distinct academic performance profiles and compare pre–post trajectories across these profiles. Given the retrospective single-cohort pre–post design and the absence of a concurrent control group, this exploratory study was not designed to establish a causal effect of the intervention. Rather, it aimed to characterize temporal changes in academic performance, examine the association between formative quiz performance and final achievement, and explore heterogeneity in students' academic trajectories following implementation of a retrieval-based formative assessment strategy.

2. Methods

2.1 Study design and setting

This study used a retrospective single-cohort pre–post design to examine changes in academic performance following the implementation of post-class formative retrieval quizzes in a first-year undergraduate course at a Chilean university. The study was conducted during the first semester of the 2026 academic period in a transversal biology course combining theoretical and practical instruction with continuous assessment. The intervention was introduced during the instructional units corresponding to Midterms 3 and 4. Academic performance during the preceding units, represented by Midterms 1 and 2, was used as the pre-implementation reference period, whereas performance during Midterms 3 and 4 represented the post-implementation period. Because the study did not include a concurrent control group or random allocation, the pre–post comparison was interpreted as evidence of temporal change rather than as a causal estimate of the effect of the quizzes. The implementation of the quizzes was part of the regular instructional process and was not introduced as a separately controlled experimental intervention. The four midterm examinations shared a common structure (3 development questions and 30–31 multiple-choice questions, 90 minutes, 42–46 total points) but covered different content areas: Midterms 1–2 assessed cellular and extracellular structures, while Midterms 3–4 assessed molecular biology and genetics. No formal psychometric analysis was conducted to equate their difficulty. This is acknowledged as a potential confounding variable in the Limitations section.

2.2 Participants

A total of 51 first-year students were initially enrolled in the course. Five students did not complete the course and were therefore not included in the final analytical sample. The final sample consisted of 46 students, including 29 males (63.0%) and 17 females (37.0%). The available academic records did not contain sufficient information to establish the specific reasons why the five students did not complete the course. Therefore, no assumptions were made regarding whether non-completion was related to academic performance, withdrawal from the university, programme change, course withdrawal, or other circumstances. The final analytical sample represented the

students who completed the course and had the academic records required for the principal pre–post analysis.

The sample size corresponds to the entire cohort enrolled in the course, which is typical of context-based educational studies. However, the small sample size limits statistical precision and generalizability, particularly for sex-stratified and cluster analyses, which should be considered exploratory. Regarding non-completion, the five students who did not complete the course had a mean Midterm 1 score of 3.2 (SD = 0.8), compared with 3.99 (SD = 1.28) for the 46 students who completed it. Although this difference is descriptively notable, it should be interpreted with caution: the group of non-completers is very small (n = 5), and the comparison group shows substantially greater dispersion, indicating that it also includes students with low scores. Thus, the observed difference does not by itself constitute evidence of selection bias; rather, it is a preliminary signal that would require larger samples and access to the reasons for non-completion to be properly examined. Reasons for non-completion were not available in the academic records, which precludes any causal interpretation.

2.3 Course assessments and formative retrieval quizzes

Academic records were obtained from the official course grading system following institutional authorization and ethics committee approval (CEC-UVM, protocol 99-26). All grades were expressed on the Chilean 1.0–7.0 grading scale, with 4.0 representing the passing threshold. Four summative midterm examinations were administered throughout the course. For the principal pre–post analysis, two composite indicators were calculated: Pre = mean of Midterms 1 and 2, Post = mean of Midterms 3 and 4. Four post-class formative retrieval quizzes were administered during the instructional units corresponding to Midterms 3 and 4, with two quizzes associated with each unit. The quizzes assessed recently presented course material and were designed as low-stakes formative assessments rather than graded components of the final course mark.

The four midterm examinations were designed by the teaching team following a common structure and evaluation criteria, covering the contents of the respective instructional units. However, no formal psychometric analysis was conducted to equate their difficulty. This is acknowledged as a potential confounding variable in the Limitations section. Each formative retrieval quiz consisted of 14 multiple-choice questions. Quizzes were administered immediately after the corresponding class session, with a duration of 15–20 minutes. Scores ranged from 1.0 to 7.0, but the quizzes were low-stakes and did not contribute to the final course grade. They were developed by the course instructors and aligned with the learning objectives of each session.

The quizzes were administered immediately following the corresponding instructional sessions. Feedback was subsequently provided and included correction of quiz items, identification of concepts requiring further attention, and study recommendations. The intervention therefore combined retrieval practice with formative feedback. Because these components were implemented together, the study does not permit their individual contributions to subsequent academic performance to be separated. Additional academic indicators included control assessments, laboratory assessments, the final examination, and the final course grade.

2.4 Procedure

The post-class formative retrieval quiz strategy was implemented during the instructional weeks corresponding to Midterms 3 and 4. Two quizzes were administered within each instructional unit. Students completed the quizzes immediately following the presentation of the corresponding class content. Once the questionnaires were completed, the instructors reviewed quiz items, discussed incorrect responses, highlighted concepts requiring additional attention, and provided study recommendations before the corresponding midterm assessment. The intervention

was therefore embedded within the regular course structure rather than delivered as an additional academic support programme.

2.5 Missing observations

Missing observations occurred primarily at the level of individual formative quizzes among students who completed the course. According to the available course records, missing quiz observations occurred because students did not attend the corresponding class session, experienced connectivity problems, or chose not to complete the assessment. These missing quiz observations were distinguished from course non-completion. The five students who did not complete the course were excluded from the final analytical sample, whereas missing quiz observations among the 46 course completers were retained as missing at the individual assessment level. No imputation was performed. Analyses involving quiz-specific variables were conducted using the available observations. Missing observations represented 27.7% for Quiz 1 (13 students), 25.5% for Quiz 2 (12 students), 14.9% for Quiz 3 (7 students), and 17.0% for Quiz 4 (8 students). Because the reasons for missing quiz observations may be related to attendance, connectivity, or students' decisions to participate, missingness cannot necessarily be assumed to be completely random. The average of the formative quizzes was calculated using the available observations for each student (i.e., the mean of the quizzes they completed). No imputation was performed. The effective n for each quiz was as follows: Quiz 1 (n = 37/46), Quiz 2 (n = 43/46), Quiz 3 (n = 40/46), Quiz 4 (n = 41/46). The overall formative quiz average was computed for the 46 students who completed the course.

2.6 Statistical analysis

Descriptive statistics were calculated for the principal academic variables. The primary outcome was the within-student difference between post-implementation and pre-implementation midterm performance: Difference = Post – Pre, where Pre represented the mean of Midterms 1 and 2 and Post represented the mean of Midterms 3 and 4. The distribution of within-student differences was assessed using the Shapiro–Wilk test. A paired-samples t test was used to compare pre- and post-implementation midterm performance in the total analytical sample. Cohen's d for paired observations was calculated using the mean within-student difference and the standard deviation of the paired differences, with 95% confidence intervals. The same paired analysis was conducted separately for male and female students as an exploratory sex-stratified analysis. Because no formal sex-by-period interaction was estimated, differences in statistical significance between male and female subgroups were not interpreted as evidence of differential effects by sex.

Pearson's correlation coefficient was used to examine the association between formative quiz average and final course grade. A simple linear regression was additionally fitted to characterize this association and obtain the coefficient of determination (R^2). This regression was interpreted as an exploratory association rather than as evidence that quiz performance caused differences in final grades. An exploratory K-means cluster analysis was conducted to identify heterogeneous student performance profiles. The number of clusters was selected using the elbow criterion and silhouette index, with variables standardized before analysis. The resulting clusters were described according to midterm performance, formative quiz performance, final course grade, and other available academic indicators. The profiles were labelled Higher Performance, Intermediate Performance, and Below Performance based on their observed academic achievement. The Below Performance profile was defined specifically as students whose performance fell below the course passing threshold, rather than indicating risk of university dropout or withdrawal. One-way ANOVAs with Tukey HSD post hoc comparisons were used to characterize differences among the identified profiles. Because some variables used to characterize the clusters were also involved in their construction, these analyses were interpreted descriptively rather than as independent validation of the clustering solution. Statistical significance was defined as $\alpha = .05$. Analyses were conducted using Python-based statistical software.

3. Results

3.1 Participants and descriptive performance

Fifty-one students were initially enrolled in the course. Five students did not complete the course and were therefore not included in the final analytical sample. The principal analysis included 46 students, comprising 29 males and 17 females. The mean final course grade among the 46 students was 4.50 (SD = 1.10), with observed grades ranging from 2.0 to 6.7. Mean midterm performance increased between the pre-implementation and post-implementation periods. The mean pre-implementation score, calculated from Midterms 1 and 2, was 3.79 (SD = 1.28), whereas the mean post-implementation score, calculated from Midterms 3 and 4, was 4.24 (SD = 1.44). Formative quiz performance varied across the four assessments. Mean quiz scores ranged from 3.76 for Quiz 3 to 4.33 for Quiz 1. Missing observations were more frequent for the first two quizzes than for the later assessments, with 27.7% missing for Quiz 1, 25.5% for Quiz 2, 14.9% for Quiz 3, and 17.0% for Quiz 4 (Table 1).

Table 1. Performance Characteristics of Student Clusters by Sex, Midterm Exams, Formative Quizzes, and Overall Averages.

Grades	Total (M, SD)	Cluster 0	Cluster 1	Cluster 2
n (%)	46	10 (21.7%)	20 (43.5%)	16 (34.8%)
Male	29	5	13	11
Female	17	5	7	5
Midterm Exams				
Midterm 1	3.99±1.3	5.6	3.9	3.1
Midterm 2	3.59±1.4	5.5	3.3	2.7
Midterm 3	3.95±1.6	6.0	4.0	2.6
Midterm 4	4.54±1.5	6.2	4.7	3.2
Midterm average (1-4)	4.02±1.3	5.9	4.0	2.9
Formative quizzes				
Quiz 1	4.33±1.7	5.6	4.3	3.6
Quiz 2	4.06±1.3	4.8	4.3	3.3
Quiz 3	3.76±1.6	5.1	3.9	2.8
Quiz 4	4.07±1.5	5.2	4.0	3.4
Formative average	4.06±0.9	5.2	4.1	3.3
Pre-Post Performance				
Pre (Midterm 1-2)	3.79±1.28	5.6	3.6	2.9
Post (Midterm 3-4)	4.24±1.4	6.1	4.4	2.9
Difference Post-Pre	+0.45	+0.5	+0.8	0.0
Final course grade	4.50±1.1	5.9	4.7	3.4

Note. Values represent cluster means. Cluster 0 corresponds to the Higher Performance group. Cluster 1 corresponds to the Intermediate Performance, showing substantial improvement from pre- to post-intervention. Cluster 2 corresponds to the Below Performance group, with the lowest performance and minimal improvement. Pre = average of Midterm 1 and Midterm 2; Post = average of Midterm 3 and Midterm 4. Formative quizzes were administered immediately after each class session corresponding to Midterms 3 and 4.

3.2 Pre–post change in midterm performance

Mean midterm performance increased from 3.79 (SD = 1.28) during the pre-implementation period to 4.24 (SD = 1.44) during the post-implementation period. The mean within-student increase was 0.45 points and was statistically significant, $t(45) = -3.04$, $p = 0.004$, with a moderate standardized effect size (Cohen's d = 0.45, 95% CI (0.16, 0.74)) (Figure 1A; Table 2). Thus, students' academic performance was higher during the assessment period following implementation of the

formative retrieval quizzes. However, because the study lacked a concurrent control group, the observed increase cannot be attributed specifically to the quizzes.

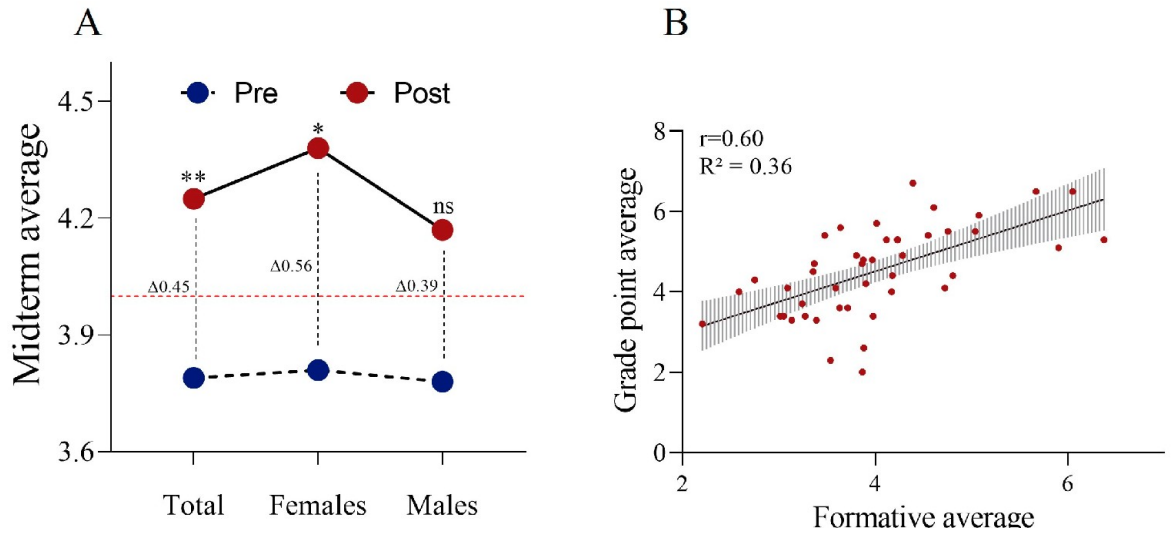


Figure 1. Pre–post change in midterm performance.

Table 2. Pre- and Post-Intervention Midterm Performance with Effect Sizes and Confidence Intervals.

Group	n	Pre (M, SD)	Post (M, SD)	Post-Pre	t	p-value	Cohen's d
Total	46	3.79 (1.28)	4.24 (1.44)	+0.45	-3.04	0.004	0.45
Female	17	3.81 (1.44)	4.38 (1.52)	+0.56	-2.72	0.015	0.66
Male	29	3.78 (1.21)	4.17 (1.41)	+0.39	-1.90	0.068	0.35

Note. Pre = average of Midterm 1 and Midterm 2; Post = average of Midterm 3 and Midterm 4. A paired-samples t-test was used to compare means. SD = Standard Deviation; CI = Confidence Interval. Cohen's d was calculated as (Post - Pre) / SD_difference. Interpretation based on Cohen's d benchmarks (0.2 = small, 0.5 = moderate, 0.8 = large).

3.3 Exploratory analysis by sex

Female students showed an increase in mean midterm performance from 3.81 (SD = 1.44) before implementation to 4.38 (SD = 1.52) afterward, corresponding to a mean increase of 0.56 points, $t(16) = -2.72$, $p = .015$, Cohen's d = 0.66, 95% CI (0.16, 1.12). Male students increased from 3.78 (SD = 1.21) to 4.17 (SD = 1.41), corresponding to a mean increase of 0.39 points, $t(28) = -1.90$, $p = 0.068$, Cohen's d = 0.35, 95% CI (−0.02, 0.72) (Figure 1A; Table 2). The numerical increase was therefore larger among female students. However, because no sex-by-period interaction was formally tested, the difference in statistical significance between the two subgroup analyses should not be interpreted as evidence that the magnitude of change differed significantly by sex.

3.4 Association between formative quiz performance and final course grade

Formative quiz performance was positively associated with final course grade ($r = 0.60$, $p < 0.001$). The corresponding simple linear regression accounted for 36.5% of the variance in final course grade ($R^2 = 0.365$). The fitted regression model was: Final course grade = 1.958 + 0.627 × Formative Quiz Average. The regression coefficient was statistically significant ($p < 0.001$). This association indicates that students with higher formative quiz performance tended to obtain higher final course grades (Figure 1B). However, the analysis was exploratory and associative. Both

variables were embedded within the same course and may reflect common underlying factors such as academic preparation, engagement, attendance, and mastery of course content. The regression therefore should not be interpreted as evidence that higher quiz performance caused higher final grades.

3.5 Exploratory cluster analysis

K-means clustering, supported by the elbow criterion and silhouette index (0.52), identified three student performance profiles: Higher Performance (n = 10, 21.7%), Intermediate Performance (n = 20, 43.5%), and Below Performance (n = 16, 34.8%), with the latter defined as students scoring below the passing threshold (Figure 2). The Higher Performance profile showed the highest indicators (midterm M = 5.9, quiz M = 5.2, final M = 5.9), with a pre–post increase of 0.5 points (5.6 to 6.1). The Intermediate Performance profile, the largest group (midterm M = 4.0, quiz M = 4.1, final M = 4.7), showed the largest improvement (+0.8 points, from 3.6 to 4.4). The Below Performance profile (midterm M = 2.9, quiz M = 3.3, final M = 3.4) remained essentially unchanged (2.9 to 2.9) (Table 1). Thus, the overall pre–post improvement was heterogeneous, with the largest gains in the Intermediate group and no measurable change in the Below Performance group.

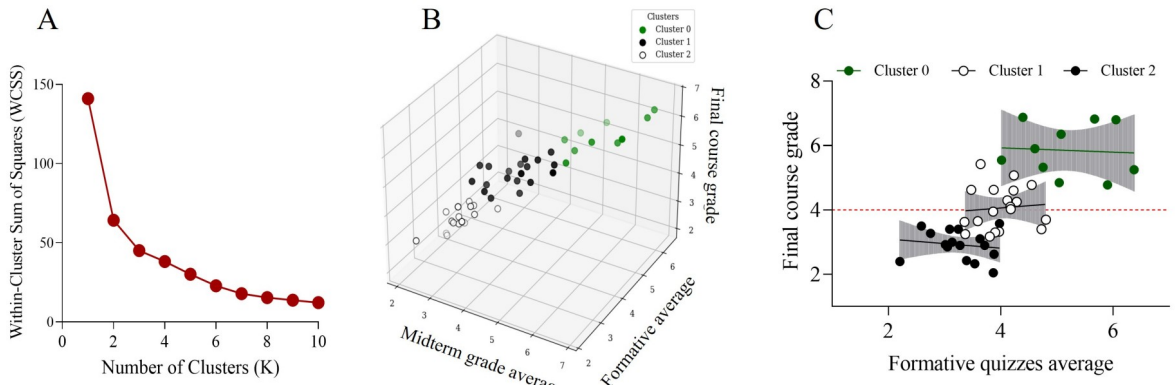


Figure 2. Exploratory cluster analysis.

3.6 Differences between performance profiles

One-way ANOVAs demonstrated substantial differences among the three profiles in midterm average, formative quiz average, and final course grade. Mean midterm average differed significantly across clusters, $F(2,43) = 75.82$, $p < 0.001$. Formative quiz average also differed significantly, $F(2,43) = 20.98$, $p < 0.001$. Final course grade showed significant differences across profiles, $F(2,43) = 77.16$, $p < 0.001$. Tukey HSD comparisons indicated significant pairwise differences between all three profiles for the academic variables examined ($p < 0.05$) (Table 3). These analyses characterize the academic differences represented by the clustering solution. Because the variables were also involved in defining the student profiles, these ANOVAs should not be interpreted as independent confirmation of the clustering solution.

Table 3. Cluster Differences in Midterm Averages, Formative Averages, and Final Grades.

Variable	Cluster 0	Cluster 1	Cluster 2	F(2, 43)	p-value	R^2
Midterm average	5.9	4.0	2.9	75.82	<0.001	0.775
Formative average	5.2	4.1	3.3	20.98	<0.001	0.488
Final course grade	5.9	4.7	3.4	77.16	<0.001	0.778

Note. Values represent cluster means. Cluster 0 corresponds to the Higher Performance group, Cluster 1 to the Intermediate Performance group, and Cluster 2 to the Below Perfomance group. F-values correspond to one-

way ANOVAs with 2 and 43 degrees of freedom. All pairwise cluster differences were significant according to Tukey HSD post hoc tests ($p < 0.05$).

4. Discussion

This study examined changes in academic performance following the implementation of post-class formative retention quizzes in a first-year undergraduate course and explored whether students exhibited heterogeneous performance trajectories. Three findings are particularly relevant. First, mean midterm performance increased from the pre-implementation period to the post-implementation period, with a mean improvement of 0.45 points and a moderate standardized effect size (Cohen's d = 0.45). Second, formative quiz performance was positively associated with final course grade (r = 0.60). Third, the exploratory cluster analysis indicated that the observed improvement was not uniform across students, with the largest descriptive increase occurring in the Intermediate Performance profile (+0.8 points), whereas the Below Performance group showed virtually no change.

4.1 Overall change in academic performance

The observed increase in midterm performance is broadly consistent with evidence supporting retrieval-based learning and formative assessment. Retrieval practice has been shown to benefit learning in applied educational settings, although the magnitude of these benefits varies according to implementation characteristics (7). Similarly, evidence from higher education indicates that low-stakes quizzing can be a promising formative assessment strategy, particularly when assessment is integrated into instruction and accompanied by feedback (3,8). However, the beneficial effects of retrieval practice without feedback may depend on question format; short-answer questions have been shown to enhance retention more than restudying, whereas multiple-choice questions do not confer the same advantage in authentic educational settings (13). Given that our quizzes used a multiple-choice format and feedback was provided immediately, this may have influenced the magnitude and pattern of the observed improvement. The present intervention incorporated these elements by administering brief quizzes immediately after instruction and providing immediate feedback during the subsequent session. Repeated retrieval may have provided students with opportunities to actively reconstruct recently learned information, while immediate feedback may have helped identify misconceptions and areas requiring further study. These mechanisms are consistent with established accounts of retrieval-based learning and formative feedback (5,21). The observed increase might be partially explained by the motivational benefits of receiving immediate feedback, as students often lose engagement when feedback is delayed beyond a certain point (11). Additionally, the structured nature of the feedback—provided during the subsequent session with item-by-item correction and study recommendations—may have contributed to the improvement, as previous research in medical education has demonstrated that systematic feedback delivery enhances student skill acquisition and performance in subsequent assessments (12).

However, the observed improvement cannot be attributed specifically to the quizzes. The study used a single-cohort pre-post design without a concurrent control group, and other factors may have contributed to the increase, including differences in course content or examination difficulty, greater familiarity with the course, changes in study behaviour, or other temporal effects. The four midterm examinations shared a common structure (3 development questions and 30–31 multiple-choice questions, 90 minutes, 42–46 total points) but covered different content areas: Midterms 1–2 assessed cellular and extracellular structures, while Midterms 3–4 assessed molecular biology and genetics. No formal psychometric analysis was conducted to equate their difficulty. This is a potential confounding variable that could have influenced the pre–post comparison. Moreover, retrieval practice and feedback were implemented together, preventing their individual contributions from being separated. The most defensible interpretation is therefore that academic performance increased during the period in which the quizzes were implemented, rather than that the quizzes caused the observed improvement.

4.2 Formative quiz performance and final course achievement

Formative quiz performance showed a moderate positive association with final course grade (r = 0.60), with the corresponding regression model accounting for 36.5% of the observed variance in final grades. This finding suggests that performance on low-stakes formative assessments may provide useful information about students' developing mastery of course content (22). Nevertheless, this association should not be interpreted as evidence that quiz performance independently determines final achievement. Both variables were obtained within the same course and are likely influenced by overlapping factors, including prior academic preparation, engagement, attendance, and general mastery of the subject. Thus, the regression is best interpreted as an exploratory characterization of the relationship between formative assessment performance and overall academic achievement rather than as evidence of a causal or independent predictive effect. In an exploratory multiple linear regression that included prior performance (mean of Midterms 1 and 2) as a predictor, prior performance was significantly associated with final course grade (B = 0.52, $p < 0.001$), and the formative quiz average remained a significant predictor (B = 0.34, $p = 0.012$), although the strength of the association was attenuated compared with the simple linear regression. The model accounted for 52.3% of the variance in final grades ($R^2 = 0.523$). This suggests that the association between quiz performance and final grade is partially explained by prior academic performance, reinforcing the need for cautious interpretation.

4.3 Heterogeneity in student performance trajectories

The exploratory cluster analysis provides an important complement to the overall pre–post comparison. Three performance profiles were identified: Higher Performance, Intermediate Performance, and Below Performance. The largest descriptive improvement occurred among students in the Intermediate Performance profile, whose mean midterm performance increased from 3.6 to 4.4. In contrast, students in the Higher Performance profile increased from 5.6 to 6.1, while the Below Performance profile remained essentially unchanged at approximately 2.9. The relatively large improvement among students in the Intermediate Performance profile may indicate that students with intermediate baseline performance had greater opportunity to benefit from repeated retrieval and feedback. However, this interpretation remains speculative because the study did not measure students' use of feedback, study behaviour, prior knowledge, or engagement with the quizzes. The smaller change among high-achieving students may partly reflect a ceiling effect, given their relatively high baseline performance. The lack of meaningful change among students in the Below Performance profile is particularly relevant. Their lack of improvement suggests that brief retrieval activities embedded within regular instruction may not be sufficient, by themselves, for students experiencing more substantial academic difficulties. Rather than indicating that retrieval practice is ineffective for these students, the finding suggests that formative quizzes may need to be complemented by more intensive academic support, such as targeted tutoring, individualized feedback, or structured study guidance. This interpretation is consistent with the broader view that formative assessment is most useful when assessment information is translated into subsequent learning or instructional actions (1,2). The heterogeneous performance trajectories observed in our study highlight a critical point emphasized in the literature: formative feedback strategies must be adapted to individual student needs to ensure equity and effectiveness (4). Our findings suggest that a one-size-fits-all approach may leave students "Below Performance" behind.
Importantly, these cluster-based findings should be regarded as exploratory. The sample was small, the profiles were derived from academic indicators within the same cohort, and the study was not designed to estimate differential intervention effects across student groups. The observed differences should not be interpreted as evidence that the quizzes caused the improvement, nor that the strategy was more effective for the Intermediate group and ineffective for the Below Performance group. Rather, these observations generate educational hypotheses to be confirmed in future studies. Moreover, the clustering solution was derived from three variables (midterm

average, formative quiz average, and final course grade) that were also used to describe the resulting profiles. Therefore, the one-way ANOVAs comparing clusters should be interpreted descriptively rather than as independent validation of the clustering solution.

4.4 Educational implications

The findings suggest three potentially relevant implications for first-year higher education. First, brief post-class retrieval quizzes can be embedded within routine teaching and combined with feedback without requiring a separate academic support programme. Second, examining student-level performance patterns may provide information that is not visible in cohort-level mean comparisons. The apparent improvement among students in the Intermediate Performance profile suggests that formative retrieval activities may be particularly relevant for students who are approaching the passing threshold but have not yet achieved consistent mastery. Conversely, the lack of improvement among students in the Below Performance profile suggests that formative quizzes may also function as an early identification mechanism, indicating when additional academic support may be required. Our findings that students in the "Intermediate Performance " group showed the most improvement align with research suggesting that experiencing formative assessment during university training can be particularly effective, as it helps students reflect on assessment complexity and motivates them to use diverse methods in their future professional roles (16). Third, the variability in outcomes across student profiles highlights the importance of not only providing feedback but also ensuring it is delivered in a structured and actionable manner; training faculty in systematic feedback protocols may enhance the effectiveness of formative interventions, particularly for students who are struggling (12). These implications should be considered provisional because the present study did not experimentally compare different support strategies or determine whether students actually used the feedback provided.

From the perspective of health sciences education, these findings suggest that brief retrieval quizzes may be particularly useful for students approaching the passing threshold, while students with more substantial academic difficulties may require additional targeted support. This has implications for the design of formative assessment strategies in medical and health professions curricula.

4.5 Limitations and future research

Several limitations should be considered. First, the retrospective single-cohort pre–post design and absence of a concurrent control group prevent causal inference. The evidence on the effectiveness of the intervention is therefore weak and should be interpreted as a description of temporal changes, not as a causal effect**.** Second, the final analytical sample was small (n = 46), limiting statistical precision and making the sex-stratified and cluster analyses exploratory. Third, five of the 51 initially enrolled students did not complete the course and were excluded from the final analysis. Their mean Midterm 1 score was lower than that of completers (3.2 vs. 3.97), suggesting a possible selection bias. Reasons for non-completion were not available in the academic records. Fourth, missing observations occurred for individual formative quizzes because students did not attend the corresponding session, experienced connectivity problems, or chose not to complete the assessment. These mechanisms may be related to student engagement, and missingness therefore cannot necessarily be assumed to be random. No imputation was performed. Fifth, although we provided immediate feedback following each quiz, we did not assess or standardize the quality and structure of the feedback delivered, unlike previous studies that have trained faculty in structured feedback protocols (12). This lack of standardization may have contributed to the heterogeneous outcomes observed across student profiles. Sixth, our quizzes used a multiple-choice format, which prior research suggests may be less effective for promoting retrieval-based learning than short-answer questions when feedback is not provided immediately (13). The use of multiple-choice questions may therefore have attenuated the potential benefits of

the retrieval practice intervention, particularly for students in the Below Performance profile who may have required more effortful retrieval to consolidate learning. Seventh, we did not assess attendance, use of feedback, motivation, or prior academic preparation, which limits the interpretability and transferability of the results. Finally, the intervention combined retrieval practice and feedback, preventing the independent contribution of each component from being determined. The cluster analysis should also be considered exploratory given the small sample and the use of overlapping academic indicators in cluster formation and subsequent characterization.

Future studies should use controlled longitudinal or quasi-experimental designs with larger samples and include measures of attendance, prior academic preparation, engagement, use of feedback, and study behaviour. Such designs would help determine whether the observed improvement is attributable to retrieval-based formative assessment and whether students with different baseline performance levels respond differently to the intervention. In particular, future research should examine whether combining retrieval quizzes with targeted academic support produces greater benefits for students who remain below the course passing threshold.

Funding: The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflict of interest: The authors declare that they have no conflict of interest.

Author contributions: MP, GUC contributed to the conception and design of the study, and to the acquisition and analysis of the data. MP wrote the manuscript, and GUC contributed to proofread the final version of the manuscript. All authors contributed to the article and approved the submitted version.

5. References

1. Black P, Wiliam D. Assessment and Classroom Learning. *Assess Educ Princ Policy Pract*. **1998**, 5, 7–74. https://doi.org/10.1080/0969595980050102
2. Nicol DJ, Macfarlane-Dick D. Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Stud High Educ*. **2006**, 31, 199–218. https://doi.org/10.1080/03075070600572090.
3. Morris R, Perry T, Wardle L. Formative assessment and feedback for learning in higher education: A systematic review. *Rev Educ*. **2021**, 9, e3292. https://doi.org/10.1002/rev3.3292.
4. Solis-Trujillo BP, Velarde-Camaqui D, Gonzales Nuñez CA, et al. The current landscape of formative assessment and feedback in graduate studies: a systematic literature review. *Front Educ*. **2025**, 10, 1509983. https://doi.org/10.3389/feduc.2025.1509983.
5. Karpicke JD, Roediger HL. The Critical Importance of Retrieval for Learning. *Science*. **2008**, 319, 966–968. https://doi.org/10.1126/science.1152408.
6. Roediger HL, Karpicke JD. The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspect Psychol Sci*. **2006**, 1, 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x.
7. Agarwal PK, Nunes LD, Blunt JR. Retrieval Practice Consistently Benefits Student Learning: A Systematic Review of Applied Research in Schools and Classrooms. *Educ Psychol Rev*. **2021**, 33, 1409–1453. https://doi.org/10.1007/s10648-021-09595-9.
8. Yang C, Luo L, Vadillo MA, et al. Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychol Bull*. **2021**, 147, 399–435. https://doi.org/10.1037/bul0000309.
9. Serra MJ, Kaminske AN, Nebel C, et al. The Use of Retrieval Practice in the Health Professions: A State-of-the-Art Review. Behav Sci. 2025, 15, 974. https://doi.org/10.3390/bs15070974.
10. Yang BW, Razo J, Persky AM. Using Testing as a Learning Tool. *Am J Pharm Educ*. **2019**, 83, 7324. https://doi.org/10.5688/ajpe7324.
11. Fisher DP, Brotto G, Lim I, et al. The Impact of Timely Formative Feedback on University Student Motivation. *Assess Eval High Educ*. **2025**, 50, 622–631.

https://doi.org/10.1080/02602938.2025.2449891.
12. Shinde SA, Joshi UU, Karanje NC, et al. Effectiveness of Structured Feedback Training on Student's Performance in Formative Assessment of Physiology Practicals: A Comparative Study. *Cureus*. **2025**, 17, e80176. https://doi.org/10.7759/cureus.80176.
13. Greving S, Richter T. Practicing retrieval in university teaching: short-answer questions are beneficial, whereas multiple-choice questions are not. *J Cogn Psychol*. **2022**, 34, 657–674. https://doi.org/10.1080/20445911.2022.2085281.
14. Bates LE, Myers SJ, DeLosh EL, et al. The Learning Benefits of Mastery Quizzing Versus Other Quizzing Methods. *Psychol Lear Teach*. **2024**, 24, 74–91. https://doi.org/10.1177/14757257241301733.
15. Roelle J, Schweppe J, Endres T, et al. Combining Retrieval Practice and Generative Learning in Educational Contexts: Promises and Challenges. *J Dev Psychol Educ Psychol*. **2022**, 54, 142–150. https://doi.org/10.1026/0049-8637/a000261.
16. Parmigiani D, Nicchia E, Murgia E, et al. Formative assessment in higher education: an exploratory study within programs for professionals in education. *Front Educ*. **2024**, 9, 1366215. https://doi.org/10.3389/feduc.2024.1366215.
17. Bulut O, Gorgun G, Yildirim-Erbasli SN. The impact of frequency and stakes of formative assessment on student achievement in higher education: A learning analytics study. *Comput Assist Learn*. **2025**, 41, e13087. https://doi.org/10.1111/jcal.13087.
18. Tomlinson A, Simpson A, Killingback C. Student expectations of teaching and learning when starting university: A systematic review. *J Furth High Educ*. **2023**, 47, 1054–1073. https://doi.org/10.1080/0309877X.2023.2212242.
19. Sirianansopa K. Evaluating students' learning achievements using the formative assessment technique: a retrospective study. *BMC Med Educ*. **2024**, 24, 1373. https://doi.org/10.1186/s12909-024-06347-5.
20. Wu J, Yu X. The influence of formative assessment on academic performance,8exploring the role of teachers' emotional support. *Front Psychol*. **2025**, 16, 1567615. https://doi.org/10.3389/fpsyg.2025.1567615.
21. Hattie J, Timperley H. The Power of Feedback. *Rev Educ Res*. **2007**, 77, 31. https://doi.org/10.3102/003465430298487
22. Pérez MÁ, Urrejola-Contreras G, Tiemann P. Low-Stakes Quizzing as a Predictor of Academic Outcomes: The Role of Motivation in Undergraduate Health Sciences Education. *Rev Esp Edu Med*. **2025**, 6, 10. https://doi.org/10.6018/edumed.680941.

© 2026 University of Murcia. Submitted for open access publication under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 Spain License (CC BY-NC-ND) (http://creativecommons.org/licenses/by/4.0/)