Educational Assessment, Evaluation and Accountability
https://doi.org/10.1007/s11092-026-09506-y

Tracking calibration and confidence: continuous and group-based insights from two high-stakes exams

Samuel P. Leon[1] · **Anastasiya Lipnevich**[2]

Received: 1 December 2025 / Accepted: 9 July 2026

Abstract

Accurate self-assessment, or calibration, is a key component of students' metacognitive monitoring and self-regulated learning. This study examined calibration and prediction confidence across two high-stakes course exams (initial and make-up) using both continuous and group-based analyses. Undergraduate students in early childhood education completed Exam 1 ($n = 128$), predicting their grade (0–10) and reporting confidence (1–5). Students who failed ($n = 45$) took a make-up exam one month later and repeated their self-assessment. In Exam 1, calibration followed a pattern consistent with the Dunning–Kruger Effect (DKE): low-performing students overestimated their performance, mid-range students were most accurate, and high-performing students slightly underestimated their results. Among students who took the make-up exam, absolute calibration accuracy improved and confidence increased, and these changes remained robust after controlling for regression-to-the-mean effects. Together, the findings show that performance-level calibration patterns consistent with the DKE persist under authentic assessment conditions, while absolute miscalibration and confidence levels can shift following feedback and repeated testing. These results highlight adaptive shifts in metacognitive monitoring accuracy across repeated high-stakes assessments and provide practical implications for supporting self-assessment in higher education.

Keywords Calibration · Self-assessment · Confidence judgment · Exam performance, actual performance, predicted performance

✉ Anastasiya Lipnevich
alipnevich@nbme.org

Samuel P. Leon
sparra@ujaen.es

[1] University of Jaén, Jaén, Spain

[2] National Board of Medical Examiners, Philadelphia, USA

1 Introduction

Metacognitive monitoring, or the capacity to evaluate one's own knowledge and performance, plays a central role in self-regulated learning (SRL) (Nelson, 1990; Zimmerman, 2013). Within this framework, self-assessment occupies a pivotal position as both a reflective and formative process, fostering a deeper understanding of one's strengths and areas for improvement (DeLuca, 2025; Lew et al., 2009; Yan et al., 2023). When students evaluate their own performance, they gain a sense of ownership and responsibility, learn to set realistic goals, and plan their learning with greater autonomy (Brown & Harris, 2013, 2014; Kostons et al., 2012). Beyond these immediate benefits, self-assessment is linked to broader educational and life outcomes through its contribution to the development of effective self-regulated learners (Aryadoust, 2015; Dunning et al., 2004; Mendoza & Yan, 2023).

Despite its recognized benefits, mastering accurate self-assessment is challenging (Raaijmakers et al., 2019). Cognitive biases such as overconfidence, anchoring, or the Dunning–Kruger effect often distort students' judgments (Dunning et al., 2004; León et al., 2023; Lipnevich et al., 2023). Remarkably, people tend to recognize such biases in others but remain unaware of distortions in their own evaluations (Pronin, 2008; Pronin & Hazel, 2023). As a result, students' self-assessments can diverge substantially from actual performance, undermining their formative potential. Research on metacognitive self-monitoring therefore seeks to understand not only whether learners can assess themselves accurately but also why accuracy varies across contexts and individuals (Ehrlinger et al., 2008).

Calibration, operationalized as the discrepancy between predicted and actual performance, captures the accuracy of self-assessment (Bol et al., 2005; Miller & Geraci, 2019; Oluk & Korkmaz, 2024). The smaller the difference, the better calibrated the student, with calibration being identified as a core metacognitive skill associated with learning outcomes, study regulation, and academic achievement (Bol & Hacker, 2012; Efklides & Vlachopoulos, 2012; Nietfeld et al., 2006; Zimmerman et al., 2011).

A substantial body of research has examined the relationship between calibration and academic performance using experimental, cross-sectional, longitudinal, and quasi-experimental designs (e.g., Bol & Hacker, 2012; Rinne & Mazzocco, 2014). These studies provide important evidence that monitoring accuracy is meaningfully associated with achievement. However, much of this work has been conducted in laboratory settings, low- or moderate-stakes classroom assessments, or structured intervention contexts. Consequently, less is known about whether these associations and performance-level patterns replicate and evolve in authentic high-stakes university exams, where judgments carry direct academic consequences.

Findings from laboratory or low-stakes classroom assessments may not fully generalize to high-stakes examination contexts, as evaluative stakes can shape motivation, affect, and regulatory effort, thereby influencing metacognitive monitoring. For example, anticipating a graded examination rather than a pass-fail examination has been shown to increase confidence and produce greater improvements in monitoring accuracy over time (Barenberg & Dutke, 2013). Similarly, classroom research indicates that retrieval practice and feedback can enhance monitoring accuracy and reduce overestimation (Butler et al., 2008; Cogliano et al., 2019). At the same time,

repeated testing does not automatically eliminate overestimation biases, as overconfidence may persist even after multiple exam experiences (Foster et al., 2017). Together, these findings highlight the need to examine whether performance-level calibration patterns and changes replicate in authentic high-stakes university exams.

An additional layer of complexity comes from confidence judgments, defined as students' ratings of how certain they are in their performance predictions (Blais et al., 2005; Chen & Bonner, 2019; León et al., 2021). Confidence judgments are valuable because they capture the certainty learners assign to their estimates and can reveal the strength of their metacognitive monitoring. Following traditions in experimental psychology, where participants are often asked not only to predict an outcome but also to rate confidence in that prediction Dunlosky & Metcalfe, 2009; León et al., 2011), we included confidence judgments to disambiguate self-assessment. Asking both what grade students expect to get and how confident they are, helps to distinguish aspirational responses from genuine monitoring judgments. Importantly, calibration reflects the discrepancy between predicted and actual performance, whereas confidence captures the certainty attached to that prediction. Learners at different achievement levels may therefore show systematic discrepancies between predicted and actual outcomes, highlighting the complexity of monitoring accuracy (Eriksson et al., 2024; Hacker et al., 2008; Stankov & Lee, 2014). Understanding this divergence is key to comprehending the dynamics of monitoring accuracy.

In the broader context of SRL, calibration accuracy is not static. Experience and feedback may play critical roles in refining it. For instance, in their meta-analysis León et al. (2023) demonstrated that feedback and prior self-assessment experience can improve accuracy of self-assessment, yet they also highlighted the shortage of robust studies conducted in authentic educational contexts. Longitudinal research suggests that calibration predicts later achievement independent of prior performance (Rinne & Mazzocco, 2014), yet most studies examine calibration at single time points or across long intervals (Oluk & Korkmaz, 2024; Yan et al., 2023). Consequently, within-student changes in calibration across repeated assessments remain insufficiently understood in authentic educational contexts, particularly when examinations are separated by weeks rather than examined at daily time scales. Recent theoretical perspectives suggest that metacognitive monitoring may exhibit a degree of short-term adaptability in response to performance feedback and evaluative experiences (Efklides, 2011; Metcalfe, 2017). Empirical syntheses further indicate that feedback and prior self-assessment experience can enhance monitoring accuracy, particularly when learners are confronted with discrepancies between expected and actual performance (León et al., 2023). In this study, we refer to this potential adaptability as *metacognitive plasticity*, defined here as the capacity to adjust monitoring accuracy over relatively short time intervals following performance discrepancies. Rather than conceptualizing calibration as a fixed trait, this perspective frames it as a dynamic process shaped by feedback, task demands, and reflective experience within authentic learning contexts. From an error-driven learning perspective, discrepancies between expected and actual performance generate cognitive conflict that may trigger adjustments in monitoring processes. Within self-regulated learning models, such discrepancies activate feedback loops linking monitoring and control processes (Efklides, 2011; Metcalfe, 2017). These mechanisms provide a theoretical basis for

understanding short-term recalibration in authentic assessment contexts. The present study addresses this gap by examining calibration and confidence in a naturalistic, high-stakes course context with two consecutive exams, providing insight into the short-term dynamics of monitoring accuracy.

Cross-sectional research has repeatedly documented a robust performance-based calibration gradient, commonly associated with the Dunning-Kruger Effect (Kruger & Dunning, 1999). In this pattern, lower-performing students tend to overestimate their performance, mid-performing students show relatively accurate judgments, and higher-performing students slightly underestimate their results (León et al., 2021; Miller & Geraci, 2019; Sheldrake et al., 2022).

Although this gradient has been observed across laboratory and classroom contexts, important questions remain regarding its generalization to authentic high-stakes university exams, the evolution of calibration and confidence across repeated assessments, and the extent to which observed changes reflect genuine adjustments in monitoring accuracy rather than statistical artifacts such as regression to the mean (Barnett & Ceci, 2002). The present study addresses these issues by examining calibration and confidence across two consecutive exams within a naturalistic course setting.

2 Research questions

RQ1. How do actual exam score and confidence judgments relate to calibration in authentic university exams with academic consequences?
RQ2. How do calibration and confidence judgments evolve when students who fail initially face a make-up exam?
RQ3. Do low-performing students overestimate, mid-performing students demonstrate accuracy, and high-performing students underestimate their performance, consistent with classic calibration patterns?
RQ4. Do observed changes persist once potential statistical artefacts (e.g., regression to the mean, ceiling/floor effects) are considered?

By combining theoretical and methodological insights from both educational and experimental metacognition research, this study examines how calibration and confidence change across repeated, high-stakes assessments in an authentic university course. The findings clarify whether self-assessment accuracy can improve over short intervals and help to distinguish metacognitive development from statistical artifacts, thereby informing evidence-based approaches to support students' monitoring accuracy in real educational settings.

2.1 Methods

2.2 Participants and context

Participants were 128 undergraduates enrolled in an Early Childhood Education course at a Spanish public university (M_{age} = 22.4 years, SD = 3.45; 86.7% female). This gender distribution is consistent with national enrollment patterns in education programs in Spain (Spanish National Statistics Institute, 2020). A subset of 45 students completed a required make-up exam approximately one month later (M_{age} = 22.8 years, SD = 4.26; 82.2% female). Of these, 38 had also taken the first exam, whereas 7 were first-time takers who had not attended the initial exam. Analyses of change therefore uses the subset of 38 students with data for both exams.

Students were required to complete the exams as part of their course requirements, but participation in the study (i.e., completing self-assessment measures) was voluntary and did not affect their grades. Students could opt out of data use at any time without penalty. Although the overall sample size (n = 128) provided adequate power (> 0.80) to detect medium-sized effects ($f \approx 0.25$) at $\alpha = 0.05$, the smaller subsample of repeaters (n = 38) offered only moderate power for small effects. Therefore, results regarding within-student changes should be interpreted with caution. Power estimates were obtained using G*Power 3.1 (Faul et al., 2007).

2.3 Exams and scoring

The exam analyzed was the regularly scheduled end-of-term examination for this course and followed the standard item-bank procedure routinely applied in prior years. Each exam included 20 three-option multiple-choice questions covering course content. To correct for guessing, a standard correction-for-guessing procedure was applied, whereby incorrect responses were penalized. Each correct response received + 0.5 points, each incorrect response − 0.25 points, and blank responses 0 points. With 20 items, the maximum possible score was 10. Because incorrect responses were penalized, total scores could take negative values (theoretical range − 5 to + 10). Negative values were retained and were not truncated. This instructor-graded score constituted the Actual Exam Score (AES) used in all analyses.

Exam forms were assembled from an instructor-authored item bank of approximately 200 three-option items spanning six course topics. The number of items drawn from each topic was proportional to its weight in the course syllabus, ensuring content representativeness across forms. For each administration, four parallel forms (a, b, c, d) were generated by randomly sampling the required number of items from each topic-specific pool without replacement within forms. Exam 1 and Exam 2 forms were generated independently using the same proportional procedure. Although formal equating analyses were not conducted, all forms were constructed following identical content specifications and sampling rules. Reliability estimates were comparable across administrations (λ_2 = 0.81 and 0.83), supporting reasonable consistency across exam versions.

2.4 Measures and study variables

Actual exam score (AES) represented the instructor-graded score on a 0–10 scale. Predicted exam score (PES) reflected each student's self-predicted grade immediately after completing the exam, using the same 0–10 scale. In metacognitive terms, this measure constitutes a global assessment of performance, as students estimated their overall exam outcome after responding to all items rather than providing item-level or prospective judgments (Nelson, 1990; Dunlosky & Metcalfe, 2009). Confidence Judgment (CJ) captured the student's confidence in that prediction, rated on a 1–5 scale (1 = *not at all confident*, 5 = *completely confident*). Calibration was computed as PES − AES, with positive values indicating overestimation, near-zero values accuracy, and negative values underestimation. For descriptive purposes, performance levels were also grouped as Fail (< 5), Low pass (5–<7.5), and High pass (≥ 7.5), but all primary analyses treated performance as a continuous variable.

2.5 Procedure

At the start of each exam, the instructor explained the procedure and scoring rules (+ 0.5 for correct, − 0.25 for incorrect, 0 for blank), and informed students that after answering they would be asked to record a predicted grade (0–10) and their confidence (1–5). Immediately after submitting their responses, students reported their predicted exam score and a confidence judgment on the last page of the exam. After grading, scores were posted on the learning management system, together with the answer key and scoring rules, ensuring baseline feedback access for all students. Optional exam review meetings were offered, so that students could examine their graded exams and discuss remaining questions with the instructor. Attendance at these sessions was not logged and individual feedback uptake was therefore unmeasured. Based on typical trends in this course, students who failed the first exam were more likely to attend review sessions. Approximately one month later students who had failed the first exam were required to take the make-up exam. All procedures complied with the Declaration of Helsinki (World Medical Association, 2013) and were approved by the [University of Jaen Ethics Committee (JUN.22/3.PRY).

2.6 Analysis plan

Analyses proceeded in several steps. First, descriptive statistics and correlations were computed for the actual exam score, predicted exam score, confidence judgment, and calibration in each exam. Second, multiple linear regression analyses were conducted separately for each exam to examine whether calibration (PES − AES) could be predicted from Actual Exam Score (AES) and Confidence Judgment (CJ), with both predictors entered simultaneously. To facilitate comparison of effect sizes, standardized regression coefficients (β) were obtained by fitting models to z-standardized variables. Multicollinearity diagnostics were examined using variance inflation factors (VIF), which were below conventional thresholds, indicating no problematic collinearity. Model assumptions were inspected through visual examination of residual plots. Third, complementary group-based comparisons (fail, low pass, high

pass) provided an interpretable summary of how calibration varied systematically across performance levels. (León et al., 2021). Fourth, paired-sample analyses examined within-student changes in calibration and confidence judgment between the first and the make-up exam, accompanied by estimation plots and bootstrap confidence intervals (Raaijmakers et al., 2019). Finally, sensitivity checks addressed possible regression-to-the-mean artefacts using Oldham's correlation and related approaches (Barnett et al., 2005). The raw data and R script for the analyses, as well as the supplementary material, are available through open and public access in the project created for this purpose in the OSF repository (https://osf.io/drtj4/?view_only=673560c5a431421caa3ca88ada4dc1ad).

3 Results

3.1 RQ1. Relations among performance, confidence, and calibration

Table 1 presents the results of the descriptive statistics for each of the two tests, as well as the change in these statistics between exams.

In Exam 1 ($n = 128$), descriptive statistics are presented in Table 1. Correlations indicated that the actual exam score was strongly negatively related to calibration ($r = -.81$), such that lower performers overestimated more strongly. Confidence judgments were modestly correlated with predicted exam scores ($r = .21$), but not strongly associated with calibration in the regression model. A multiple regression model was estimated for Exam 1, with standardized actual exam score and confidence judgment entered simultaneously as predictors of calibration, actual exam score was a strong negative predictor of calibration ($\beta = -0.83$, 95% CI [−0.93, −0.72], $p < .001$), whereas confidence judgment was not statistically significant ($\beta = 0.09$, 95% CI [−0.02, 0.19], $p = .096$). The model explained 67% of the variance in calibration ($R^2 = 0.67$). Multicollinearity diagnostics indicated no evidence of collinearity between predictors (VIFs ≈ 1.02). Figure 1 illustrates the linear decline in overestimation as achievement increases, consistent with the classic calibration pattern reported in previous literature.

For Exam 2 (make-up exam) ($n = 45$), descriptive statistics are presented in Table 1. Correlations again showed a negative association between actual exam score and calibration ($r = -.69$), whereas confidence judgment was positively related to calibration ($r = .62$). A separate multiple regression model was estimated for Exam

Table 1 *Descriptive Statistics for Exam 1 and Exam 2*

	Exam 1			Exam 2					
Measure	M	SD	*n*	M	SD	*n*	ΔM	ΔSD	n_{paired}
AES	4.97	2.16	128	7.07	1.96	45	+2.10	2.04	38
PES	5.91	1.26	128	6.85	1.47	45	−0.38	1.52	38
CJ	2.62	0.76	128	3.33	0.80	45	+0.76	0.71	38
Calibration	+0.94	1.77	128	−0.21	1.77	45	−1.25	1.86	38

Note. *AES* Actual Exam Score, *PES* Predicted Exam Score, *CJ* Confidence Judgment, Calibration PES – AES, ΔM Change in the mean between Exam 1 and Exam 2, ΔSD Change in the Standard Deviation between Exam 1 and Exam 2

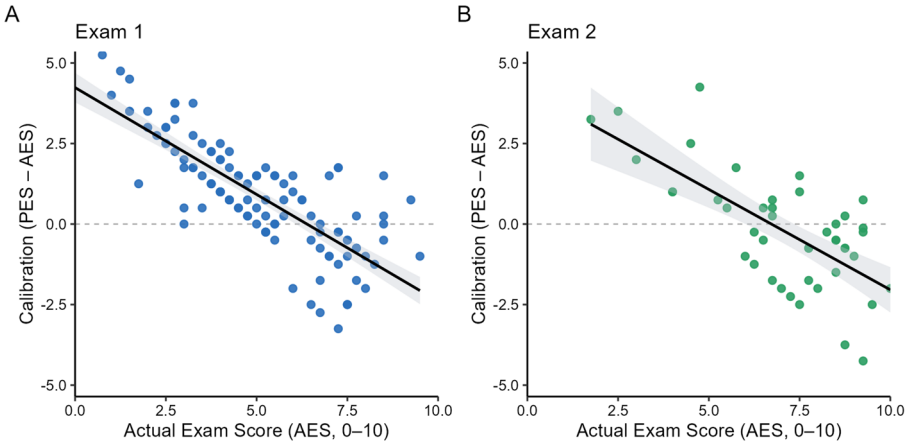

Fig. 1 Relation between actual exam performance and calibration across exams. Scatterplots illustrate the association between students' actual exam scores (AES) and their calibration (Predicted - Actual score) for Exam 1 and Exam 2. Negative slopes indicate that lower-performing students tended to overestimate, whereas higher-performing students slightly underestimated their performance. Shaded areas represent 95% confidence intervals

2 using the same specification, both actual exam score ($\beta = -0.84$, 95% CI [−1.04, −0.63], $p < .001$) and confidence judgment ($\beta = 0.42$, 95% CI [0.22, 0.62], $p < .001$) significantly predicted calibration. The model explained 63% of the variance ($R^2 = 0.63$), and multicollinearity diagnostics again indicated no concern (VIFs ≈ 1.13). These results indicate that although performance level remained strongly associated with calibration, confidence judgments also contributed significantly in the make-up exam. Overall, the findings for RQ1 indicate that the classic calibration pattern persisted across exams, while confidence judgments emerged as a significant predictor of calibration in the second sitting.

3.2 RQ2. Change across repeated exams

Analyses of students who took both exams ($n = 38$) revealed significant within-student changes. Calibration improved from Exam 1 ($M = +1.02$, $SD = 2.02$) to Exam 2 ($M = -0.23$, $SD = 1.76$), with a mean difference of 1.25 points, $t(37) = 2.95$, $p = .005$, Cohen's d = 0.48, 95% CI [0.39, 2.11]. Confidence also increased significantly between exams, with a mean difference of −0.76 points (reflecting higher confidence in Exam 2), $t(37) = -4.37$, $p < .001$, Cohen's d = -0.71, 95% CI [−1.12, −0.41]. (Fig. 2A) illustrates the improvement in calibration, showing that most students shifted from overestimation toward more accurate self-assessment on the second attempt, whereas (Fig. 2B) shows the concurrent rise in confidence. Together, these paired plots indicate that students who failed the first exam adjusted both their monitoring accuracy and their confidence after additional study and exposure to feedback. Overall, the within-student analyses confirm that calibration accuracy improved between exams while confidence also increased, suggesting adaptive adjustments in metacognitive monitoring across repeated assessments.

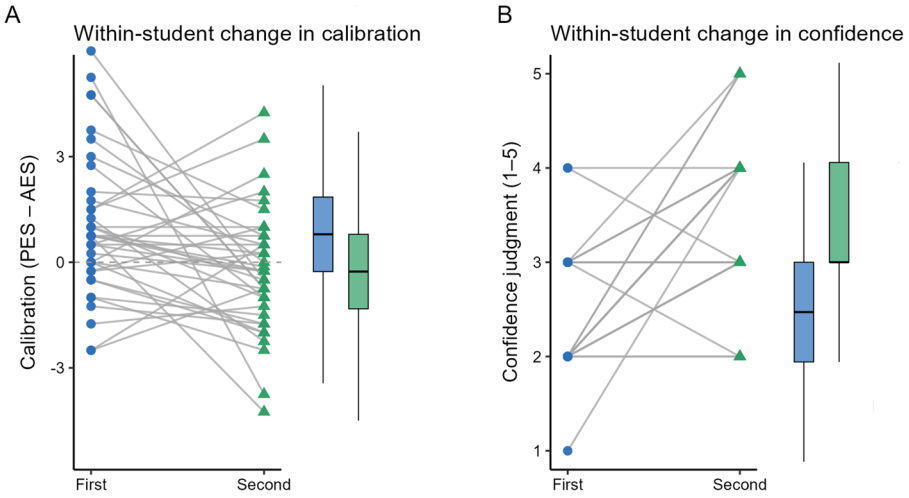

Fig. 2 Within-student changes in calibration and confidence judgments between exams. Panel **A** shows paired changes in calibration (Exam 1 → Exam 2), and Panel **B** shows corresponding changes in confidence judgments (CJ, 1–5 scale). Each line represents an individual student. Boxplots and density plots on the right summarize score distributions. Students generally shifted toward more accurate calibration and higher confidence on the second attempt

3.3 RQ3. Reproduction of classic patterns

Complementary group-based analyses descriptively reproduced the expected level of calibration. In Exam 1, failing students (actual exam score < 5) showed marked overestimation ($M = +2.17$), low passing students were close to accurate ($M = +0.11$), and high passing students slightly underestimated their performance ($M = -0.85$). In Exam 2, a similar pattern was observed, although with a smaller failing subgroup ($n = 7$): failing students continued to overestimate ($M = +2.50$), low passing students were close to accurate (M = -0.29), and high passing students underestimated (M = -0.96). Figure 3 presents raincloud plots with bootstrap-based confidence intervals, illustrating the distribution and 95% CIs of calibration for each performance group across exams. The descriptive pattern closely mirrored the continuous regression results, indicating that performance-level differences in calibration were consistent across sittings.

As an exploratory robustness check, we conducted non-parametric bootstrap resampling (5,000 iterations) to examine the stability of group-based estimates. The resampled confidence intervals were consistent with those obtained from the parametric analyses, supporting the robustness of the observed calibration pattern (see Supplementary Appendix A for detailed estimation plots and intervals).

3.4 RQ4. Robustness to artefacts

Finally, sensitivity analyses examined whether the observed improvements could be attributed to statistical artefacts rather than substantive change. Oldham's correlation between calibration change and the average of both exams was small and

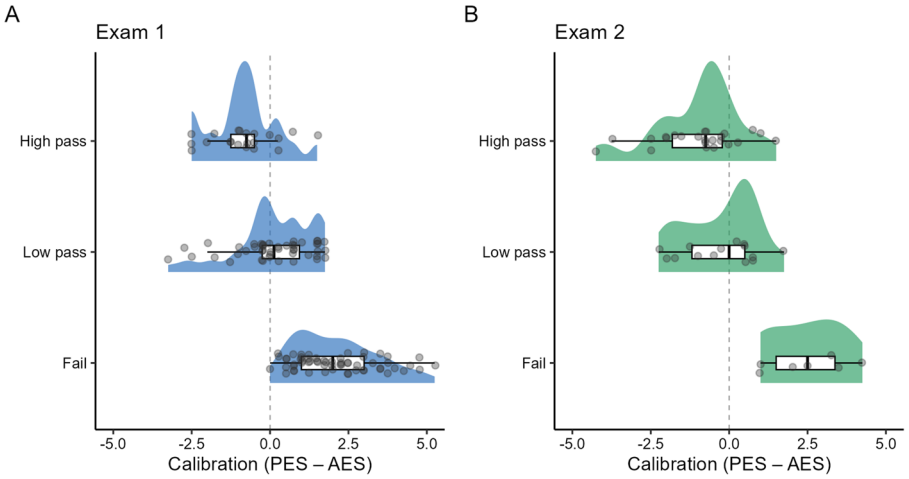

Fig. 3 Calibration pattern by performance group with bootstrap confidence intervals. Mean calibration (Predicted – Actual) is displayed for three performance groups (Fail, low pass, high pass) in Exam 1 and Exam 2. Ribbons depict 95% bootstrap confidence intervals based on 5,000 resamples. The classic pattern (i.e., overestimation among lower performers and underestimation among higher performers) persisted across both exams

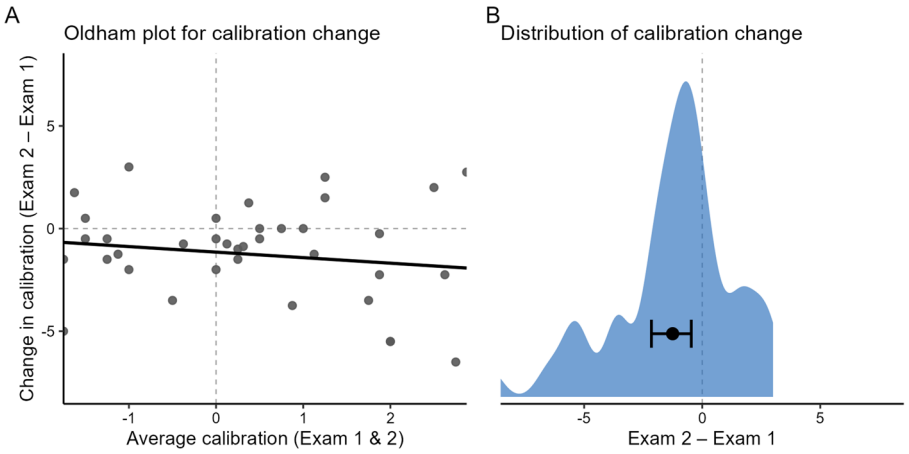

Fig. 4 Sensitivity analyses testing robustness to statistical artefacts. Panel **A**: Oldham plot showing the relationship between mean calibration (across exams) and change in calibration (Exam 2 – Exam 1), illustrating minimal regression-to-the-mean bias. Panel **B**: Distribution of calibration change scores with mean ± 95% CI. Both panels indicate that improvements in calibration reflect genuine metacognitive adjustment rather than artefactual effects

negative ($r = -.14$), suggesting that regression-to-the-mean effects were unlikely to fully account for the observed pattern. Additional regressions predicting calibration in Exam 2 from Exam 1 yielded consistent directional estimates, and standardizing calibration by baseline *SD* produced comparable results. Figure 4 illustrates these analyses: Panel A presents the Oldham plot, while Panel B shows the distribution of

individual calibration changes with mean and bootstrap confidence intervals. Taken together, these analyses indicate that the improvements in calibration and confidence are unlikely to be solely explained by statistical artefacts.

4 Discussion

In this study we examined students' calibration accuracy and confidence judgments in an authentic, high-stakes assessment context, focusing on two consecutive university exams. We aimed to examine whether the well-documented calibration pattern across performance levels would replicate under authentic course conditions and whether calibration accuracy would improve with repeated testing and feedback opportunities. The findings revealed three consistent patterns. First, calibration accuracy varied systematically with performance level, such that low performers overestimated, average-performing students were relatively accurate, and high performers slightly underestimated their performance. Second, calibration and confidence jointly improved among students who took the make-up exam, indicating adaptive monitoring over time. Third, these changes persisted even after accounting for potential statistical artefacts, suggesting genuine shifts in self-evaluation rather than regression-to-the-mean effects.

These results are largely consistent with prior work, but they also extend the Dunning–Kruger framework by emphasizing its ecological validity in high-stakes educational settings. Rather than simply replicating the pattern under laboratory or low-stakes conditions, this study demonstrated that the calibration pattern persisted even when students' performance had real academic consequences, offering a more context-sensitive understanding of metacognitive miscalibration. This pattern aligns with prior research demonstrating systematic, performance-based differences in calibration (Kruger & Dunning, 1999; León et al.,2021; Miller & Geraci, 2011). Across both experimental and classroom settings, lower-achieving students tend to overestimate their abilities, often due to limited metacognitive insight or knowledge deficits that hinder accurate self-evaluation. In contrast, high-achieving students may underestimate their performance reflecting elevated internal standards or conservative self-judgment tendencies (Stankov & Lee, 2014). Our findings extend these findings to naturalistic university exams, suggesting that even in consequential settings, where performance carries academic and emotional weight, the classic calibration pattern remains remarkably robust. However, not all studies have found such improvement or stability in calibration. For instance, Bol et al. (2012) and Dinsmore and Parkinson (2013) reported mixed or null effects of feedback and practice on calibration accuracy, suggesting that contextual and methodological factors may moderate outcomes.

The within-student improvement observed among students who took the make-up exam can also be interpreted within the frameworks of error-driven learning and self-regulated learning. (e.g., Zimmerman, 2000; Metcalfe, 2017). These models propose that discrepancies between expected and actual outcomes act as feedback signals that drive metacognitive adaptation. In this sense, the within-student improvement observed among students who took the make-up exam aligns with the notion of metacognitive plasticity introduced earlier, referring to learners' short-term capacity to

adjust monitoring accuracy in response to performance feedback, cognitive conflict, or affective cues. Such adaptability emerges through the interplay among feedback timing, task demands, and emotional factors that jointly foster recalibration of judgments and refinement of learning strategies. Experiencing failure on the first exam likely provided concrete feedback on prior misjudgments, whereas the opportunity to retake the exam functioned as a feed-forward mechanism, supporting subsequent monitoring and regulation processes.

This pattern resonates with research showing that repeated testing with feedback can foster recalibration and more accurate confidence judgments (Hacker et al., 2008; León et al., 2023; Yan et al., 2023). It also aligns with the notion that calibration is not merely a static trait but a skill that develops through experience, reflection, and exposure to evaluative feedback. Importantly, the concurrent increase in confidence and accuracy among repeaters suggests that students learned not only to *know more*, but to *know better what they knew*, which represents a core indicator of improved metacognitive monitoring.

A deeper look into recent models of metacognitive development further supports these interpretations. For example, Efklides' (2011) Metacognitive and Affective Model of Self-Regulated Learning posits that metacognitive monitoring and control processes are continuously shaped through feedback loops involving both cognitive and emotional components. In this light, the improvement observed in students who had to retake the exam can be seen as an adaptive response to dissonance between expected and actual outcomes. In other words, students who initially failed may have recalibrated not only their knowledge estimates but also their motivational orientation toward accuracy. Similarly, the improvement in confidence observed here mirrors findings from recent work on confidence–accuracy calibration (De Bruin & van Merriënboer, 2017), which emphasizes the reciprocal nature of monitoring and control: as learners refine their judgments, they also enhance their ability to allocate effort strategically.

From an educational perspective, these findings underscore the importance of designing assessment practices that support metacognitive development. Opportunities for self-assessment, combined with structured feedback, can help students calibrate their judgments more precisely (Van der Linden et al., 2023). Instructors can leverage exam reviews, formative assessments, or reflective prompts to encourage students to compare their predicted and actual performances, thereby transforming evaluation moments into learning opportunities (Arianto & Hanif, 2024; Teng et al., 2024). Embedding such reflective cycles in higher education may foster a culture of metacognitive awareness, promoting more self-regulated, confident, and accurate learners.

4.1 Limitations

Despite these promising insights, several limitations warrant attention. The study was conducted within a single course and discipline, with a predominantly female sample typical of Early Childhood Education programs in Spain. Although this homogeneity aids internal validity, it may constrain generalizability across disciplines or cultures. Additionally, the modest size of the make-up exam subgroup ($n = 38$) limits statisti-

cal power for detecting smaller effects, even though observed changes were consistent and theoretically meaningful. Although all students had access to feedback following the first exam (e.g., answer keys, scoring rules, and optional review sessions), individual feedback uptake was not systematically measured. Consequently, the extent to which recalibration was directly attributable to feedback engagement versus additional study or motivational factors cannot be fully determined. Future research should replicate these findings in diverse academic settings and incorporate direct indicators of feedback engagement (e.g., attendance logs, self-reports, or learning analytics) alongside longitudinal designs including multiple assessment cycles. Moreover, integrating qualitative measures (e.g., students' explanations of their judgments) could deepen insights into the mechanisms underlying calibration shifts.

5 Conclusion

Overall, this study not only reinforces the robustness of calibration patterns but also offers a bridge between theoretical insight and classroom practice. By integrating mechanisms such as error-driven learning, metacognitive plasticity, and self-regulated learning cycles, it extends recent models (e.g., Efklides, 2011; Metcalfe, 2017) and contributes to ongoing debates about how feedback timing and confidence monitoring shape learning outcomes. This study contributes to a growing body of evidence that calibration accuracy, though imperfect, can improve through structured assessment experiences. By situating calibration research within authentic educational practice, our findings highlight the complex links among cognition, motivation, and feedback in shaping students' self-evaluation. Encouraging learners to reflect on both *what* they know and *how well* they know it may be one of the most powerful, yet underused, routes toward deeper learning and academic resilience.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11092-026-09506-y.

Author contributions S.L. and A.L. wrote manuscript text, S.L. did data analysis. Both authors reviewed the manuscript.

Funding No funding was received for this study.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Arianto, A., & Hanif, M. (2024). Metacognitive and self-regulated strategies as predictors of primary school students' self-efficacy and problem-solving skills in science learning. *International Journal of Educational Research*, *115*, 102–118. https://files.eric.ed.gov/fulltext/EJ1441218.pdf

Aryadoust, V. (2015). Self-and peer assessments of oral presentations by first-year university students. *Educational Assessment*, *20*(3), 199–225. https://doi.org/10.1080/10627197.2015.1061989

Barenberg, J., & Dutke, S. (2013). Metacognitive monitoring in university classes: Anticipating a graded vs. a pass-fail test affects monitoring accuracy. *Metacognition and Learning*, *8*(2), 121–143. https://doi.org/10.1007/s11409-013-9098-3

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, *128*(4), 612–637. https://doi.org/10.1037/0033-2909.128.4.612

Barnett, A. G., Van Der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: what it is and how to deal with it. *International journal of epidemiology*, *34*(1), 215–220. https://doi.org/10.1093/ije/dyh299

Blais, A. R., Thompson, M. M., & Baranski, J. V. (2005). Individual differences in decision processing and confidence judgments in comparative judgment tasks: The role of cognitive styles. *Personality and Individual Differences*, *38*(7), 1701–1713. https://doi.org/10.1016/j.paid.2004.11.004

Bol, L., & Hacker, D. J. (2012). Calibration research: Where do we go from here? *Frontiers in psychology*, *3*, 229. https://doi.org/10.3389/fpsyg.2012.00229

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education*, *73*(4), 269–290. https://doi.org/10.3200/JEXE.73.4.269-290

Bol, L., Hacker, D. J., Walck, C. C., & Nunnery, J. A. (2012). The effects of individual or group guidelines on the calibration accuracy and achievement of high school biology students. *Contemporary Educational Psychology*, *37*(4), 280–287. https://doi.org/10.1016/j.cedpsych.2012.02.004

Brown, G. T. L., & Harris, L. R. (2013). Student self-assessment. In J. H. McMillan (Ed.), *The SAGE handbook of research on classroom assessment* (pp. 367–393). Sage.

Brown, G. T. L., & Harris, L. R. (2014). The future of self-assessment in classroom practice: reframing self-assessment as a core competency. *Frontline Learning Research*, *3*, 22–30. https://doi.org/10.14786/flr.v2i1.24

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning Memory and Cognition*, *34*(4), 918–928. https://doi.org/10.1037/0278-7393.34.4.918

Chen, P. P., & Bonner, S. M. (2019). A framework for classroom assessment, learning, and self-regulation. *Assessment in Education: Principles Policy & Practice*, *27*(4), 373–393. https://doi.org/10.1080/0969594X.2019.1619515

Cogliano, M. C., Kardash, C. A. M., & Bernacki, M. L. (2019). The effects of retrieval practice and prior topic knowledge on test performance and confidence judgments. *Contemporary Educational Psychology*, *56*, 117–129. https://doi.org/10.1016/j.cedpsych.2018.12.001

de Bruin, A. B., & van Merriënboer, J. J. (2017). Bridging cognitive load and self-regulated learning research: A complementary approach to contemporary issues in educational research. *Learning and Instruction*, *51*, 1–9. https://doi.org/10.1016/j.learninstruc.2017.06.001

DeLuca, C. (2025). Diversity: a necessary imperative for assessment research. *Assessment in Education: Principles Policy & Practice*, *32*(3), 253–256. https://doi.org/10.1080/0969594X.2025.2549158

Dinsmore, D. L., & Parkinson, M. M. (2013). What are confidence judgments made of? Students' explanations for their confidence ratings and what that means for calibration. *Learning and Instruction*, *24*, 4–14. https://doi.org/10.1016/j.learninstruc.2012.06.001

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Sage.

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

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational psychologist*, *46*(1), 6–25. https://doi.org/10.1080/00461520.2011.538645

Efklides, A., & Vlachopoulos, S. P. (2012). Measurement of metacognitive knowledge of self, task, and strategies in mathematics. *European Journal of Psychological Assessment*, *28*(3), 227–239. https://doi.org/10.1027/1015-5759/a000145

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, *105*(1), 98–121. https://doi.org/10.1016/j.obhdp.2007.05.002

Eriksson, K., Lindvall, J., & Helenius, O. (2024). When competence and confidence are at odds: A cross-country examination of the Dunning–Kruger effect. *European Journal of Psychology of Education*, *39*(3), 2641–2658. https://doi.org/10.1007/s10212-024-00804-x

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. https://doi.org/10.3758/BF03193146

Foster, N. L., Was, C. A., Dunlosky, J., & Isaacson, R. M. (2017). Even after thirteen class exams, students are still overconfident: The role of memory for past exam performance in student predictions. *Metacognition and Learning*, *12*(1), 1–19. https://doi.org/10.1007/s11409-016-9158-6

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning*, *3*, 101–121. https://doi.org/10.1007/s11409-008-9021-5

Kostons, D., van Gog, T., & Paas, F. (2012). Training self-assessment and task-selection skills: a cognitive approach to improving self-regulated learning. *Learning and Instruction*, *22*(2), 121–132. https://doi.org/10.1016/j.learninstruc.2011.08.004

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

León, S. P., & Vallejo, A. P. (2021). & Nelson. J. B. Variability in the accuracy of self-assessments among low, moderate, and high performing students in university education. *Practical Assessment, Research & Evaluation, 26*(16). Available online: https://scholarworks.umass.edu/pare/vol26/iss1/16/

León, S. P., Abad, M. J., & Rosas, J. M. (2011). Context–outcome associations mediate context-switch effects in a human predictive learning task. *Learning and Motivation*, *42*(1), 84–98. https://doi.org/10.1016/j.lmot.2010.10.001

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

Lew, M. D., Alwis, W. A. M., & Schmidt, H. G. (2009). Accuracy of students' self-assessment and their beliefs about its utility. *Assessment & Evaluation in Higher Education*, *35*(2), 135–156. https://doi.org/10.1080/02602930802687737

Lipnevich, A. A., Khorramdel, L., & Smith, J. K. (2023). In R. J. Tierney, F. Rizvi, & K. Erkican (Eds.), *Assessment, evaluation, and accountability: A brief introduction* (Vol. 13). Elsevier. International encyclopedia of educationhttps://doi.org/10.1016/B978-0-12-818630-5.09004-7

Mendoza, N. B., & Yan, Z. (2023). Exploring the moderating role of well-being on the adaptive link between self-assessment practices and learning achievement. *Studies in Educational Evaluation*, *77*, 101249. https://doi.org/10.1016/j.stueduc.2023.101249

Metcalfe, J. (2017). Learning from errors. *Annual review of psychology*, *68*, 465–489. https://doi.org/10.1146/annurev-psych-010416-044022

Miller, T. M., & Geraci, L. (2011). Unskilled but aware: reinterpreting overconfidence in low-performing students. Journal of experimental psychology: learning, memory, and cognition, 37(2), 502.

Miller, T. M., & Geraci, L. (2019). Opportunities for self-evaluation increase student calibration in an introductory biology course. *CBE—Life Sciences Education*, *18*(3), ar44. https://doi.org/10.1187/cbe.18-10-0202

Nelson, T. O. (1990). Metamemory: A theoretical framework and new findings. *Psychology of learning and motivation* (Vol. 26, pp. 125–173). Academic.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning*, *1*(2), 159–179. https://doi.org/10.1007/s10409-006-9595-6

Oluk, A., & Korkmaz, H. (2024). Who can predict their performance more accurately? An investigation of undergraduate students' self-assessment behavior in mathematics courses. *Metacognition and Learning*, *19*, 315–347. https://doi.org/10.1007/s11409-024-09381-2

Pronin, E. (2008). How we see ourselves and how we see others. *Science*, *320*, 1177–1180. https://doi.org/10.1126/science.1154199

Pronin, E., & Hazel, L. (2023). Humans' bias blind spot and its societal significance. *Current Directions in Psychological Science*, *32*(5), 406–414. https://doi.org/10.1177/09637214231178745

Raaijmakers, S. F., Baars, M., Paas, F., van Merriënboer, J. J. G., & van Gog, T. (2019). Effects of self-assessment feedback on self-assessment and task-selection accuracy. *Metacognition and Learning*, *14*(1), 21–42. https://doi.org/10.1007/s11409-019-09189-5

Rinne, L. F., & Mazzocco, M. M. M. (2014). Knowing right from wrong in mental arithmetic judgments: Calibration of confidence predicts the development of accuracy. *Plos One*, *9*(7), e98663. https://doi.org/10.1371/journal.pone.0098663

Sheldrake, R., Mujtaba, T., & Reiss, M. J. (2022). Implications of under-confidence and over-confidence in mathematics at secondary school. *International Journal of Educational Research*, *116*, 102085. https://doi.org/10.1016/j.ijer.2022.102085

Spanish National Statistics Institute (2020). *Females in the Teaching Body According to the Grade Level They Teach*. Available in: http://www.ine.es/ss/Satellite?L=es_ES&c=INESeccion_C&cid=1259925481851&p=1254735110672&pagename=ProductosYServicios%2FPYSLayout&param3=1259924822888

Stankov, L., & Lee, J. (2014). Quest for the best non-cognitive predictor of academic achievement. *Educational Psychology*, *34*(1), 1–8. https://doi.org/10.1080/01443410.2013.858908

Teng, F., Lin, H., & Chen, W. (2024). Developing metacognition-based student feedback literacy through formative assessment. *Studies in Educational Evaluation*, *83*, 102–115. https://doi.org/10.1016/j.stueduc.2024.102115

Van der Linden, J., Veenman, M., & Van den Berg, E. (2023). Teachers' conceptions of formative assessment and their impact on students' self-regulated learning in higher education. *Journal of Formative Assessment in Education*, *10*(2), 45–62. https://doi.org/10.1007/s41686-023-00082-8

World Medical Association. (2013). World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *Jama*, *310*(20), 2191–2194. https://doi.org/10.1001/jama.2013.281053

Yan, Z., Wang, X., Boud, D., & Lao, H. (2023). The effect of self-assessment on academic performance and the role of explicitness: a meta-analysis. *Assessment & Evaluation in Higher Education*, *48*(1), 1–15. https://doi.org/10.1080/02602938.2021.2012644

Zimmerman, B. J. (2000). Attaining Self-Regulation: A Social Cognitive Perspective. In *Handbook of Self-Regulation* (pp. 13–39). Academic Press. https://doi.org/10.1016/B978-012109890-2/50031-7

Zimmerman, B. J., & Schunk, D. H. (Eds.). (2013). *Self-regulated learning and academic achievement: Theoretical perspectives*. Routledge.

Zimmerman, B. J., Moylan, A., Hudesman, J., White, N., & Flugman, B. (2011). Enhancing self- reflection and mathematics achievement of at-risk urban technical college students. *Psychological Test and Assessment Modeling*, *53*(1), 141–160. Available online athttp://www.psychologie-aktuell.com/journale/psychological-test-and-assessment-modeling.html