Metacognitive Monitoring Accuracy and Self-Regulated Learning Outcomes in Undergraduate STEM Cohorts: An Empirical Analysis of Calibration Bias Under Distributed Practice Schedules

Dana Williams
Associate Professor
University of Toronto
27 King's College Circle, Toronto, Ontario M5S 1A1, Canada

Sam Walker
PhD
Utrecht University
Heidelberglaan 8, 3584 CS Utrecht, Netherlands

Kim Roberts
Professor
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

Abstract. Metacognitive calibration — the alignment between learners' confidence judgments and actual performance — remains insufficiently characterized within distributed practice contexts in tertiary STEM education. This study examined 318 undergraduate students across three research-intensive universities, employing pre/post judgment-of-learning (JOL) protocols, spaced retrieval tasks, and hierarchical linear modeling to isolate the effects of calibration bias on summative assessment outcomes. Findings indicate that over-confident learners (positive calibration bias ≥ 0.25) demonstrated a statistically significant 14.3% decrement in end-of-semester examination scores compared to well-calibrated peers ($p < .001$), independent of prior academic achievement. Instructional scaffolding targeting metacognitive awareness attenuated this bias by 38.6% across a 12-week intervention. These results underscore the urgency of embedding systematic calibration training within undergraduate STEM curricula to mitigate performance gaps attributable to inaccurate self-assessment.

Keywords: metacognitive calibration, judgment of learning, self-regulated learning, distributed practice, calibration bias, undergraduate STEM education, spaced retrieval practice, hierarchical linear modeling, metacognitive scaffolding

Introduction

The capacity to accurately monitor one's own comprehension and knowledge state — a process formally designated as metacognitive monitoring — constitutes a foundational determinant of self-regulated learning (SRL) in higher education. In the context of rapidly evolving STEM disciplines, where the volume and complexity of conceptual material continue to escalate, learners who cannot reliably assess the boundaries of their own

understanding are systematically disadvantaged, not merely in formative assessments but across the full arc of professional preparation. The global expansion of tertiary STEM enrolments between 2020 and 2024, documented extensively by the UNESCO Institute for Statistics, has intensified institutional pressure to address persistent achievement gaps that cannot be explained solely by socioeconomic or preparatory factors. Accumulating psychometric evidence implicates metacognitive calibration accuracy — operationally defined as the concordance between prospective confidence ratings and objective task performance — as a primary mediating variable in these disparities. Calibration bias, specifically the overconfidence effect wherein learners systematically overestimate their mastery of domain-specific material prior to retrieval-based assessment, has been robustly documented in laboratory paradigms since the seminal work of Koriat and Bjork (2005). However, the ecological validity of such findings within authentic undergraduate instructional environments, particularly those incorporating distributed practice schedules — a pedagogical arrangement shown to enhance long-term retention through the spacing effect — remains an empirical lacuna. Distributed practice introduces temporal discontiguity between encoding and retrieval episodes, which both attenuates subjective familiarity-based confidence and alters the metacognitive cues available to the learner during JOL elicitation. Whether this environmental manipulation systematically modulates calibration bias trajectories across a full academic semester, and whether instructional scaffolding can remediate maladaptive monitoring patterns at scale, are questions of immediate translational relevance for curriculum designers and educational psychologists operating in 2024–2026 institutional contexts. The present investigation directly addresses these open questions through a longitudinal, multi-site empirical analysis involving 318 undergraduate participants enrolled in introductory biochemistry, discrete mathematics, and engineering mechanics courses. Participants completed bi-weekly JOL assessments anchored to spaced retrieval prompts, permitting Gamma correlation-based calibration indices to be computed at the individual level across twelve time-points. Hierarchical linear models were specified to partition variance attributable to between-student calibration trajectories, course-level instructional scaffolding conditions, and time-on-task covariates. The paper is organized as follows: Section 2 reviews the theoretical frameworks underpinning metacognitive monitoring, calibration bias, and spaced practice interactions; Section 3 details the multi-site sampling procedure, measurement instruments, and analytic strategy; Section 4 presents the hierarchical linear modeling results alongside descriptive calibration trajectory plots; Section 5 discusses implications for undergraduate STEM curriculum design and metacognitive intervention development; and Section 6 concludes with recommendations for institutional assessment policy and directions for future longitudinal replication.

References

1. Гурко О.В. Лексика графічного дизайну в українській мові кінця XX – початку XXI століття : дис. … канд. філол. наук : 10.02.01. Дніпропетровськ, 2008. 185 с.
2. Гурко, О. В. (2023). Інтенція персуазивності в мовленні В. Зеленського. Англістика та американістика: зб. наук. пр./Дніпр. нац. ун-т ім. Олеся Гончара. Вип, 20, 13-17.
3. Дячук, Ю. С. (2014). Украинско-российские параллели в переводе современной французской художественной прозы. Язык и культура, (2 (26)), 74-88.