

Supplementary Materials

To ensure the psychometric adequacy of the instruments used in this study, a systematic evaluation of their structural validity and internal consistency was conducted prior to the main analyses. When instruments had not been previously validated in the present context or had undergone structural adaptation, their latent structure was examined within a structural equation modeling (SEM) framework.

Exploratory Factor Analysis (EFA) was conducted when the dimensionality of the scale was uncertain, followed by Confirmatory Factor Analysis (CFA) to test the stability and adequacy of the identified factorial structure. Model fit was evaluated using multiple complementary indices, including the Comparative Fit Index (CFI), Tucker–Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA) with 90% confidence intervals, and the Standardized Root Mean Square Residual (SRMR), following the recommendations of Hair et al. (2019).

Internal consistency was assessed using both Cronbach's alpha (α) and McDonald's omega (ω). Reporting both coefficients provides a more robust estimate of reliability, particularly when the tau-equivalence assumption may be violated. Reliability coefficients were interpreted using the conventional benchmarks proposed by George and Mallery (2003). When applicable, item-level diagnostics were examined to ensure that all indicators contributed meaningfully to construct representation.

This procedure ensured adequate structural validity and reliability of the instruments, supporting the robustness of the subsequent analyses.

1. Self-regulated learning scale

The scale used in this study to assess students' self-regulated learning is a composite of three scales, each reflecting how the individual regulates the specific dimensions proposed by Wolters et al. (2005): Strategies for the Regulation of Academic Cognition, Strategies for the Regulation of Academic Behavior and Strategies for the Regulation of Academic Motivation.

1.1. Strategies for the Regulation of Academic Cognition

The first scale, Strategies for the Regulation of Academic Cognition, comprises two subscales: one assessing cognitive strategies and the other assessing metacognitive strategies that students use in mathematics learning. These subscales are grounded in social-cognitive models of self-regulated learning and have been widely used to measure students' strategic regulation of cognition in academic contexts.

Cognitive self-regulation was measured using a 14-item scale that assessed students' use of learning strategies to process, organize, and retain academic information. These items were

organized into three dimensions: Rehearsal strategies (4 items), Elaboration strategies (6 items), and Organization strategies (4 items). These dimensions reflect increasing levels of cognitive engagement, from more surface-level processing (e.g., rehearsal: *I practice by repeating the material several times*) to deeper processing strategies (e.g., elaboration and organization: *I try to connect math ideas with those from other subjects whenever possible*, and for teachers, e.g., *When I study from the textbook, I underline the material to help organize my thoughts.*).

Metacognitive self-regulation was assessed using a 12-item scale (e.g., *"I set goals to direct my activities for each study period"*) designed to measure students' planning, monitoring, and regulation of their learning processes. This dimension reflects higher-order self-regulatory activities.

All items were rated on a five-point Likert scale, with higher scores indicating more frequent use of cognitive or metacognitive self-regulation strategies. Scale scores were computed by averaging item responses within each dimension, with higher values reflecting greater strategic engagement in learning.

A confirmatory factor analysis (CFA) was conducted to test the original four-factor structure of the Strategies for the Regulation of Academic Cognition scale (Memorization, Organization, Elaboration, and Metacognitive Regulation). The model showed acceptable fit, $\chi^2(253) = 433$, $p < .001$, CFI = .917, TLI = .902, SRMR = .049, RMSEA = .048, 90% CI [.041, .056]. However, correlations among the cognitive factors were extremely high (ranging from .888 to 1.017), indicating substantial overlap and limited discriminant validity.

Given these findings, a more parsimonious two-factor solution was specified, distinguishing Cognitive Self-Regulation (14 items) and Metacognitive Self-Regulation (12 items). Fit indices indicated an acceptable model fit (CFI = .919, TLI = .906, SRMR = .044, RMSEA = .046). All standardized loadings were statistically significant ($p < .001$) and of moderate to high magnitude.

Internal consistency was good for both dimensions: Cognitive Self-Regulation ($\alpha = .857$; $\omega = .859$) and Metacognitive Self-Regulation ($\alpha = .817$; $\omega = .822$). The global scale showed excellent reliability ($\alpha = .911$; $\omega = .912$), supporting the construct's internal coherence.

1.2. Strategies for the Regulation of Academic Behavior

The behavioral regulation instrument, proposed by Wolters et al. (2005), comprises 12 items organized into two dimensions: Effort regulation and Regulation of Time and Study Environment. Regulation of time and study environment was measured using eight items that assess students' ability to manage when, where, and under what conditions they study. These strategies reflect students' proactive management of external conditions to sustain engagement in mathematical work. Effort regulation was assessed using four items that measure students' persistence and self-discipline when faced with mathematical challenges. Effort regulation is considered a core behavioral component of self-regulated learning, as it reflects students' ability to translate intentions and goals into sustained academic action. All items were rated on a five-point Likert scale, with higher scores indicating greater use of behavioral self-regulation

strategies. Scores for each dimension were computed by averaging the corresponding item responses, with higher values reflecting more effective regulation of time and study environment, as well as effort.

A confirmatory factor analysis was conducted using the original two-factor structure of the behavioral self-regulation scales, Time and Study Environment Management (TSE) and Effort Regulation (ER). Although this model demonstrated acceptable global fit (CFI = .955; TLI = .911; RMSEA = .062; SRMR = .041), the two subscales had only modest reliability ($\alpha = .677$; $\alpha = .695$), and several TSE items had weak standardized loadings (e.g., .236, .293). These findings indicated limited discriminant validity between the two constructs. The behavioral regulation scale was then examined using a unidimensional model that included all 12 items. The model showed acceptable fit, $\chi^2(48) = 106$, $p < .001$, CFI = .932, TLI = .907, SRMR = .049, RMSEA = .063, 90% CI [.047, .080]. Standardized loadings ranged from .218 to .709 and were statistically significant.

Internal consistency was adequate ($\alpha = .809$; $\omega = .817$), supporting retention of the unidimensional structure for subsequent analyses.

1.3. Strategies for the Regulation of Academic Motivation

Motivational self-regulation strategies were assessed using the Academic Self-Regulation Questionnaire (SRQ-A), in its Portuguese validated version developed by Gomes et al. (2019), grounded in Self-Determination Theory (SDT, Ryan & Deci, 2000). Previous validation studies with Portuguese student samples (Gomes et al., 2019) have demonstrated that this 16-item version of the SRQ-A shows adequate internal consistency, a simplex pattern of correlations consistent with SDT assumptions, and good construct validity, supporting its use for assessing both specific motivational regulations and overall motivational self-regulation in educational research.

The scale comprises four dimensions, each assessed with four items, representing distinct regulatory styles: External Regulation, reflecting engagement driven by external demands, rewards, or avoidance of punishment; Introjected Regulation, capturing internally pressurized behavior regulated by guilt, shame, or contingent self-worth; Identified Regulation, representing engagement based on the personal importance and perceived value of the activity; Intrinsic Regulation, reflecting engagement driven by inherent interest and enjoyment. All items were rated on a five-point Likert-type frequency scale, with higher scores indicating stronger endorsement of each regulatory style. Scores for each dimension were computed by averaging the corresponding item responses.

Given its established factorial validity, only internal consistency was evaluated in the present sample. Internal consistency coefficients were acceptable to good across dimensions: External Regulation ($\alpha = .716$; $\omega = .736$), Introjected Regulation ($\alpha = .699$; $\omega = .707$), Identified Regulation ($\alpha = .856$; $\omega = .859$), and Intrinsic Regulation ($\alpha = .789$; $\omega = .797$).

In addition, a Relative Autonomy Index (RAI) was computed to provide a global indicator of motivational self-regulation. The RAI is a scoring method proposed initially by Grolnick and Ryan (1989) and widely used in SDT-based research to represent the degree to which individuals' motivation is relatively autonomous versus controlled. Higher RAI values indicate a more self-determined motivational profile, characterized by greater reliance on identified and intrinsic regulation relative to external and introjected regulation. More autonomous forms of regulation receive positive weights, whereas more controlled forms receive negative weights. The RAI was calculated using the following formula: $RAI = (-2 \times \text{External}) + (-1 \times \text{Introjected}) + (1 \times \text{Identified}) + (2 \times \text{Intrinsic})$

2. Formative assessment scale

2.1. Students' version (SV)

The FA scale was developed in two versions to capture both teacher and student perceptions of classroom pedagogical and evaluative practices. The instrument was originally grounded in Wiliam and Thompson (2007) five strategies, which served as the initial theoretical dimensions: 1) Clarifying, Sharing, and Understanding Learning Intentions and Criteria for Success: This dimension includes 4 items assessing the extent to which learning goals and success criteria are made explicit, understandable, and visible to students (e.g., *My teacher talks with us about the steps to complete a task successfully*); 2) Engineering Effective Classroom Discussions, Activities, and Tasks that Elicit Evidence of Learning: This dimension consists of 9 items measuring the use of questioning, discussions, and learning tasks designed to elicit evidence of students' understanding and to inform instructional decisions (e.g., *The teacher gives students enough time to ask questions*); 3) Providing Feedback that Moves Learning Forward: This dimension includes 16 items assessing the provision of feedback focused on improvement rather than evaluation, emphasizing guidance on how to close the gap between current and desired performance (e.g., *When I don't understand a topic, my teacher tries to explain it to me in a different way*) ; 4) Activating Students as Instructional Resources for One Another: This dimension comprises 4 items capturing practices that promote peer interaction and peer assessment, such as collaborative work, peer feedback, and shared problem solving (e.g., *In math class, we exchange and correct each other's tasks*); 5) Activating Students as Owners of Their Own Learning: This dimension includes 6 items assessing the extent to which students are encouraged to self-assess, set learning goals, monitor their progress, and take responsibility for their own learning processes, thus supporting self-regulated learning (e.g., *My teacher lets me think about how I want to learn mathematics*) .

The initial scale consisted of 39 items, comprising 29 items adapted from the Student and Teacher Assessment for Learning Questionnaires developed by Pat-El et al. (2013) and 10 new items developed to complement the intended analytical scope. To allow for direct comparison of perceptions, parallel items were authored for the teacher version (Pat-El et al., 2013).

The factorial structure of the formative assessment scale was initially examined using Exploratory Factor Analysis (EFA) with the minimum residual (minres) extraction method and

oblimin rotation, which allows the factors to correlate. The assumptions for conducting the EFA were satisfied.

The Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy indicated an excellent overall value (KMO = .959), suggesting that the correlation matrix was highly suitable for factor analysis. Individual KMO values for the items were also very high, ranging from .909 to .980, indicating that all items showed adequate shared variance. In addition, Bartlett’s test of sphericity was statistically significant, $\chi^2(351) = 6050$, $p < .001$, confirming that the correlation matrix differed significantly from an identity matrix and was therefore appropriate for factor extraction.

The factor solution revealed three factors with eigenvalues greater than 1, explaining 45.5% of the total variance. The first factor had an eigenvalue of 6.08, accounting for 22.5% of the variance; the second had an eigenvalue of 4.76, explaining 17.6%; and the third had an eigenvalue of 1.46, explaining 5.4%.

Factor loadings indicated that items associated with Feedback loaded primarily on the first factor, with relatively high loadings ranging from .49 to .84, suggesting a robust dimension related to feedback practices that support learning. The second factor grouped items related to tasks, learning objectives, and the elicitation of evidence of learning, with moderate-to-high loadings generally ranging from .48 to .68. The third factor included items associated with shared regulation processes, such as peer assessment and student participation in the regulation of learning, showing moderate loadings.

The inter-factor correlations indicated moderate relationships among the dimensions. The first factor correlated with the second ($r = .663$) and with the third ($r = .495$), whereas the correlation between the second and third factors was lower ($r = .244$).

These results suggest that, although the dimensions are related within the broader construct of formative assessment, they retain sufficient conceptual distinction to be considered separate factors.

Overall, the EFA provided empirical support for a three-factor structure of the scale, consistent with the different components of formative assessment practices. The resulting structure demonstrated both statistical adequacy and conceptual coherence, serving as the basis for the subsequent Confirmatory Factor Analysis (CFA) conducted to test the stability and fit of the identified factorial model.

Before the final confirmatory analysis, the scale underwent systematic refinement. Item reduction was guided by both statistical and theoretical criteria. Specifically, items were removed when they (a) showed low factor loadings, indicating a weak contribution to the latent construct, (b) exhibited substantial cross-loadings across factors, suggesting limited discriminant validity, (c) were redundant with conceptually overlapping items, or (d) lacked conceptual clarity within the emerging factorial structure. The aim of this refinement process was to obtain a more parsimonious instrument while preserving the theoretical coverage of the formative assessment construct.

As a result of this process, 12 items were removed, yielding a refined 27-item version of the scale. These results indicate that the 27-item structure represents the final validated version of the scale used in the present study, balancing statistical fit, internal consistency, and theoretical coherence.

The refined 27-item model was subsequently tested via CFA and demonstrated satisfactory fit: $\chi^2(315) = 552, p < .001$; CFI = .933; TLI = .925; SRMR = .049; RMSEA = .050 (90% CI [.043, .057]).

Standardized factor loadings ranged from .390 to .753 and were all statistically significant ($p < .001$). Internal consistency was good for Clarifying and Elicitation ($\alpha = .821$; $\omega = .823$), excellent for Feedback ($\alpha = .925$; $\omega = .925$), and acceptable for Shared Regulation ($\alpha = .706$; $\omega = .712$).

The internal consistency of the student version was also examined by modeling it as a unidimensional structure, labeled Global Formative Assessment (Global FA). This decision was grounded in evidence that Assessment for Learning can be modeled as both a higher-order construct and a set of distinct yet related components (Pat-El et al., 2013). Accordingly, perceived formative assessment was operationalized primarily through this Global FA score for hypothesis testing. Results indicated excellent internal consistency, with Cronbach's alpha ($\alpha = .931$) and McDonald's omega ($\omega = .931$), supporting the use of a global formative assessment score in subsequent analyses.

The scale was therefore composed of 27 items distributed across three dimensions:

Clarification and Elicitation (11 items). This dimension assesses the extent to which learning objectives and criteria are clearly communicated, understood, and made explicit to students, and how the teacher uses tasks, questioning, and classroom discussions to elicit evidence of learning. It includes practices such as clarifying intended learning outcomes, aligning tasks with objectives, and using instructional strategies that enable teachers to monitor students' understanding throughout the learning process (e.g., Students Version(SV) *"When the math teacher gives me a task to do, I clearly know what I can learn from that task"*/ Teacher Version(TV) *"I ensure my students know what they can learn from their tasks"*).

Feedback (12 items). This dimension assesses the quality and function of the teacher's feedback, with a focus on its formative role. It examines the extent to which feedback is improvement-oriented, provides specific guidance on how to progress, and helps students understand the gap between their current performance and the desired level of achievement. It also includes elements that encourage reflection and support the development of learner autonomy (e.g., (SV) *"My teacher talks to me about how to discover my strengths in order to improve"*/ (TV) *"I discuss with my students how to use their strengths to improve their tasks"*).

Shared Regulation (4 items). This dimension assesses the extent to which responsibility for the learning process is shared with students. It encompasses practices that promote active student participation, such as self-assessment, peer assessment, and involvement in decisions about their own learning. It therefore reflects the promotion of self-regulation and shared

responsibility in managing students' academic development (e.g., (SV) *"My teacher let me think about how I want to learn mathematics"* / (TV) *I involve my students in thinking about how they want to learn mathematics*).

All items were rated on a five-point Likert scale, with higher scores indicating a stronger perceived presence of formative assessment practices. Scores were computed by averaging item responses within each dimension.

2.2. Teachers' version (TV)

As part of the analysis of the psychometric properties of the instruments, a Confirmatory Factor Analysis (CFA) was conducted on the teacher version of the Formative Assessment scale, using the three-dimensional structure previously identified and validated in the student sample (Clarification and Elicitation, Feedback, and Shared Regulation). The aim was to test the model's adequacy for teachers' perceptions, thereby ensuring structural comparability between the two versions of the instrument.

The teacher sample comprised 42 participants. CFA conducted on the three-factor structure derived from the student model yielded very poor fit indices (CFI = .039, TLI = -.072, SRMR = .214, RMSEA = .741, 90% CI [.723, .759]). Similarly, the unidimensional Global FA model showed inadequate fit (CFI = .036, TLI = -.065, SRMR = .216, RMSEA = .727, 90% CI [.720, .733]).

However, the fit indices from the CFA of the teacher version were weak. These results should be interpreted in light of an important methodological limitation: the small sample size ($N = 42$). CFA is sensitive to sample size and generally requires larger samples to produce stable parameter estimates and robust fit indices. Small samples yield less stable parameter estimates, inflated standard errors, and consequently poorer model fit indicators, even when the underlying structure is theoretically sound. Therefore, it is plausible that the weak fit indices primarily reflect limitations in statistical power rather than a conceptual inadequacy of the model (Brown, 2015; Kline, 2016).

Despite the poor model fit, the internal consistency estimates were acceptable: Clarifying and Elicitation ($\alpha = .770$; $\omega = .760$), Feedback ($\alpha = .750$; $\omega = .760$), Shared Regulation ($\alpha = .700$; $\omega = .750$), and Global FA ($\alpha = .729$; $\omega = .747$) indicating acceptable internal reliability, particularly given the limited number of participants.

Given these limitations, the fit indices should be interpreted cautiously. Nevertheless, the teacher version of the scale was retained because of its theoretical coherence, acceptable reliability, and its analytical relevance for comparing teachers' and students' perceptions of formative assessment practices (Brown, 2015; Kline, 2016).

Overall, the psychometric analyses provide evidence of the instruments' structural validity and reliability in the present sample. Where necessary, theoretically guided model refinements were implemented and empirically supported, ensuring robust measurement for subsequent analyses.