International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(23s) 2026 pp. 1330-1349
https://doi.org/10.70917/ijcisim-2026-5696
Article

Data-Driven Evaluation of Mathematics Teacher Development: Within-Teacher Change and Territorial Heterogeneity in Panama

Luisa Morales-Maure [1], [5], *, Orlando García-Marimón[2],[5], Rosa Duran-González[3], Laura Miriam Franco Sánchez[4]

[1]Department of Mathematics, Faculty of Natural, Exact and Technological Sciences, Universidad de Panamá, Panama City, Panama. ORCID: http://orcid.org/0000-0003-3905-9002
[2]Department of Mathematics, Faculty of Natural, Exact and Technological Sciences, Universidad de Panamá, Panama City, Panama. ORCID: https://orcid.org/0000-0002-0334-6133
[3]Institute of Social Sciences and Humanities, Universidad Autónoma del Estado de Hidalgo, Pachuca, Hidalgo, México. ORCID: https://orcid.org/0000-0001-8121-5019
[4]Institute of Social Sciences and Humanities, Universidad Autónoma del Estado de Hidalgo, Pachuca, Hidalgo, México. ORCID: https://orcid.org/0000-0002-0676-9510
[5]National Research System (SNI), SENACYT, Panama City, Panama
*Corresponding Author: E-mail: luisa.morales@up.ac.pa

Abstract: Large-scale professional-development programmes generate longitudinal educational data that can be used to examine both within-teacher change and territorial heterogeneity. This study presents a data-driven assessment of perceived professional learning across six dimensions of didactic suitability among 420 Panamanian in-service mathematics teachers enrolled in the Estrategias Didácticas para la Enseñanza de la Matemática (EDEM) programme in 2017, 2018, 2020, and 2021. A total of 409 teachers contributed linkable pre- and post-training self-assessments. A 45-item common core was harmonised across questionnaire versions, creating a reproducible longitudinal dataset analysed through reliability estimation, paired Wilcoxon tests, rank-biserial effect sizes, multiplicity-adjusted subgroup comparisons, and distribution-sensitive visual analytics. Self-ratings increased across all six dimensions. Affective ($\Delta = 0.74$; $r = .739$) and interactional ($\Delta = 0.69$; $r = .765$) changes were largest, whereas mediational change was small ($\Delta = 0.10$; $r = .134$; Holm-adjusted $p = .031$). Affective, mediational, and interactional change varied across adequately represented provinces, while gender differences did not survive multiplicity adjustment. The results show how longitudinal, data-driven analysis can reveal patterns of professional learning and territorial heterogeneity that would be obscured by aggregate post-training scores alone. The findings are interpreted as changes in teacher self-perception rather than causal programme effects because the study lacks a comparison group and direct observation of classroom practice.
Keywords: Didactic Suitability; Mathematics Teacher Development; Data-Driven Evaluation; Longitudinal Analysis; Territorial Heterogeneity; Teacher Self-Assessment; Panama

1. INTRODUCTION

Large-scale professional-development programmes produce repeated educational measurements that can support more informative evaluation than end-point averages alone. When pre- and post-training records can be linked at the teacher level, they permit direct examination of individual change; when those records retain province and cohort identifiers, they can also reveal whether patterns are distributed uniformly across territories and implementation periods. The analytical challenge is therefore not simply to estimate a national post-training profile, but to construct a longitudinal dataset, quantify within-teacher change, test subgroup heterogeneity, and visualise distributions without

converting descriptive differences into causal claims. This study applies that data-driven longitudinal perspective to EDEM, a nationwide mathematics professional-development programme in Panama.

This distinction is important because the quality of mathematics teaching cannot be reduced to disciplinary knowledge alone. Teachers must select and connect mathematical meanings, anticipate students' interpretations, organise classroom discourse, foster participation, respond to errors, mobilise material and digital resources, and adapt instruction to institutional and sociocultural conditions. The Onto-Semiotic Approach to Mathematical Knowledge and Instruction brings these requirements together through the construct of didactic suitability, understood as a reasoned assessment of the coherence and adequacy of an instructional process across six interdependent dimensions: epistemic, cognitive, interactional, affective, ecological, and mediational (Godino, 2009, 2011; Font et al., 2013). From this perspective, instructional quality depends less on the isolated presence of particular techniques than on the relationships established among mathematical content, learners, interaction, resources, emotions, and context.

The reflective function of this framework is especially relevant for teacher education. Didactic suitability criteria provide educators with a shared professional language for examining why an instructional decision may be mathematically appropriate, cognitively accessible, affectively productive, interactionally viable, adequately mediated, and responsive to its institutional environment. Research on prospective teachers' reflections about mathematical errors has similarly shown that the analysis of classroom incidents can reveal how teachers mobilise criteria to identify weaknesses, justify alternatives, and propose improvements to instruction (Font et al., 2024).

Although the six dimensions are theoretically integrated, evidence from professional development programmes is often reported through decontextualised national averages. This practice creates at least two blind spots. First, high mean scores may conceal unequal access to the institutional and material conditions required to enact what teachers have learned. Second, a programme may coincide with changes in knowledge, confidence, and reflective capacity while leaving structural constraints on classroom dialogue, technological mediation, curricular adaptation, and contextual relevance largely unchanged. Territorial analysis is therefore necessary for distinguishing perceived professional learning from the local capacity to enact it. EDEM offers an informative Panamanian context: the 2017, 2018, 2020, and 2021 cohorts represented all ten provinces, and broader programme records included three indigenous comarcas. The national scope supported a shared language for analysing mathematics instruction while bringing together teachers working under markedly different infrastructural, institutional, and sociocultural conditions. The paired assessments permit analysis of within-participant change in perceived didactic suitability, and the participation records situate those changes territorially.

Recent studies reinforce the need to connect professional reflection with the design and enactment of mathematically meaningful activity. Rodríguez-Nieto et al. (2026) show that students' work with frequency distribution tables is organised through systems of connections among representations, procedures, concepts, meanings, and extra-mathematical contexts. Their findings suggest that epistemic quality is not equivalent to possessing isolated content knowledge; it involves recognising, coordinating, and promoting relationships among mathematical objects and situations. In professional development, teachers' epistemic growth should therefore be interpreted in terms of their capacity to design learning opportunities that make such connections visible and discussable.

This relational understanding is complemented by studies of practical argumentation and reflective judgement. Cortés et al. (2026) demonstrate that didactic suitability criteria can organise teachers' analysis of gamification and game-based learning, redirecting reflection from enthusiasm for an innovation towards its mathematical, cognitive, affective, mediational, interactional, and ecological consequences. Ledezma et al. (2026) likewise show that prospective teachers' decisions in mathematical modelling are supported by practical arguments involving purposes, anticipated consequences, contextual relevance, and pedagogical values. Taken together, these contributions position teacher learning as the development of an ability to justify instructional decisions rather than merely reproduce recommended methods. Research involving educational robotics and computational thinking further extends this argument. Sala-Sebastià et al. (2026a) describe a BlueBot-based simulation designed to develop computational

thinking while incorporating a gender perspective. The relevance of this work to the present study lies not simply in the use of a digital device, but in the coordination required among task design, mathematical and computational meanings, student participation, technological mediation, and equity considerations. Such experiences illustrate that mediational suitability cannot be inferred from the availability of technology alone. It depends on whether technological resources are incorporated into pedagogically coherent activities that broaden participation and respond to diverse learners.

A related study by Sala-Sebastià et al. (2026b) examines the criteria adopted by teacher educators when reflecting on a robot-based problem-solving activity. Its findings strengthen the view that technology-rich instruction must be evaluated relationally: the educational value of robots depends on the mathematical opportunities they generate, the forms of interaction they support, their cognitive accessibility, and their correspondence with wider curricular purposes. This is particularly important for territorially diverse professional-development programmes, because access to devices, connectivity, technical assistance, and opportunities for classroom experimentation may vary significantly across regions.

The professional-learning literature also suggests that sustained change depends on collaborative structures extending beyond short training experiences. Richit et al. (2026) analyse the professional learning of primary teachers through lesson studies and emphasise learning generated through collaborative planning, observation, analysis, and revision of lessons. This perspective is especially pertinent to EDEM because it distinguishes exposure to theoretical principles from the iterative development of practice. A nationally shared framework may introduce criteria for analysing instruction, but their stable enactment is more likely when teachers can test ideas, examine classroom evidence, and revise decisions with colleagues over time.

Collectively, these studies point towards a broader conception of mathematics teacher development. Professional learning involves the construction of mathematical connections, the formulation of practical arguments, critical reflection on errors, the pedagogically justified use of technologies, attention to gender and participation, and sustained collaborative inquiry. These capacities cannot be examined independently of the environments in which teachers are expected to exercise them. Digital tools may enable computational activity in one setting but remain inaccessible or pedagogically unsupported in another. Collaborative lesson analysis may flourish where institutional time and professional networks are available yet prove difficult to sustain in isolated or under-resourced schools. The same training programme can therefore produce shared conceptual resources while generating uneven possibilities for enactment.

Against this background, this study aimed to develop a data-driven longitudinal assessment of perceived professional learning in a national mathematics teacher-development programme and to identify patterns of within-teacher change and territorial heterogeneity across six dimensions of didactic suitability. Two questions guided the analysis: (1) how did teachers' self-reported epistemic, cognitive, interactional, affective, ecological, and mediational suitability change from pre- to post-training; and (2) to what extent did individual change scores vary by province, cohort year, and gender? The analytical emphasis was on linked observations, effect magnitude, measurement reliability, multiplicity-controlled subgroup comparisons, and distribution-sensitive visualisation, while avoiding causal interpretation of self-reported change.

2. LITERATURE REVIEW AND ANALYTICAL FRAMEWORK

The analytical framework links didactic suitability with a contextual reading of professional learning. Rather than treating instructional quality as a list of independent competencies, it considers how mathematical meanings, student understanding, interaction, affect, resources, and institutional conditions operate together. This relational perspective provides the basis for interpreting both within-teacher change and variation across territorial settings.

2.1 Didactic suitability as a relational judgement

Didactic suitability is sometimes represented as six adjacent criteria, but its theoretical value lies in their interdependence. Epistemic suitability concerns the representativeness, coherence, and richness of the mathematical

meanings made available. Cognitive suitability concerns the relation between those meanings and learners' prior knowledge, difficulties, and potential development. Interactional suitability addresses how discourse, participation, and feedback permit conflicts of meaning to become visible and productive. Affective suitability concerns interests, emotions, beliefs, values, and willingness to engage. Mediational suitability examines whether time, technologies, representations, and material resources support the intended activity. Ecological suitability situates instruction in relation to curricula, institutions, community practices, cultural expectations, and wider educational demands (Godino, 2009, 2011; Font et al., 2013).

No single dimension can compensate fully for weakness in another. Mathematically rich tasks may remain inaccessible without cognitive adaptation; digital tools may add little when they are not integrated into mathematical reasoning; positive affect may encourage participation but cannot replace conceptual coherence. Suitability is therefore a reasoned judgement about a configuration rather than an additive score. This interpretation is consistent with work showing that the criteria orient practical reflection and teacher argumentation in lesson design (Calle et al., 2024; Campuzano et al., 2024; Campo-Meneses et al., 2024; Sol et al., 2024).

Cortés et al. (2026) make this function especially clear: when teachers examine gamification and game-based learning through didactic suitability criteria, reflection moves beyond enthusiasm for an innovation towards analysis of its mathematical, cognitive, affective, interactional, mediational, and ecological consequences. The relevance to EDEM is direct. A national programme should not be judged only by whether participants value it, but by whether it supports balanced reasoning across the criteria and whether territorial settings permit that reasoning to be enacted.

2.2 Epistemic quality, mathematical connections, and practical argumentation

Epistemic suitability is broader than accuracy. It concerns the range and organisation of mathematical practices that learners can encounter problem solving, representation, argumentation, modelling, generalisation, and connections among objects. Pino-Fan et al. (2017) illustrate how the OSA can identify the meanings mobilised in teaching and the tensions created when definitions or representations are treated narrowly. This perspective helps explain why professional development that incorporates structured task analysis may produce strong epistemic self-perceptions.

Rodríguez-Nieto et al. (2026) extend this argument in statistics education by documenting how students connect representations, procedures, concepts, and contextual information while solving problems involving frequency distribution tables. Their "new view" of connections is pertinent here because it clarifies what teachers' epistemic growth should ultimately enable: not more content in isolation, but richer networks of mathematical meaning. Within EDEM, a high epistemic profile can therefore be interpreted as increased perceived capacity to recognise and organise such connections. It cannot, however, establish that these connections were observed in classroom practice, an important boundary of the present self-report evidence.

Ledezma et al. (2026) add a complementary dimension. Their analysis of practical arguments in modelling shows that pedagogical design involves warrants, values, and anticipated consequences, not only technical decisions. Teachers transform tasks by arguing why a context, representation, or modelling cycle is educationally appropriate. This links epistemic and ecological suitability: mathematical authenticity and contextual relevance must be justified together. It also links epistemic and interactional suitability because argumentation becomes visible through discourse, questioning, and comparison of solutions.

2.3 Affect as a condition of mathematical participation

Affect in mathematics education encompasses emotions, attitudes, beliefs, values, motivation, and identities (McLeod, 1989; Hannula et al., 2004; Zan et al., 2006; Leder & Forgasz, 2006; Hannula, 2012). These are not external influences on cognition. They shape what learners notice, whether they persist, how they interpret error, and whether they regard participation as legitimate. For teachers, affect also influences the selection of tasks, tolerance for uncertainty, management of discussion, and confidence in adopting unfamiliar resources.

The affective criterion within didactic suitability is thus best understood as an organising condition of engagement. A supportive emotional climate can make demanding mathematical activity possible, while anxiety or low expectations can narrow it. Sánchez Aguilar (2022) and Godino et al. (2022) similarly position affective factors as consequential for instructional decision-making. The co-occurrence of strong epistemic and affective profiles would therefore be theoretically meaningful: conceptual reflection and professional confidence may reinforce one another when teachers analyse tasks collaboratively and connect them with classroom experience.

2.4 Territorial conditions as opportunity structures for enactment

Territory is not treated here as a label attached after the analysis. It is an opportunity structure composed of institutional presence, connectivity, travel, resource circulation, language, community relationships, and local histories of programme participation. These conditions shape mediational and ecological suitability directly, but they also influence interactional opportunities and the sustainability of professional communities.

Rico and Lupiáñez (2008) and Gómez-Flores et al. (2023) argue that teacher education should respond to socio-territorial realities rather than assume that centralised provision is equally usable everywhere. This becomes especially important when professional development includes digital or hybrid components. Access to technology is only one layer; pedagogical mediation also requires functional equipment, time, technical support, relevant materials, and opportunities to test and revise practice. Territorial dispersion in mediational and ecological self-perceptions can consequently indicate not weak teacher commitment, but unequal enactment conditions.

The framework therefore treats each teacher's paired pre- and post-training self-assessments as conceptually situated within a territorial participation ecology. The questionnaires indicate how participants perceived their suitability before and after the diploma; the participation records indicate where the programme had greater or lesser institutional penetration. Read together, these sources support cautious interpretation of perceived change alongside locally differentiated opportunity, without presuming that the territorial mechanisms were measured.

2.5 Data-driven longitudinal evaluation and visual analytics

In this study, data-driven evaluation denotes an auditable analytical workflow in which archived educational records are harmonised, linked at the participant level, checked for measurement consistency, analysed as repeated observations, and examined across contextual subgroups. Visual analytics is used as a distribution-sensitive exploratory layer rather than as decorative reporting: paired profiles, boxplots, and territorial displays are used to inspect overlap, dispersion, outliers, and spatially patterned heterogeneity before those patterns are interpreted alongside inferential tests. The objective is explanation and pattern identification rather than prediction. Because the research questions are theory-defined and the linked sample is moderate in size, machine-learning, clustering, and other predictive procedures were not introduced merely to increase computational complexity; they would not provide a more valid answer to the longitudinal questions addressed here.

3. MATERIALS AND METHODS

This section describes the methodological procedures used to examine teachers' perceived development across the six dimensions of didactic suitability and to relate these perceptions to the territorial implementation of the EDEM programme. It outlines the study design, participant profile, instrument, data preparation, statistical procedures, visualisation strategy, and ethical safeguards.

3.1 Study design and analytic logic

The study used a retrospective longitudinal design with a within-participant pre–post component and a complementary descriptive analysis of territorial participation. Quantitative self-assessment records were used to estimate change in perceived didactic suitability from the beginning (A) to the end (D) of each diploma cohort. Province-by-year participation records contextualised programme reach. Because there was no untreated comparison

group or random assignment, temporal change is interpreted as an association with the training period, not as a causal effect of EDEM.

Two analytical layers were retained. The first examined paired change across the six didactic-suitability dimensions and selected post-training distributions. The second examined the distribution of EDEM participation by province and year. The layers were integrated at interpretation rather than collapsed into a single index; observed participation patterns were not treated as evidence of unmeasured mechanisms such as facilitation intensity, professional networks, or institutional memory.

3.2 Dataset and cohort construction

Programme records defined an analytic cohort of 420 in-service mathematics teachers enrolled in EDEM across 2017, 2018, 2020, and 2021. Of these, 409 completed linkable A and D questionnaires (97.4% of the reported cohort), including 319 women and 90 men. Eleven participants voluntarily withdrew because completion of the diploma was not mandatory. Participants taught mainly in primary and lower-secondary education in public and semi-public schools and reported between five and more than twenty years of experience. The paired records represented Panamá, Panamá Oeste, Colón, Coclé, Herrera, Los Santos, Veraguas, Chiriquí, Bocas del Toro, and Darién. The indigenous comarcas were represented in broader programme records but not in the paired item-level subset used for change-score comparisons.

EDEM was implemented through collaboration between academic institutions and Panama's Ministry of Education. Its curriculum combined theoretical study, practice-oriented workshops, reflective analysis, and classroom implementation projects. The 2020 cohorts operated in hybrid or online mode. Most participants had not previously worked with a systematic framework such as the OSA, so the programme also introduced a common vocabulary for analysing instruction.

3.3 Instrument harmonisation and measurement reliability

Each archived cohort workbook contains an A sheet administered before the diploma and a D sheet administered after it (2017-A/2017-D, 2018-A/2018-D, 2020-A/2020-D, and EDEM-A/EDEM-D for 2021). The questionnaire measured teachers' self-perceptions across the six OSA criteria: epistemic, cognitive, interactional, affective, ecological, and mediational suitability. The 2017–2020 workbooks used a 63-item version and the 2021 workbook used a 90-item version; both stored responses on an ordinal 0–3 scale. Item wording and dimensional labels were reviewed item by item, producing a common core of 45 conceptually equivalent items: 7 epistemic, 9 ecological, 8 cognitive, 6 affective, 5 mediational, and 10 interactional items. Items without a defensible cross-version match were excluded. Exact participant identifiers linked 409 complete A–D pairs: 12 from 2017, 150 from 2018, 103 from 2020, and 144 from 2021.

Content validity was supported by explicit alignment with OSA indicators and expert review by mathematics educators. Cronbach's α and McDonald's ω were computed separately for each dimension and time point in the 45-item common core (n = 409). Omega was estimated from a one-factor principal-axis solution based on Pearson correlations. Table 1 reports post-training coefficients and 95% percentile bootstrap confidence intervals from 2,000 resamples stratified by questionnaire version; baseline coefficients are reported in the text to show the precision of the change analysis. Because α and ω are unchanged by a common positive linear transformation, coefficients were calculated from the archived 0–3 item codes. No missing item values occurred within the 409 retained pairs.

3.4 Longitudinal linkage and data processing

The reported programme cohort was N = 420 and the paired analysis retained 409 participants. An audit of the raw workbooks found 430 A-sheet rows and 409 D-sheet rows: all 409 D records linked to an A record, while 21 A records had no D counterpart. Programme information identifies 11 voluntary withdrawals from the reported cohort because the diploma was not compulsory. The retained files do not contain an exclusion log that explains the status of the other 10 baseline rows relative to the reported N = 420; they were therefore not assigned an invented status and

were excluded from inference. Analyses were restricted to the 409 unambiguous A–D pairs. Gender, age, institution, and province codes were decoded with the supplied codebook. Common-core dimension means were calculated on the original 0–3 scale and linearly transformed to 1–5 for presentation using 1 + (4/3)x; this transformation does not alter test statistics or reliability coefficients.

3.5 Within-teacher change analysis

Within each dimension, scores A and D were compared with paired Wilcoxon signed-rank tests. Effect sizes were reported as paired rank-biserial correlations and standardised mean change (Cohen's dz). This within-teacher layer was treated as the primary longitudinal analysis because it preserves the dependency between each participant's pre- and post-training observations.

3.6 Territorial, cohort-year, and gender heterogeneity analysis

Individual change scores (D − A) were compared by province and cohort year with Kruskal–Wallis tests and epsilon-squared (ε^2), and by gender with Mann–Whitney tests and rank-biserial correlations. Provincial inference was restricted a priori to provinces with at least 20 linked pairs (seven provinces; n = 397); Bocas del Toro (n = 3), Darién (n = 1), and Veraguas (n = 8) were retained for descriptive display only. Significant provincial omnibus tests were followed by pairwise comparisons so that territorial heterogeneity was evaluated from individual longitudinal change rather than from static post-training means.

3.7 Multiple-comparison control and inferential safeguards

Holm correction controlled multiplicity across the six dimensional tests and across follow-up comparisons where applicable. The analyses estimate changes in self-perception and subgroup heterogeneity; they do not verify changes in teaching practice or student learning, establish longitudinal or cross-version measurement invariance, or identify causal programme, territorial, cohort, or gender effects. Cohort year remains confounded with implementation conditions and, in 2021, questionnaire version.

3.8 Data visualisation and exploratory analytics

Visual displays were retained only when they added distributional information not already carried by the statistical tables. The figures were treated as exploratory visual analytics for examining paired profiles, post-training dispersion, provincial heterogeneity, gender overlap, and historical provincial patterns. Their role was to reveal concentration, overlap, skew, outliers, and within-group variability that a single mean or aggregate index could conceal. Annual participation by province and year is therefore reported numerically rather than through additional graphics, keeping the visual evidence subordinate to the paired longitudinal analysis. Harmonisation, statistical analysis, and figure generation were conducted with reproducible Python scripts available from the corresponding author on reasonable request.

3.9 Ethics and data governance

Initial phases received approval from the Bioethics Committee of the University of Panama (CBUP/298/2019; CBUP/279/2021; CBUP/448/2021). For the 2025 analytical extension, oversight was transferred to CBI–USantander (M–087–2025) because of a conflict-of-interest concern in the Mathematics academic area; no new intervention was undertaken before resolution. Participation was voluntary, informed consent was obtained, and individual responses were anonymised. Only aggregate records are reported.

4. RESULTS

This section presents the findings in a sequence that moves from the overall national profile to dimensional variability, territorial contrasts, participation patterns, and gender-related distributions. The analysis first identifies the dominant configuration of didactic suitability and then examines where greater dispersion and contextual heterogeneity become visible.

4.1 Paired pre–post change across all six dimensions

The paired common-core analysis shows positive change in teachers' self-reported didactic suitability across all six dimensions (Figure 1). On the 1–5 presentation scale, epistemic suitability increased from 3.86 to 4.25, ecological from 4.06 to 4.40, cognitive from 3.91 to 4.35, affective from 3.67 to 4.41, mediational from 4.00 to 4.10, and interactional from 3.70 to 4.40. The largest mean changes were affective (Δ = 0.74) and interactional (Δ = 0.69); cognitive, epistemic, and ecological changes were moderate in descriptive magnitude (Δ = 0.45, 0.39, and 0.34), and mediational change was small (Δ = 0.10). These values describe increased self-ratings during the training period, not direct observations of improved teaching.

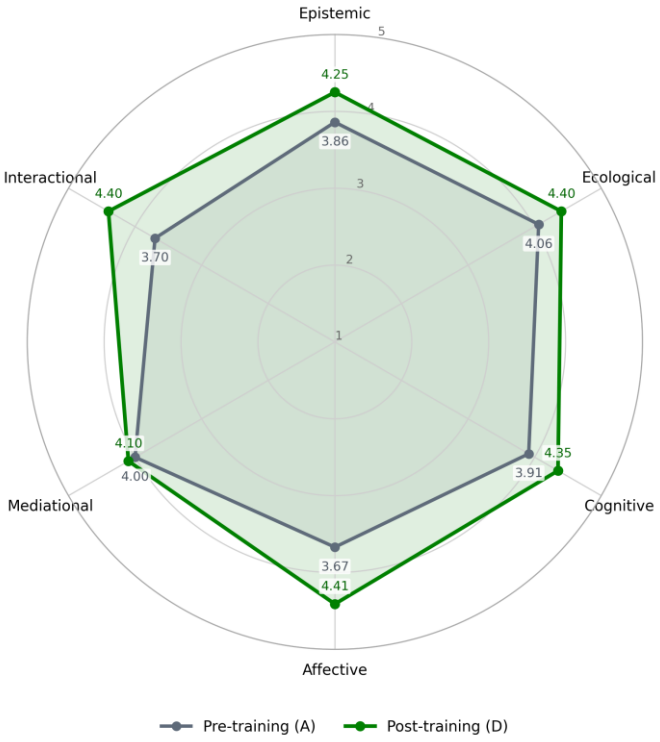

Figure 1. Paired pre- and post-training mean profiles in the harmonised 45-item common core (n = 409).

4.2 Dimension-specific reliability of the harmonised item-level subset

Post-training internal consistency was strongest for interactional (α = .862; ω = .865) and ecological suitability (α = .832; ω = .834); mediational suitability had the lowest post-training coefficients (α = .696; ω = .725) (Table 1). At baseline, coefficients ranged from α = .449 to .745 and ω = .478 to .745. Baseline epistemic (α = .449; ω = .478) and mediational scores (α = .535; ω = .594) were comparatively weak, and ecological and affective scores were moderate. Consequently, change estimates in those less reliable baseline dimensions require caution. The higher post-training coefficients must not be interpreted causally, and neither set of coefficients establishes longitudinal or cross-version measurement invariance.

Table 1. Post-training internal consistency of the harmonised OSA dimensions (n = 409).

Dimension	**k**	**α**	**95% CI**	**ω**	**95% CI**
Epistemic	7	.754	[.679, .809]	.765	[.694, .818]
Ecological	9	.832	[.800, .857]	.834	[.803, .859]
Cognitive	8	.781	[.741, .814]	.807	[.774, .835]

Affective	6	.793	[.749, .831]	.800	[.759, .836]
Mediational	5	.696	[.630, .754]	.725	[.670, .776]
Interactional	10	.862	[.828, .890]	.865	[.832, .893]

Note. CI = percentile bootstrap confidence interval based on 2,000 resamples stratified by questionnaire version. McDonald's ω was estimated from a one-factor principal-axis solution.

4.3 Within-participant change in the harmonised common core

Paired Wilcoxon tests confirmed positive changes in all six self-reported dimensions after Holm correction (Table 2). Rank-biserial effects were largest for interactional (r = .765) and affective suitability (r = .739), followed by epistemic (r = .612), cognitive (r = .556), and ecological suitability (r = .536). Mediational suitability showed a small positive effect (r = .134; adjusted p = .031). The convergence of mean change, signed-rank tests, and paired effect sizes supports a robust shift in perceived affective and interactional suitability, while the mediational result should be interpreted conservatively.

Table 2. Paired pre–post changes in perceived didactic suitability (n = 409).

Dimension	**Pre M (SD)**	**Post M (SD)**	**Mean Δ**	**W**	**Holm p**	**r_rb**	**dz**
Epistemic	3.86 (0.50)	4.25 (0.60)	0.39	13458.0	< .001	.612	.524
Ecological	4.06 (0.47)	4.40 (0.50)	0.34	17419.5	< .001	.536	.502
Cognitive	3.91 (0.67)	4.35 (0.49)	0.45	15556.0	< .001	.556	.510
Affective	3.67 (0.75)	4.41 (0.51)	0.74	9529.0	< .001	.739	.779
Mediational	4.00 (0.65)	4.10 (0.59)	0.10	25996.5	.031	.134	.111
Interactional	3.70 (0.65)	4.40 (0.51)	0.69	8969.5	< .001	.765	.803

Note. Means and standard deviations are presented on the transformed 1–5 scale; tests used the original 0–3 composites. $\Delta = D - A$; W = Wilcoxon signed-rank statistic; r_rb = paired rank-biserial correlation; dz = standardised mean change. Positive values indicate higher post-training self-ratings, not observed classroom effects.

4.4 Post-training concentration and dispersion

The historical post-training boxplots in Figure 2 add distributional context to the paired findings. Content knowledge and assessment concentrate toward the upper end of their retained scales, whereas didactic decision-making and classroom management extend across broader ranges. These displays do not measure individual change and are not numerically interchangeable with the 45-item common-core composites; they show only how post-training self-perceptions were distributed in the original descriptive analysis.

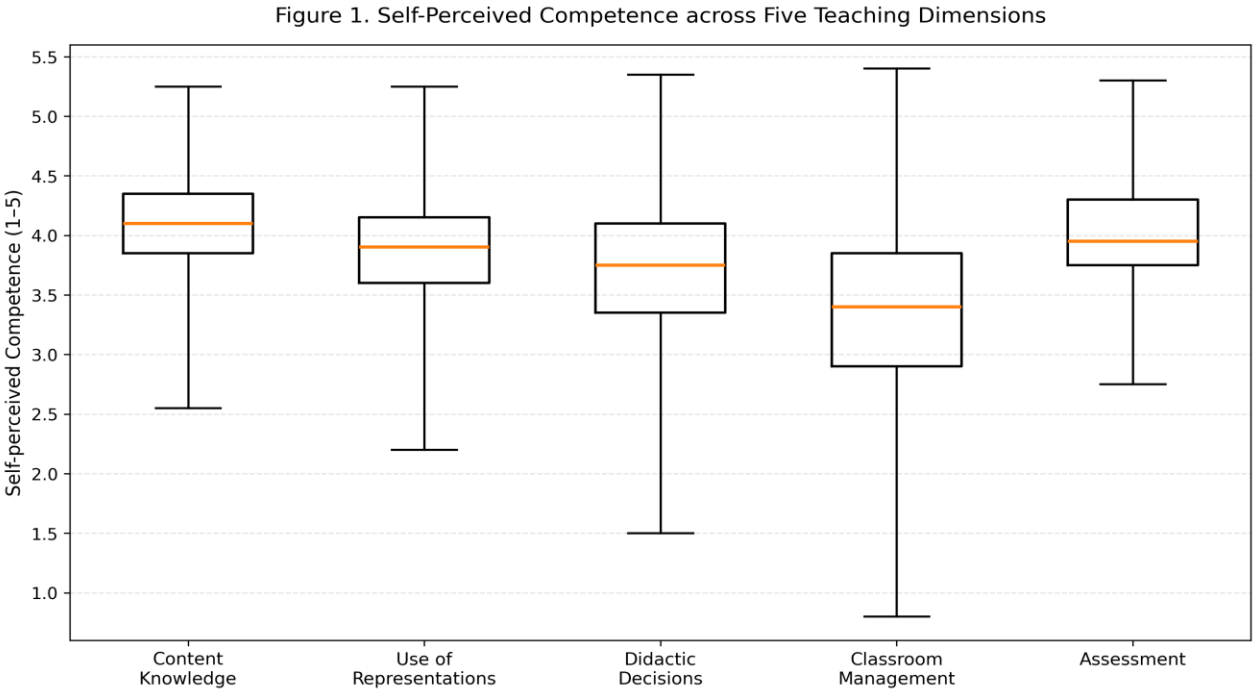

Figure 2. Boxplot distributions of teachers' self-reported development across five pedagogical subcompetencies.

Figure 3 preserves the original post-training dimensional distributions as boxplots. It should be read as a descriptive view of medians, interquartile ranges, and overall spread rather than as the source of the paired means in Figure 1 and Table 2. The paired analysis provides the stronger evidence about change; the boxplots remain useful for showing within-dimension heterogeneity among teachers' self-reports.

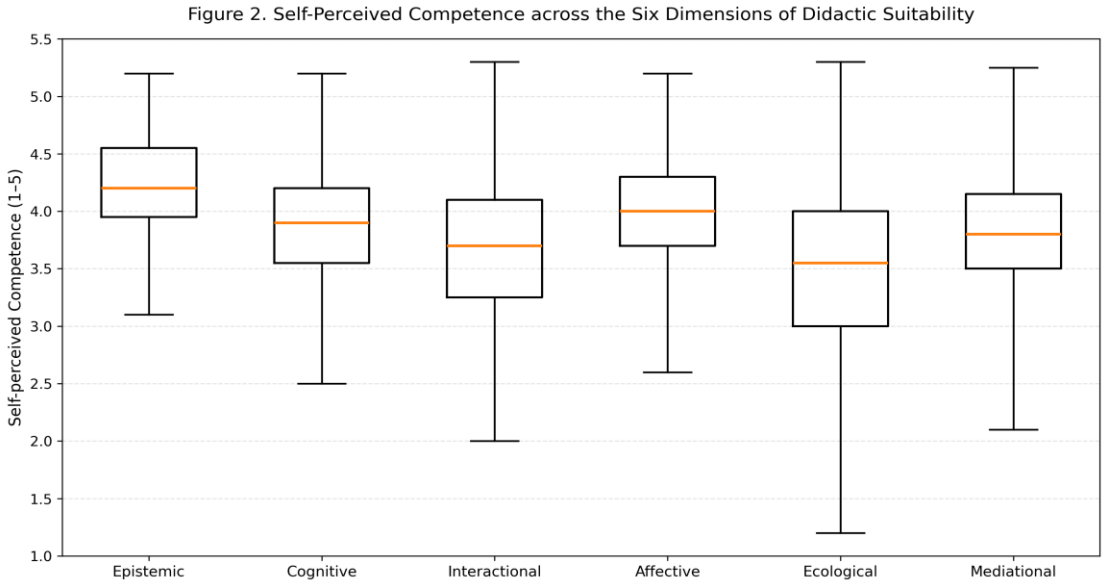

Figure 3. Boxplot distributions across the six dimensions of didactic suitability.

4.5 Provincial differences in paired change

Figure 4 presents province-level boxplots of post-training aggregate scores and shows substantial overlap among territories. It should be read as a descriptive display of medians, interquartile ranges, and heterogeneity, not as

a provincial ranking. Province-specific differences are evaluated below using individual D − A change scores from the harmonised pairs.

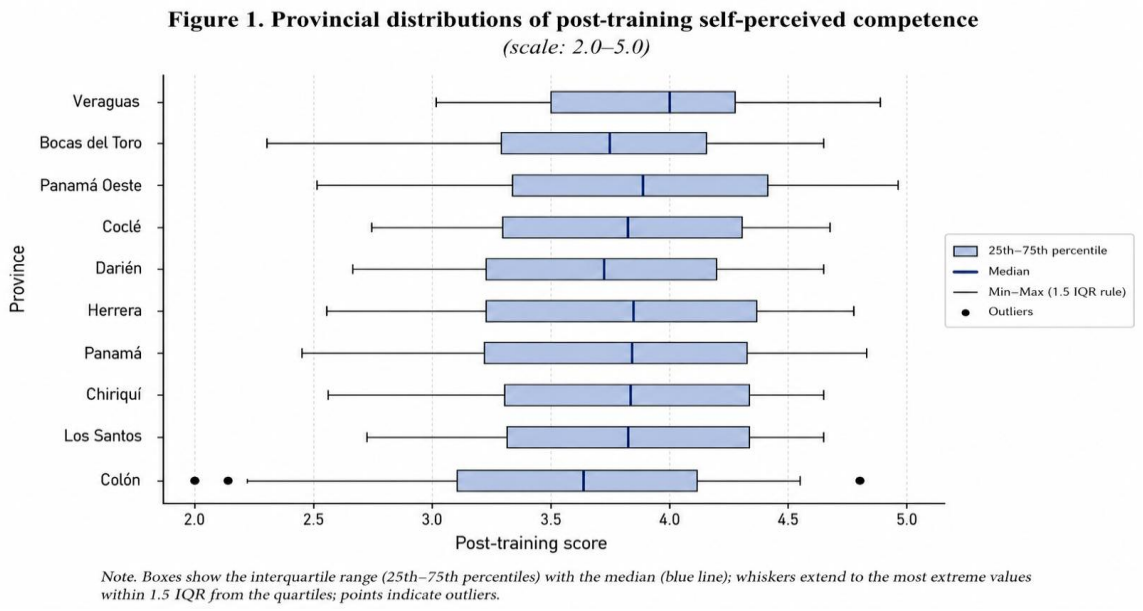

Figure 4. Province-level boxplots of post-training didactic competence across Panamanian provinces.

Figure 5 provides a second province-level boxplot representation with explicit median markers and comparable axes. Differences in medians, interquartile ranges, whiskers, and outliers show that a single provincial average would obscure substantial within-province variation. The display reinforces the need to examine dispersion alongside central tendency, especially in territories where connectivity, travel, school size, and access to professional communities differ markedly across local settings.

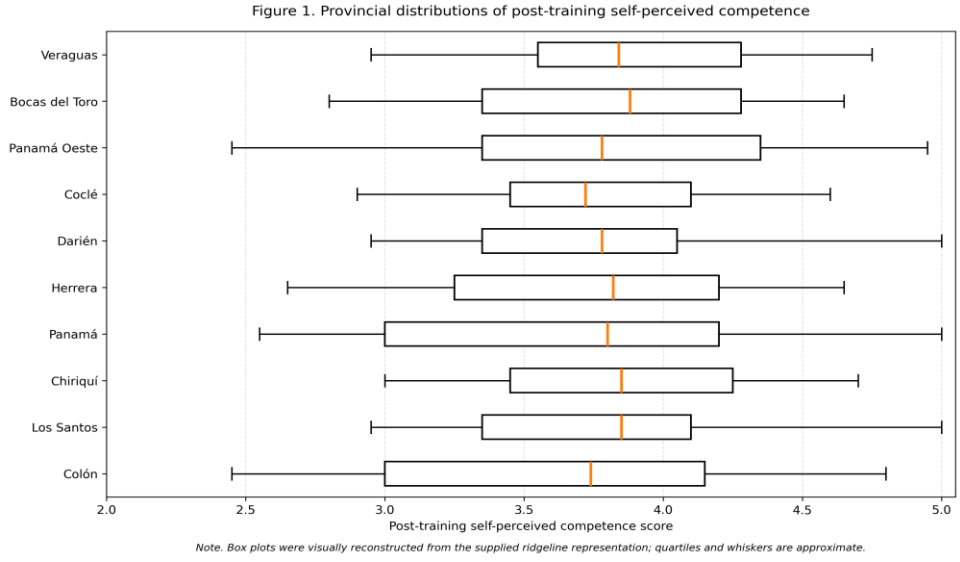

Figure 5. Province-level boxplots with median markers for post-training didactic competence.

Table 3. Mean paired change in perceived didactic suitability by province (n = 409).

Province	n	Epist.	Ecol.	Cogn.	Affect.	Mediat.	Interact.	Overall

Bocas del Toro†	3	+0.06	+0.00	−0.39	−0.30	+0.09	−0.04	−0.10
Coclé	21	+0.62	+0.20	+0.13	−0.23	+0.15	+0.06	+0.16
Colón	50	+0.42	+0.31	+0.64	+0.78	+0.31	+0.88	+0.56
Chiriquí	25	+0.26	+0.27	+0.53	+0.91	+0.14	+1.11	+0.54
Darién†	1	+1.14	+0.89	+1.00	+1.33	+1.60	+0.93	+1.15
Herrera	30	+0.32	+0.33	+0.46	+0.24	+0.44	+0.26	+0.34
Los Santos	27	+0.37	+0.30	+0.22	+0.06	+0.18	+0.16	+0.21
Panamá	204	+0.35	+0.40	+0.38	+0.92	−0.10	+0.77	+0.45
Panamá Oeste	40	+0.58	+0.30	+0.82	+1.18	+0.36	+0.89	+0.69
Veraguas†	8	+0.31	+0.15	+0.56	+0.47	+0.70	+0.60	+0.47

Note. Values are mean paired changes (D − A) on the transformed 1–5 scale. Positive values indicate higher post-training self-ratings. Overall is the unweighted mean of the six-dimensional changes within each participant. † n < 20: descriptive only and excluded from provincial inference; these small cells should not be ranked. No paired item-level cases were available for the comarcas.

Among the seven provinces with at least 20 paired cases (n = 397), change scores differed for affective suitability, $H(6) = 63.45$, Holm-adjusted $p < .001$, $\varepsilon^2 = .147$; mediational suitability, $H(6) = 20.37$, adjusted $p = .010$, $\varepsilon^2 = .037$; and interactional suitability, $H(6) = 44.04$, adjusted $p < .001$, $\varepsilon^2 = .098$. No adjusted provincial difference was detected for epistemic, ecological, or cognitive change (adjusted $p \geq .122$). Affective gains were comparatively pronounced in Panamá Oeste, Panamá, Chiriquí, and Colón, whereas Coclé showed a negative mean affective change. Interactional gains were largest in Chiriquí, Panamá Oeste, Colón, and Panamá. Representative Holm-adjusted contrasts included Panamá over Coclé in affective change ($p < .001$; $r = .728$) and Chiriquí over Coclé in interactional change ($p = .002$; $r = .678$). The mediational omnibus test was significant, but no individual provincial pair survived post-hoc Holm correction. These findings identify heterogeneous perceived change and do not justify ranking provinces or attributing differences to unmeasured territorial mechanisms.

4.6 Cohort-year variation in paired change

Change scores differed across the four cohort years for every dimension after Holm correction: epistemic, $H(3) = 9.48$, adjusted $p = .024$, $\varepsilon^2 = .016$; ecological, $H(3) = 11.33$, adjusted $p = .020$, $\varepsilon^2 = .021$; cognitive, $H(3) = 72.27$, adjusted $p < .001$, $\varepsilon^2 = .171$; affective, $H(3) = 162.22$, adjusted $p < .001$, $\varepsilon^2 = .393$; mediational, $H(3) = 121.07$, adjusted $p < .001$, $\varepsilon^2 = .292$; and interactional, $H(3) = 175.48$, adjusted $p < .001$, $\varepsilon^2 = .426$. The largest cohort-associated heterogeneity therefore occurred in interactional, affective, and mediational change. These differences are not interpreted as a temporal trend in programme effectiveness because cohort year also captures implementation conditions and, in 2021, a different questionnaire version.

4.7 Similar perceived changes across gender groups

Figure 6 presents the retained post-training distribution by gender. The paired change analysis provides the relevant longitudinal comparison: among 319 women and 90 men, no gender difference in any of the six change scores survived Holm correction (adjusted $p \geq .126$; $|\text{rank-biserial } r| \leq .159$). The boxplots are therefore used to visualise dispersion and overlap, not to imply equivalence of programme benefit.

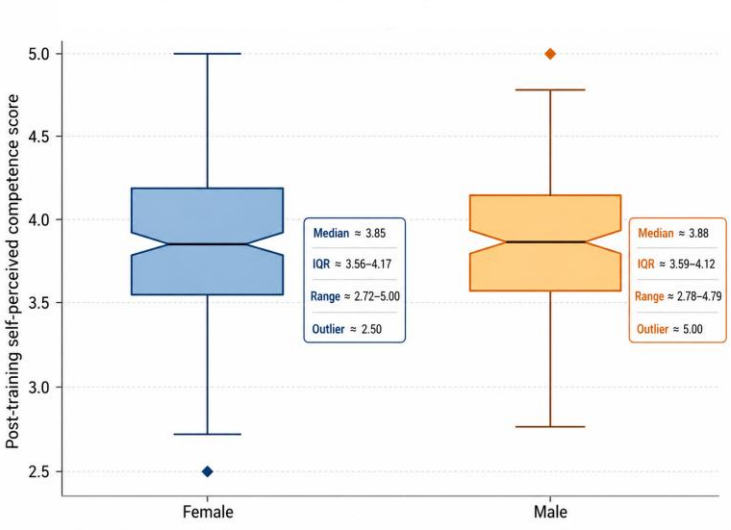

Figure 6. Notched boxplots of post-training didactic competence by gender.

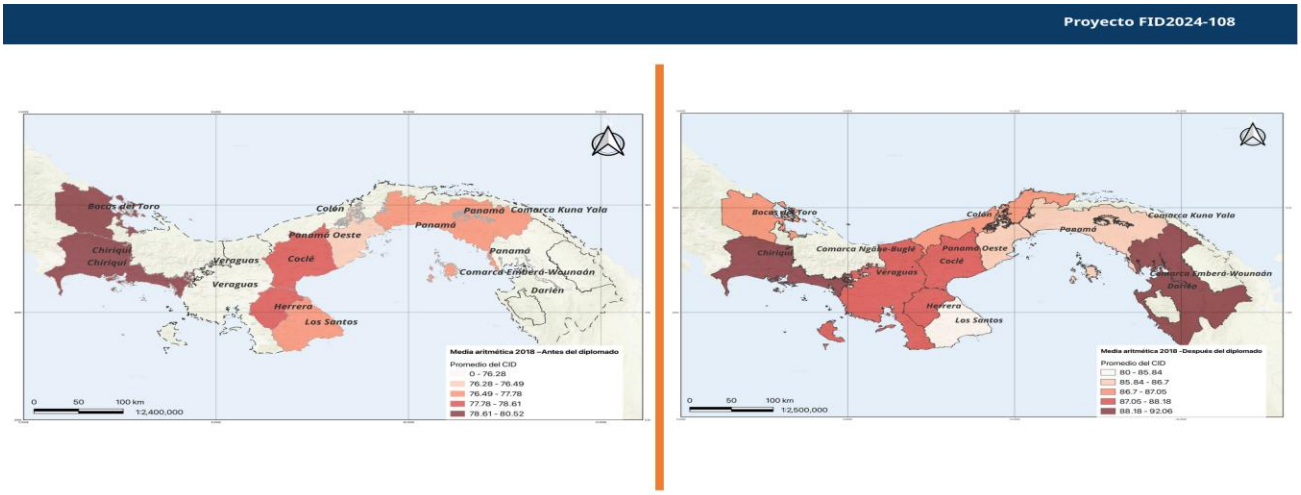

Figure 7. Provincial mean scores on the Teacher Competency Index after participation in the EDEM professional development programme.

The provincial means in Figure 7 are historical aggregate post-training scores and should not be numerically combined with the harmonised common-core changes in Table 3. The aggregate figure describes retained provincial index levels, whereas Table 3 documents within-participant changes. The paired data support increased self-ratings in several dimensions but do not establish that EDEM compensated for territorial inequality, because no comparison group or direct measure of territorial mechanisms was available.

Table 4. Summary of exploratory change-score comparisons in the harmonised subset.

Comparison	Dimension(s)	Statistic	Holm p	Effect size
Province (7 groups)	Affective	$H(6) = 63.45$	$< .001$	$\varepsilon^2 = .147$
Province (7 groups)	Mediational	$H(6) = 20.37$	.010	$\varepsilon^2 = .037$
Province (7 groups)	Interactional	$H(6) = 44.04$	$< .001$	$\varepsilon^2 = .098$
Province (7 groups)	Other 3 dimensions	$H = 3.38–13.15$	$\geq .122$	$\varepsilon^2 \leq .018$

Gender	All 6 dimensions	U = 12,776.5–16,639.5	≥ .126	|r| ≤ .159
Year (4 groups)	Epistemic	H(3) = 9.48	.024	ε^2 = .016
Year (4 groups)	Ecological	H(3) = 11.33	.020	ε^2 = .021
Year (4 groups)	Other 4 dimensions	H = 72.27–175.48	< .001	ε^2 = .171–.426

Note. All comparisons use individual D − A change scores. Province tests included only provinces with n ≥ 20 (n = 397); gender and year tests used n = 409. Representative adjusted provincial contrasts were Panamá > Coclé for affective change and Chiriquí > Coclé for interactional change. Results are exploratory and do not establish territorial or cohort causation.

The gender and territorial patterns also motivate questions beyond the study's initial focus on didactic suitability. Gender, family responsibilities, school conditions, and territory may interact in shaping participation and perceived professional development, but these mechanisms were not measured in the retained records. They are therefore treated as hypotheses for targeted future investigation rather than explanations of the observed scores.

This interpretation is consistent with the historical perspective developed in the coordinated work of Tirado Villegas, Rivera Gómez, and Gómez García (2023), which shows that women have repeatedly had to overcome structural and symbolic barriers in science, education, and public life. These barriers include social beliefs, institutional practices, unequal recognition, and the persistent invisibility of women's educational and intellectual work. From this perspective, the slightly wider dispersion observed among women may motivate investigation of a broader diversity of participation conditions, including family responsibilities, access to professional support, and opportunities to translate learning into classroom practice; these mechanisms were not measured or tested in the retained dataset.

Accordingly, the gender findings should be interpreted as an absence of detectable differences in the paired change scores, not as evidence of equal programme benefit. Future studies should replicate this comparison prospectively with objective outcomes, interviews, explicit measures of relational conditions, and sufficient power for intersectional and territorial analyses.

5. DISCUSSION

The discussion is organised around five implications of the paired analysis: where perceived change was concentrated, which dimensions remained constrained, how interactional change can be interpreted, how provincial heterogeneity should be read, and what the longitudinal-territorial design contributes to evaluation of national professional development. This organisation shifts attention from static post-training profiles to within-teacher change and the conditions under which such change may be enacted.

5.1 Perceived gains were concentrated in interaction and affect

The clearest pattern was not uniform growth across all dimensions. Affective suitability increased by 0.74 points and interactional suitability by 0.69 points on the 1–5 presentation scale, with the largest paired rank-biserial effects. Epistemic, cognitive, and ecological self-ratings also rose, but less strongly, while mediational change remained small. The profile therefore suggests that EDEM was experienced most strongly in dimensions related to engagement, confidence, dialogue, and participation rather than in the material and technological conditions that support enactment. Because the outcomes are self-reports, this is evidence of perceived professional development, not verified classroom change.

Teacher perspectives are especially useful when they are treated as diagnostic evidence rather than as direct proxies for student outcomes. Idrus et al. (2026) used mathematics teachers' reports to identify conceptual and instructional needs in geometry, showing how practitioner judgements can reveal where reasoning-oriented pedagogy requires additional support. The present longitudinal analysis extends that diagnostic logic by identifying the didactic-suitability dimensions in which teachers perceived the greatest change. This interpretation also accords with affect

research that rejects a separation between cognition and emotion (McLeod, 1989; Hannula et al., 2004; Zan et al., 2006; Chacón-Rivadeneira et al., 2026) and with Sánchez Aguilar's (2022) view of affect as a transversal element of teaching practice.

Recent large-scale evidence also suggests that teacher self-efficacy is embedded in professional-learning structures rather than functioning as an isolated individual attribute. Wang et al. (2026), analysing 15,847 primary teachers across 16 education systems, reported that the association between professional learning communities and teacher self-efficacy operated through teacher collaboration, professional development, and job satisfaction, with meaningful cross-system variation in the relative strength of those pathways. This provides a useful comparative lens for EDEM's affective gain: the increase is compatible with strengthened professional confidence and collaborative experience, but the present archive did not measure those mediating mechanisms and therefore cannot attribute the observed affective change to them.

Rodríguez-Nieto et al. (2026) help specify the epistemic content of this development. Their study of frequency distribution tables shows that successful problem solving depends on connecting representations, procedures, concepts, and contexts. In EDEM, the epistemic self-rating increased from 3.86 to 4.25 ($\Delta = 0.39$; $r = .612$), which is best interpreted as greater perceived readiness to organise such mathematical relationships, not as verified mastery or evidence that students encountered richer tasks. Future observational work should test whether these self-perceptions correspond to multiple representations, argumentation, and contextual connections in actual lessons.

5.2 Mediational suitability marks the implementation boundary

Reliability results require a differentiated reading of change. Post-training coefficients were acceptable to strong, whereas baseline reliability was weak for epistemic and mediational suitability and moderate for some other dimensions. The small mediational change must therefore be interpreted with particular caution. Rather than treating the six dimensions as equally precise or equally responsive, the findings point to both a measurement boundary and an implementation boundary: some constructs were less stable at baseline, and the resource-dependent dimension changed least.

Mediational suitability is where the distinction between professional understanding and enactment becomes most visible. Teachers may leave a programme with stronger criteria for analysing instruction while still working under constraints of time, connectivity, devices, technical assistance, or access to collaborative support. This interpretation is consistent with Zuo et al. (2025), whose review of teacher collaboration through digital tools concluded that sustained technology-enabled collaboration depends on digital infrastructure, teachers' digital competence, and organisational support rather than on tool availability alone. The EDEM result therefore should not be read as weak professional commitment; it identifies the dimension most dependent on conditions outside the individual teacher.

Cortés et al. (2026) show that didactic-suitability criteria can organise reflection on innovations such as gamification, enabling teachers to examine mathematical and pedagogical consequences rather than adopt techniques uncritically. The EDEM findings are compatible with this formative potential because self-ratings increased across all six dimensions, particularly affective and interactional suitability. Yet reflective confidence and implementation capacity are not identical. Teachers may report greater understanding of suitable processes while still lacking resources, time, or institutional support to enact them consistently.

This distinction prevents deficit interpretations of under-served provinces. Wider ecological and mediational distributions should not be attributed automatically to weaker teachers. They may instead indicate unequal resource ecologies and inconsistent opportunities to sustain practices introduced during training. Pino-Fan et al. (2017) similarly emphasise that tools and representations acquire didactic value only through their integration into mathematical meaning-making. Hardware availability alone cannot secure mediational suitability.

5.3 Interactional change as professional agency

Interactional suitability showed the strongest paired rank-biserial effect (r = .765), increasing from 3.70 to 4.40. This matters because interaction is the site where mathematical meanings, student reasoning, feedback, and participation are negotiated. The result suggests a stronger perceived capacity to elicit explanations, respond to errors, compare strategies, and organise participation. It does not establish that classroom discourse changed, but it identifies a concrete target for follow-up observation and coaching.

Ledezma et al. (2026) illuminate this mechanism through practical argumentation. Prospective teachers justified how and why application problems should be transformed into modelling problems, revealing that pedagogical decisions are structured by reasons, anticipated outcomes, and contextual considerations. In in-service development, comparable argumentation should occur both among teachers and with students. Mentoring focused on questioning, warrants, counterexamples, and comparison of models could consolidate the perceived interactional gain documented here and, crucially, permit its enactment to be evaluated through observed classroom discourse.

A complementary interpretation is offered by Zhang et al. (2026), who describe mathematical-modelling pedagogy as requiring a reconstruction of the teacher's role from knowledge transmitter toward facilitator, context designer, and supporter of students' modelling thinking, teamwork, and reflection. That role shift is analytically consistent with the demands represented by interactional suitability because it depends on orchestrating explanation, questioning, collaboration, and reflective dialogue. Accordingly, EDEM's strong interactional effect can be read as increased perceived readiness for these interactionally demanding professional roles, while remaining distinct from evidence that such role reconstruction was observed in classroom practice.

The variability in classroom management and didactic decision-making shown in Figure 2 reinforces this point. Real-time interaction is less scriptable than content analysis. It requires adaptive expertise, repeated rehearsal, observation, and feedback. A programme that wishes to convert reflective knowledge into interactional consistency should include video-based lesson analysis, peer observation, rehearsal of discussion moves, and coaching cycles that continue after formal coursework.

5.4 Territorial heterogeneity without territorial ranking

Provincial differences were selective rather than universal. Affective, mediational, and interactional change varied across adequately represented provinces, whereas epistemic, ecological, and cognitive change did not survive multiplicity correction. This pattern argues against a simple high-versus-low province interpretation. Recent evidence from rural-school reform reaches a compatible conclusion: Yoke et al. (2026) found that the feasibility of blended reform depended on teacher readiness, institutional support, infrastructure, and contextual planning. Applied cautiously to EDEM, the comparison supports interpreting territory as a set of enabling or constraining conditions rather than as a label for teacher quality.

The policy implication is to monitor the conditions under which professional learning can be enacted, not merely the geographical distribution of participants. Province-level change estimates should therefore be read together with local information on connectivity, facilitation, protected time, school support, and continuity of professional communities. Such measures would allow future evaluations to test mechanisms that the present archive can only hypothesise.

The individual-level change tests provide the empirical basis for this cautious territorial reading. Panamá, Panamá Oeste, Colón, and Chiriquí showed comparatively larger affective or interactional gains in selected contrasts, but unequal cell sizes and the exclusion of very small provincial groups preclude a territorial league table. No mediational pair survived post-hoc Holm correction, even though the omnibus test was significant. The appropriate conclusion is heterogeneity in perceived change, not evidence of province-level programme effectiveness.

Year-specific participation records remain useful as implementation context, but they should not be used to assign stable labels such as "high-access" or "low-access" province. The paired results show that professional learning varies by dimension, while programme reach varies over time. Future monitoring should therefore update territorial evidence regularly and distinguish access, perceived change, and observed enactment as separate indicators.

5.5 What the longitudinal-territorial design adds

The main contribution is methodological and interpretive. Methodologically, the study moves beyond a post-training profile by harmonising item-level records, preserving teacher-level linkage, estimating within-teacher change across six dimensions, and combining inferential statistics with distribution-sensitive visual analytics. Interpretively, it treats province and cohort as contexts for heterogeneity rather than as causal explanations. This combination yields a more defensible account of national professional development than a single aggregate score or a static map of post-training competence.

The visual-analytic layer is consequential because it changes what can be detected in the archive. Paired profiles make dimensional asymmetry visible, boxplots expose within-group dispersion and overlap, and provincial displays show why territorial averages should not be converted into league tables. Read together with multiplicity-controlled tests, these displays help distinguish visually salient patterns from statistically supported heterogeneity and reduce the risk of overinterpreting unstable subgroup contrasts.

For international readers, the Panamanian case illustrates a scalable evaluation principle: national programmes should be examined simultaneously for common professional learning and unequal enactment conditions. Dimensions closely tied to reflection and interaction may shift even when material mediation changes little. This distinction is especially relevant in geographically diverse systems that use hybrid delivery to extend reach.

Taken together with recent studies, the findings situate teacher development at the intersection of pedagogical reasoning, professional collaboration, and context-sensitive implementation. Idrus et al. (2026) show how teacher perspectives can diagnose reasoning and resource needs in mathematics; Zuo et al. (2025) highlight the infrastructure and competence required for digital collaboration; and Yoke et al. (2026) demonstrate that reform feasibility in rural settings depends on readiness and institutional conditions. The present study adds longitudinal evidence that these contextual issues can coexist with substantial perceived growth in some dimensions and modest change in others.

6. CONCLUSIONS AND IMPLICATIONS

This study developed a data-driven longitudinal assessment of perceived professional learning during EDEM and examined whether individual change scores varied across territorial and demographic contexts. Among 409 paired participants from the reported analytic cohort of 420, self-ratings increased in all six harmonised dimensions. The largest changes were affective and interactional; epistemic, cognitive, and ecological changes were also positive, while mediational change was small. Affective, mediational, and interactional change varied across adequately represented provinces, whereas gender differences in change were not detected after multiplicity correction. These findings describe longitudinal changes in self-perception and must not be equated with verified changes in teaching or student learning.

The scientific implication is that didactic suitability should be analysed as a configuration of interdependent dimensions using evidence that preserves change over time. The results also demonstrate the value of longitudinal, data-driven analysis for identifying patterns of professional learning that may remain hidden when programme evaluation relies solely on aggregate post-training scores. Pairing reliability analysis, within-teacher comparisons, multiplicity-controlled subgroup tests, and distribution-sensitive visual analytics provides a stronger basis for evaluating scaled professional development than static post-training profiles alone.

The educational implication is that the perceived affective and interactional gains should be consolidated while mediational support is strengthened. Future cycles should combine lesson-rehearsal protocols, analysis of student reasoning, observation of classroom discourse, and coaching focused on questioning and argumentation with reliable access to time, materials, connectivity, and technical support. Mathematical connections, in the sense developed by Rodríguez-Nieto et al. (2026), can serve as observable targets for both task design and classroom interaction.

For public policy, uniform provision should be replaced by differentiated support within a common national framework. Priority actions include provincial professional-learning hubs; offline and low-bandwidth resource

packages; technical and pedagogical support for digital mediation; protected time for collaborative planning; and an equity dashboard combining participation continuity with ecological, mediational, and interactional indicators. Indigenous and remote jurisdictions should participate in the co-design of materials so that ecological suitability is built into the programme rather than assessed only after implementation.

Future research should prospectively replicate the paired design with a comparison group and triangulate self-perceptions with classroom observation, teacher interviews, student work, and learning outcomes. It should test longitudinal and cross-version measurement invariance, model school- and province-level variation, document reasons for non-completion with a complete participant flow, and directly measure infrastructure, facilitation, professional networks, and institutional support. Comparative studies across Latin American national programmes would clarify which dimensions travel well through common curricula and which require locally governed infrastructures.

The study has six central limitations. First, A and D assessments permit within-participant analysis, but the absence of a comparison group or random assignment precludes causal attribution to EDEM. Second, all outcomes are teachers' self-perceptions; classroom practices, student learning, and actual changes in teaching were not directly evaluated. Third, 409 of the reported cohort of 420 completed both assessments; the 11 voluntary withdrawals occurred because the diploma was not mandatory, so self-selection and attrition bias remain possible. Fourth, the raw A sheets contain 430 baseline rows, 10 more than the reported cohort, and the retained files provide no exclusion log or status for those additional records; the analysis therefore used only the 409 unambiguous A–D pairs. Fifth, harmonisation retained 45 common items, baseline reliability was weak for some dimensions, and longitudinal and cross-version measurement invariance was not established. Sixth, provincial cells were highly unequal, three provinces had fewer than 20 pairs, no paired cases were available from the comarcas, and a stable invitation denominator was unavailable. Accordingly, the findings support conclusions about changes in reported perceptions, not programme effectiveness, territorial causation, population-level gender equivalence, or unmeasured institutional mechanisms.

The broader conclusion is that scalable teacher education cannot be evaluated only by what it teaches. It must also be evaluated by where, with what resources, through which professional relationships, and under what conditions teachers are able to turn that learning into mathematics instruction.

Academic Contribution. This study extends the application of didactic suitability from lesson-level analysis to the territorial evaluation of a nationwide mathematics teacher-development programme. Its methodological contribution is an auditable data-driven workflow that combines cross-version item harmonisation, participant-level longitudinal linkage, reliability estimation, non-parametric within-teacher inference, multiplicity-controlled heterogeneity testing, and distribution-sensitive visual analytics. This workflow identifies variation that aggregate national indicators can conceal while preserving a clear boundary between perceived change, territorial heterogeneity, and causal programme impact.

7. Author contributions

Luisa Morales-Maure: Conceptualisation, methodology, project administration, funding acquisition, supervision, writing original draft, and writing review and editing. Orlando García-Marimón: Formal analysis, software, data visualisation, data curation, and writing review and editing. Rosa Durán-González: Investigation, validation, data curation, and writing review and editing. Laura Franco Sánchez: Statistical validation, interpretation of results, visualization review, and writing review and editing. All authors reviewed and approved the final manuscript and accepted responsibility for the integrity of the work.

Conflicts of Interest: The authors declare no financial, commercial, professional, or personal relationships that could have influenced the design, conduct, analysis, interpretation, or reporting of this study. The authors further declare that the transfer of ethical oversight to CBI–USantander was implemented as an institutional safeguard to

address a potential conflict-of-interest concern within the academic area and did not affect the independence of the research, the analysis of the data, or the conclusions presented in the manuscript.

Funding: This study received financial support from PID2021-127104NB-I00; SENACYT–Panama through projects VIP-01-90-2024-01, VIP-01-90-2024-02, VIP-01-90-2024-03, and FID2024-108; institutional project VIP-040405-2025-008. The funders had no role in design, execution, analysis, interpretation, manuscript preparation, or publication decision.

Acknowledgments: The authors thank the mathematics teachers who participated in the EDEM programme and shared their professional reflections and self-assessment data. The authors also acknowledge the academic, administrative, technical, and territorial support provided by the Universidad de Panamá and the collaborating institutions involved in programme implementation, data management, and ethical oversight.

8. Data availability

De-identified province-level aggregate data, the instrument, codebook, and analysis scripts will be deposited in an open repository upon acceptance. Until then, they are available from the corresponding author on reasonable request, subject to ethical and institutional restrictions. Individual-level responses cannot be shared because of confidentiality obligations.

REFERENCES

1. Calle Palomeque, C. E., Breda, A., Font, V., & Arias, A. L. (2024). Análisis de la reflexión en la enseñanza de las funciones racionales, utilizando criterios de idoneidad didáctica. *Acta Latinoamericana de Matemática Educativa*. *37(2), 1–11*.
2. Campo-Meneses, K. G., Font Moll, V., & García-García, J. (2024). Criterios que orientan la práctica de un profesor al enseñar las funciones exponencial y logarítmica. *Educação e Pesquisa, 50*(1), 61. e267455. https://doi.org/10.1590/S1678-4634202450267455es
3. Campuzano, T. S., Font, V., & Breda, A. (2024). Criterios de idoneidad didáctica en la argumentación práctica de docentes sobre el diseño de una unidad didáctica. *Acta Latinoamericana de Matemática Educativa*. 37(2).
4. Carvajal, S., Rodríguez, J. G., Moll, V. F., & Sánchez, A. (2024). Assessing preservice teachers' digital creativity in mathematics. *REDIMAT: Journal of Research in Mathematics Education*. 13(1), 38-58.
5. Cortés, A., Sánchez, A., Breda, A., & Font, V. (2026). Didactic suitability criteria to organise mathematics teachers' reflection aimed at improving instructional processes in mathematics: The case of gamification and game-based learning. In C. Fernández, S. Llinares, A. Gutiérrez, & N. Planas (Eds.), Innovations in initial mathematics teacher education: New developments and insights (pp. 247–272). Springer. https://doi.org/10.1007/978-3-032-01520-4_12
6. D'Amore, B. (2024). En torno a la publicación del libro de Juan Díaz Godino: Enfoque ontosemiótico en educación matemática. Fundamentos, herramientas y aplicaciones. *Educación matemática*, *36*(3), 300-310.
7. Espinoza Vargas, S., & Rico, L. (2022). Criterios de idoneidad didáctica para valorar la formación inicial del profesorado. *Revista Paradigma*, 43(1), 49–74.
8. Font, V., Breda, A., Sala-Sebastià, G., & Pino-Fan, L. R. (2024). Future teachers' reflections on mathematical errors made in their teaching practice. *ZDM–Mathematics Education*, 56, 1169–1181. https://doi.org/10.1007/s11858-024-01574-y
9. Font, V., Godino, J. D., & Gallardo, J. (2013). The onto-semiotic approach to research in mathematics education. *The Mathematics Enthusiast*, 10(1–2), 583–622.
10. Godino, J. D. (2009). Categorías para la descripción y evaluación de la idoneidad didáctica de procesos de enseñanza y aprendizaje de las matemáticas. *Educación Matemática*, 21(3), 97–118.
11. Godino, J. D. (2011). Indicadores de idoneidad didáctica de procesos de instrucción matemática. *Paradigma*, 32(2), 13–36.
12. Godino, J. D., Batanero, C., & Font, V. (2007). *Fundamentos de la enseñanza y el aprendizaje de las matemáticas para maestros*. Universidad de Granada.
13. Gómez-Flores, A., Méndez-López, J., & Ortiz-Navarro, H. (2023). Teacher training and its classroom impact from a territorial perspective. *Revista Latinoamericana de Estudios Educativos*, 53(1), 45–68.
14. Hannula, M. S. (2012). Exploring new dimensions of mathematics-related affect: Embodied and social theories. *Research in Mathematics Education,* 14(2), 137–161. https://doi.org/10.1080/14794802.2012.694281
15. Hannula, M. S., Evans, J., Philippou, G., & Zan, R. (2004). Affect in mathematics educationExploring theoretical frameworks. In M. J. Høines & A. B. Fuglestad (Eds.), *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 1, pp. 107–136). PME.

16. Idrus, H., Zulnaidi, H., & Abdul Rahim, S. S. (2026). A needs analysis for developing a geometry module to enhance higher-order thinking skills in primary mathematics education. Pertanika Journal of Social Sciences & Humanities, 34(4), 1677–1692. https://doi.org/10.47836/pjssh.34.4.01
17. Chacón-Rivadeneira, K., Morales-Maure, L., & Fernández-Lame, F. (2026). Didactic-Mathematical Knowledge and Instructional Design in Blended Mathematics Teaching Under Conditions of Unequal Digital Access: A Multiple Case Study in Central America. *International Journal of Special Education*, 41(12s), 467–482.
18. Leder, G. C., & Forgasz, H. J. (2006). Affect and mathematics education. In A. Gutiérrez & P. Boero (Eds.), *Handbook of research on the psychology of mathematics education* (pp. 403–427). Sense Publishers.
19. Ledezma, C., Sol, T., Sánchez, A., & Font, V. (2026). Practical arguments of prospective primary education teachers in mathematical modelling problems. *Education Sciences*, 16(1), 118. https://doi.org/10.3390/educsci16010118
20. McLeod, D. B. (1989). Beliefs, attitudes, and emotions: New views of affect in mathematics education. In D. B. McLeod & V. M. Adams (Eds.), *Affect and mathematical problem solving: A new perspective* (pp. 245–258). Springer.
21. Godino, J. D., Burgos, M., & Gea, M. M. (2022). The onto-semiotic Approach in Mathematics Education. Analysing objects and meanings in Mathematical Practice. In *Advances in the anthropological theory of the didactic* (pp. 51-60). Cham: Springer International Publishing.
22. Pino-Fan, L. R., Font, V., & Breda, A. (2017). An onto-semiotic study of the teaching of mathematical definitions. *Educational Studies in Mathematics*, 94(3), 275–293.
23. Richit, A., Tomkelski, M. L., & Breda, A. (2026). Aprendizagens Profissionais de Professores dos Anos Iniciais em Estudos de Aula. *Acta Scientiae*, 28(2).
24. Rico, L., & Lupiáñez, J. L. (2008). Mathematics teacher training and school practice. *UNO: Revista de Didáctica de las Matemáticas*, 47, 7–16.
25. Rodríguez-Nieto, C. A., Barrera-Pacheco, A., Cantillo-Rudas, B. M., Ballestas Rodríguez, M., Font Moll, V., Díaz-Levicoy, D., & Rodríguez-Vásquez, F. M. (2026). University students' mathematical connections in problem-solving on frequency distribution tables: A new view of connections in statistics education. *EURASIA Journal of Mathematics, Science and Technology Education*, 22(7), em2865. https://doi.org/10.29333/ejmste/18847
26. Sala Sebastià, G., Breda, A., Cortés da Silva, A., García Mora, E., & Font, V. (2026). *¿Cómo desarrollar el pensamiento computacional con BlueBots? Una experiencia de simulación con perspectiva de género.* En M. Díez Ojeda & M. Á. Queiruga Dios (Eds.), Reimaginando el aprendizaje para un mundo en cambio (pp. 203–207). Universidad de Burgos, Servicio de Publicaciones e Imagen Institucional. https://doi.org/10.36443/9791387585280
27. Sala-Sebastià, G., Breda, A., Seckel, M. J., & Farsani, D. (2026). Criterios asumidos por formadores de profesorado al reflexionar sobre una actividad de resolución de problemas con robots. *Bolema: Boletim de Educação Matemática*, 40, e230245. https://doi.org/10.1590/1980-4415v40a230245;
28. Sánchez Aguilar, M. (2022). Rumbo a SOMIDEM1: recuento del Precongreso SOMIDEM 2022. *Educación Matemática*, 34(3), 374–375. https://doi.org/10.24844/EM3403.15
29. Criterios de idoneidad didáctica en la argumentación práctica de formadores de profesores de matemática. (2024). Em *SciELO Preprints*. https://doi.org/10.1590/SciELOPreprints.9323
30. Villegas, G. T., Gómez, E. R., & García, L. E. G. (Eds.). (2023). *Saberes, enseñanza y poder: las mujeres rompiendo techos de cristal en el espacio público: siglos XVI al XXI*. Benemérita Universidad Autónoma de Puebla, Instituto de Ciencias Sociales y Humanidades" Alfonso Vélez Pliego".
31. Wang, H., Abdullah, Z., & Md Nor, A. R. (2026). Unpacking the mechanism: How professional learning communities influence primary school teachers' self-efficacy. *International Journal of Computer Information Systems and Industrial Management Applications*, 18, 13. https://doi.org/10.70917/ijcisim-2026-1787
32. Yoke, S. K., Mohd Rosley, S. N. A., Jamil, N. I., Abd Majid, F., & Pham, C. H. (2026). Teachers' perspectives on the feasibility of education reform in rural schools: An exploratory case study of the Versatile Learning Pathway model. *Pertanika Journal of Social Sciences & Humanities*, 34(3), 1295–1314. https://doi.org/10.47836/pjssh.34.3.09
33. Zan, R., Brown, L., Evans, J., & Hannula, M. S. (2006). Affect in mathematics education: An introduction. *Educational Studies in Mathematics*, 63(2), 113–121. https://doi.org/10.1007/s10649-006-9028-2.
34. Zhang, Y., Zulkifli, N. N., & Mohd Ayub, A. F. (2026). Teachers' roles and teaching behaviors in mathematical modeling pedagogy. *International Journal of Computer Information Systems and Industrial Management Applications*, 18, 11. https://doi.org/10.70917/ijcisim-2026-1792
35. Zuo, L., Krish, P., & Jain, J. (2025). Teacher collaboration through digital tools in the 21st century—Issues and challenges: A comprehensive literature review. *Pertanika Journal of Social Sciences & Humanities*, 33(1), 99–118. https://doi.org/10.47836/pjssh.33.1.05