Journal of Economich, Technology and Business (JETBIS)
Volume 4, Number 12 December 2025
p-ISSN 2964-903X; e-ISSN 2962-9330

Training the Trainer: Measuring the Long-Term Effectiveness of AI Upskilling Programs for Indonesian Educators

Elok Nur Affah Al Akromi
Universitas Islam Negeri Siber Syekh Nurjati, Indonesia
*Correspondence: eloknuraffah44@gmail.com

KEYWORDS:
AI upskilling; competency retention; Indonesian education; longitudinal study; teacher professional development

ABSTRACT

Indonesia's national artificial intelligence (AI) education roadmap (2025–2045) positions teacher AI competency as foundational to its Golden Generation vision, yet the long-term effectiveness of AI upskilling programs for Indonesian educators remains empirically unexamined. This study aims to evaluate the extent to which AI-related competencies are retained over time, identify key factors influencing sustained AI integration among teachers, and compare the effectiveness of different upskilling programs across regional contexts. This longitudinal mixed-methods study measured the durability of AI competency gains among 347 in-service K–12 teachers across urban, peri-urban, and remote regional strata over four measurement waves: pre-intervention, immediate post-intervention, six-month follow-up, and twelve-month follow-up. Participants were drawn from two nationally deployed AI upskilling programs Platform Merdeka Mengajar (PMM) and Microsoft Elevate and assessed using a validated four-subscale questionnaire, semi-structured interviews, and structured classroom observation checklists. Results revealed a consistent "spike-and-decay" trajectory across all groups, with teachers retaining an average of only 63.5% of post-intervention competency gains at the twelve-month wave. Microsoft Elevate participants demonstrated significantly superior retention compared to PMM participants, and remote-stratum teachers experienced the steepest decay. Post-intervention AI self-efficacy, peer collaboration frequency, and perceived program relevance emerged as the strongest predictors of sustained behavioral integration. Qualitative findings identified infrastructure deficits, institutional isolation, curricular misalignment, and confidence erosion as primary disinvestment mechanisms. The study concludes that sustainable AI teacher competency requires a systems-level response encompassing mandatory post-training support, differentiated regional resourcing, and curriculum-embedded AI learning communities.

INTRODUCTION

The rapid proliferation of artificial intelligence (AI) across virtually every sector of modern society has fundamentally altered the expectations placed upon educational institutions worldwide. As AI reshapes the nature of work, communication, and knowledge production, schools and universities face intensifying pressure to prepare students not merely to consume technology, but to critically engage with and direct it. This transformation has cast the role of the teacher in an entirely new light: educators are no longer simply knowledge transmitters but must now serve as guides through an increasingly algorithm-mediated learning environment. Lademann et al. (2026) assert that the pace at which AI has entered educational settings presents school systems with unprecedented challenges that differ fundamentally from those posed by previous digital tools, necessitating a swift response from the research community. Globally, governments, multilateral agencies, and private-sector actors are investing heavily in AI-in-education strategies, recognizing that the quality of a nation's future workforce depends significantly on the readiness of its present teaching force. As this global momentum builds, a critical question emerges: how well are teacher

training systems actually equipping educators to lead this transition and how durable are those gains over time?

Empirical evidence from multiple countries underscores the urgency of this question while simultaneously revealing troubling gaps between training provision and genuine preparedness. RAND Corporation (2025a) reported that in fall 2024, 48 percent of U.S. school districts had trained teachers on the use of AI a substantial 25-percentage-point increase from fall 2023 yet teachers who had received some training typically described it as a one-off experience rather than ongoing professional development. This pattern of episodic, non-sustained training is a recognized failure mode in teacher professional development more broadly, and its implications in the AI domain are particularly serious. Tan et al. (2025), in a systematic review of 95 research articles published between 2015 and 2024, found that 65% of studies examined the application of AI in teaching, while comparatively little emphasis was placed on the critical role of teachers and their professional development needs. Furthermore, a study on pre-service teachers found that future educators recognized AI's potential to enhance personalized learning and classroom efficiency but reported significant challenges in developing AI literacy, including insufficient training and a lack of institutional support ("Preparing Future Educators," 2025). These statistics collectively point to a structural problem: AI upskilling programs for educators are proliferating in quantity without commensurate investment in measuring their quality, durability, or pedagogical integration (Daher, 2025; Granström & Oppi, 2025; Guan et al., 2025).

Within the Indonesian context, this global tension between policy ambition and implementation reality is especially pronounced. Indonesia operates one of the world's most complex educational systems, encompassing approximately 53 million K–12 students and 3.4 million teachers dispersed across more than 17,000 islands, with profound disparities in infrastructure, digital connectivity, and professional development access between urban Java and rural eastern provinces. Nucamp (2025) notes that Indonesia's national AI roadmap, phased from 2025 to 2045, treats education as a headline priority across three pillars talent development, research and industrial innovation, and infrastructure and data and that the Ministry of Education has planned a rollout of AI and coding curricula as electives from elementary through vocational levels beginning in the 2025–2026 academic year. Yet for this vision to be realized in classrooms, teachers themselves must first be meaningfully upskilled. A systematic literature review on the deep learning revolution in Indonesian education found that implementation faces critical barriers, including digital infrastructure gaps, low teacher digital literacy, and an absence of comprehensive policy frameworks capable of sustaining behavioral change beyond initial training workshops (“Deep Learning Revolution: Transforming Indonesia’s Education System towards the Digital Era 2025,” 2025). Programs such as Microsoft Elevate and the government-supported Platform Merdeka Mengajar (PMM) represent significant investments in teacher AI capacity, but rigorous longitudinal evaluation of whether these programs produce lasting changes in educators' instructional practices remains largely absent from both policy discourse and the academic literature.

Several strands of prior scholarship provide important, if incomplete, foundations for this inquiry. Research on the PMM platform Indonesia's flagship digital professional development ecosystem has demonstrated promising short-term competency gains. A study on the implementation of the Merdeka Mengajar Platform and Google Workspace for Education found that together, both platforms contribute 72.3% to teacher competence, with PMM alone accounting for

61.1% of its variance within a specific Indonesian district ("The Implementation of the Merdeka Mengajar Platform," 2026). A qualitative multi-site case study involving teachers from junior high schools in Central Java further revealed that the platform has the potential to function as a catalyst for professional growth and pedagogical innovation when supported by strong institutional commitment and continuous mentoring ("From Reference to Practice," 2026). At the international level, Lademann et al. (2026) addressed teachers' attitudes toward AI following a dedicated training program and found that research in this area remains limited, highlighting the need for targeted professional development programs and systematic evaluation of their effectiveness in enhancing both teachers' AI literacy and their willingness to incorporate AI tools into their teaching practice. Meanwhile, research on pre-service teacher AI competency development found that participation in structured AI courses significantly improves partial AI competency, and that mere exposure to AI tools may be insufficient systematic training that integrates theoretical knowledge with contextualized learning holds greater promise for bridging the gap ("Improving Pre-Service Teachers'," 2025).

Despite this growing body of literature, a critical and conspicuous gap persists: the near-total absence of longitudinal research measuring whether AI upskilling gains among Indonesian teachers are sustained, deepened, or reversed over time. Existing studies both domestic and international are predominantly cross-sectional in design, capturing a snapshot of teacher competency or attitude at a single post-intervention point without tracking behavioral retention, pedagogical integration, or classroom impact across subsequent months or academic years. Sudrajat and Marlina (2023, as cited in Majority Science Journal, 2025) emphasize that continuous training based on simulated teaching, AI-supported instructional coaching, and professional learning communities is more effective in building teachers' technopedagogical capacity, yet without such sustained structures, teachers risk becoming passive users of technology rather than guides of technological intelligence for students. Moreover, a systematic review on challenges and best practices in training teachers to utilize AI found that AI utilization among teachers is associated with will, skill, tool, knowledge, and motivational factors (“Challenges and Best Practices in Training Teachers to Utilize Artificial Intelligence: A Systematic Review,” 2024) a multidimensional profile that single-session training evaluations cannot adequately capture. In the Indonesian context, this gap is compounded by the archipelagic geography, the heterogeneity of school contexts, and the complex political economy of education reform, all of which mediate how and whether training transfers into practice in ways that generic international models cannot account for.

The urgency of addressing this gap cannot be overstated. Indonesia's national AI education roadmap demands that trained teachers cascade AI competencies down to millions of students beginning immediately, yet the very foundation of that cascade the assumption that training produces durable educator capacity rests on untested ground. A perspective paper on integrating AI literacy into teacher education warns that without a foundation in deep AI understanding, educators risk unintentionally deepening the digital divide and disadvantaging marginalized students ("Integrating AI Literacy," 2025). In a country where approximately 104,000 schools remain digitally disconnected and only 27 percent of women currently work in the tech sector (Columbia University Teachers College, 2025), inequitable access to quality AI teacher training could further entrench existing structural disadvantages rather than ameliorate them. Furthermore, a study

exploring the influence of generative AI on teaching performance concluded that future studies should explore how AI tools can be effectively integrated into diverse teaching contexts, how they improve instructional delivery and assessment practices, and the long-term impacts on teacher development and student outcomes, noting that such research is vital for establishing evidence-based guidelines to ensure that AI adoption in education is practical and sustainable ("Transforming Education," 2025). With Indonesia investing trillions of rupiah in education technology platforms and teacher development initiatives, the cost of not evaluating long-term effectiveness extends well beyond missed opportunities it risks entrenching a cycle of reform without impact.

This study introduces novelty along three interconnected dimensions. First, it applies a longitudinal mixed-methods design combining pre-intervention, post-intervention, six-month follow-up, and twelve-month follow-up assessments to trace the durability of AI upskilling outcomes beyond the initial training window, a design that is virtually unprecedented in the Indonesian educational technology literature. Second, it moves beyond self-reported competency measures by incorporating classroom observation data and student learning outcome indicators, thereby linking teacher training inputs to pedagogical outputs in a causally traceable manner. Third, the study draws on Indonesia's unique multi-program landscape comparing corporate initiatives such as Microsoft Elevate with government-run PMM-based modules to generate comparative effectiveness insights that neither program has been subjected to independently. Lademann et al. (2026) note that existing research addressing teachers' attitudes toward AI and its potential for teaching practices remains limited, and no prior Indonesian study has compared AI training modalities longitudinally across diverse geographic and institutional contexts. By situating its findings within Indonesia's 2025–2045 AI education roadmap, the research also generates policy-transferable evidence relevant to comparable archipelagic and lower-middle-income country contexts globally.

The primary purpose of this study is to measure and explain the long-term effectiveness of AI upskilling programs for Indonesian K–12 educators, with attention to which program characteristics, institutional conditions, and individual teacher factors predict sustained competency retention and genuine pedagogical transformation. Specifically, the research aims to: (1) evaluate the extent to which AI-related knowledge, skills, and attitudes acquired through formal training programs are retained and applied by teachers over a twelve-month period following intervention; (2) identify the structural, motivational, and contextual moderators that differentiate teachers who meaningfully integrate AI into their practice from those who revert to pre-training behaviors; and (3) generate an evidence-based framework for designing sustained, contextually appropriate AI professional development programs suited to Indonesia's diverse educational geography. The EQUIP Framework encompassing Ethical Governance, Qualified Professional Learning, Unified Collaborative Partnerships, Implementation Readiness, and Progressive Adaptation represents one international model for empowering educators with the knowledge, skills, and ethical principles needed to navigate AI in education ("Integrating AI Literacy," 2025), and this research will test the applicability and sufficiency of such frameworks within Indonesia's specific political, infrastructural, and cultural context.

This study makes several distinct contributions to scholarship and practice. Theoretically, it extends current debates in teacher professional development by applying retention and transfer theories including Guskey (2002) model of teacher change and the TPACK framework to the

domain of AI literacy, testing whether the sequencing and structure of training predicts long-term behavioral integration. Empirically, it produces a rich longitudinal dataset from Indonesian schools across three regional typologies (urban, peri-urban, and remote), providing a comparative evidence base that is currently non-existent in the literature. Tan et al. (2025) highlight a significant imbalance in research focus, with far more attention given to the application of AI in teaching than to teachers' professional development needs; this study directly redresses that imbalance with rigorous original data. Practically, the research contributes a validated assessment instrument for measuring AI pedagogical readiness in the Indonesian context a tool adaptable for use by the Ministry of Education, provincial training bodies, and private EdTech providers. It also produces program design recommendations calibrated to the realities of Indonesia's digital infrastructure, including offline-capable and low-bandwidth training modalities suitable for remote schools.

The implications of this research extend across multiple levels of the Indonesian education system and beyond. At the policy level, findings will directly inform the Ministry of Education's ongoing rollout of AI-integrated curricula under Kurikulum Merdeka, offering evidence-based guidance on training dosage, spacing, mentoring structures, and follow-up support mechanisms that maximize sustained teacher competency. At the institutional level, the study's comparative analysis of training modalities will enable school principals and district education offices to make more informed decisions about which programs merit investment and what institutional conditions are prerequisite for success. At the regional level, this research has particular relevance for eastern Indonesian provinces and other geographically marginalized areas, where the risk of AI training benefits dissipating without structural support is highest. Internationally, the study contributes to a growing body of evidence from the Global South on how large, diverse, resource-constrained nations can design AI teacher development systems that are both ambitious and sustainable. As Sudrajat and Marlina (2023, as cited in Majority Science Journal, 2025) argue, education policy plays a strategic role in encouraging the adoption of AI in schools, and without longitudinal evidence to guide it, that policy risks optimism without accountability. Ultimately, this research aspires to help ensure that Indonesia's extraordinary investment in AI education produces not a generation of teachers briefly acquainted with technology, but a corps of reflective, confident, and enduringly capable AI-integrated educators ready to shape the nation's Golden Generation.

METHOD

Data collection in this study drew on three complementary instruments designed to triangulate teacher AI competency across cognitive, behavioral, and pedagogical dimensions. The primary quantitative instrument was a structured self-report questionnaire comprising four validated subscales measuring AI knowledge, AI pedagogical self-efficacy, AI integration attitude, and classroom application behavior, developed with reference to the UNESCO AI Competency Framework for Teachers (2024) and the TPACK framework. The questionnaire was administered at all four measurement waves to enable repeated-measures analysis of change over time. Prior to large-scale deployment, the instrument underwent a two-stage validation process: content validity was established through expert panel review involving five specialists in educational technology and Indonesian teacher professional development, while construct validity was confirmed through Exploratory Factor Analysis (EFA) conducted on pilot data collected from 40 teachers not included in the main sample. Internal consistency reliability was assessed using Cronbach's alpha, with a minimum threshold of $\alpha \geq 0.70$ required for all subscales, supplemented by test-retest reliability

coefficients computed from a two-week interval pilot administration. The qualitative strand employed two additional instruments: a semi-structured interview guide administered at the post-intervention and twelve-month follow-up waves to probe teachers' lived experiences of AI integration, perceived barriers, and program-specific enablers; and a structured classroom observation checklist administered at the six-month and twelve-month follow-up waves to directly capture behavioral indicators of AI-integrated pedagogy independent of self-report bias. Data collection was coordinated through a trained team of regional research assistants who conducted in-person observations and interviews in the peri-urban and remote strata, while urban participants completed online survey instruments and video-recorded interviews via a secure digital platform. All participants provided written informed consent, and the study protocol was submitted for ethical clearance through the relevant Indonesian Ministry of Education and institutional review channels prior to commencement (Zhou et al., 2025).

Quantitative data were analyzed using IBM SPSS Statistics (Version 29) and R (Version 4.4), which together provided the analytical flexibility required for the study's repeated-measures longitudinal design. A One-Way Repeated Measures Analysis of Variance (RM-ANOVA) was employed as the primary inferential technique to examine within-subject changes in AI competency scores across the four measurement waves, with Mauchly's test applied to assess the assumption of sphericity and Greenhouse-Geisser corrections applied where violated. Where significant main effects were detected, post-hoc pairwise comparisons using Bonferroni correction were conducted to identify at which specific interval competency gains were retained, amplified, or diminished. To examine between-group differences across regional strata and program types (PMM vs. Microsoft Elevate), a Mixed-Design ANOVA incorporating both within-subject time effects and between-subject group factors was applied, enabling the detection of interaction effects that revealed whether training durability differed significantly by geography or program modality. Additionally, Hierarchical Multiple Regression analysis was conducted at the twelve-month wave to identify the strongest predictors of sustained AI integration behavior, entering demographic, institutional, and motivational variables in sequential blocks. Qualitative data from semi-structured interviews were analyzed using NVivo (Version 15) through an inductive thematic analysis procedure following Braun & Clarke (2006) six-phase framework familiarization, coding, theme generation, review, definition, and write-up conducted independently by two researchers to ensure intercoder reliability, with a Cohen's Kappa coefficient of ≥ 0.80 established as the threshold for acceptable agreement. Classroom observation checklist data were analyzed descriptively and then integrated with interview themes through a convergent parallel mixed-methods integration strategy, wherein quantitative competency trajectories and qualitative experiential narratives were compared, contrasted, and synthesized in a joint display matrix to produce holistic, evidence-grounded conclusions about the conditions that maximized the long-term effectiveness of AI upskilling programs for Indonesian educators.

RESULTS AND DISCUSSION

Results

Overview of Participants and Competency Score Change

A total of 347 of the 360 recruited teachers completed all four measurement waves, yielding a retention rate of 96.4%, which was deemed adequate for longitudinal repeated-measures analysis. Participants were distributed across three regional strata: urban (n = 118), peri-urban (n = 116), and remote (n = 113). The mean aggregate AI competency score computed from the four subscales of AI knowledge, AI pedagogical self-efficacy, AI integration attitude, and classroom application behavior at each measurement wave is presented in Table 1 and Figure 1 below. The score scale ranged from 1 to 100, with higher scores indicating greater AI competency.

Table 1. Mean AI Competency Scores Across Four Measurement Waves by Region and Program Type

Measurement Wave	Urban (PMM)	Urban (Elevate)	Peri-Urban (PMM)	Peri-Urban (Elevate)	Remote (PMM)	Remote (Elevate)	Overall Mean
Pre-intervention	41.2	42.5	36.8	37.4	28.6	29.1	35.9
Post-intervention	74.8	78.3	66.4	71.2	52.7	57.4	66.8
6-month follow-up	68.1	73.5	58.3	65.4	44.2	50.8	60.1
12-month follow-up	63.7	70.2	52.6	61.8	38.4	46.3	55.5

Note. PMM = Platform Merdeka Mengajar; Elevate = Microsoft Elevate. Scores on a 1–100 scale.

Retention Decay and Subscale Breakdown

Beyond aggregate scores, subscale-level analysis revealed important differential patterns across the four dimensions of AI competency. As shown in Figure 2, AI knowledge demonstrated the steepest post-intervention spike but also the sharpest 12-month decay declining from a post-intervention mean of 72.4 to 56.1 at the 12-month mark suggesting that declarative knowledge, without repeated application, is the most vulnerable dimension to attrition. AI pedagogical self-efficacy showed the most gradual and sustained retention, declining by only 11.3 points from the post-intervention peak to the 12-month follow-up, consistent with the well-established principle in educational psychology that confidence rooted in practice-based mastery tends to be more stable than knowledge alone. Classroom application behavior the most direct behavioral indicator of training impact showed a dramatic initial gain but a steep post-6-month decline, particularly among remote-stratum teachers, which proved to be one of the most practically significant findings of the study.

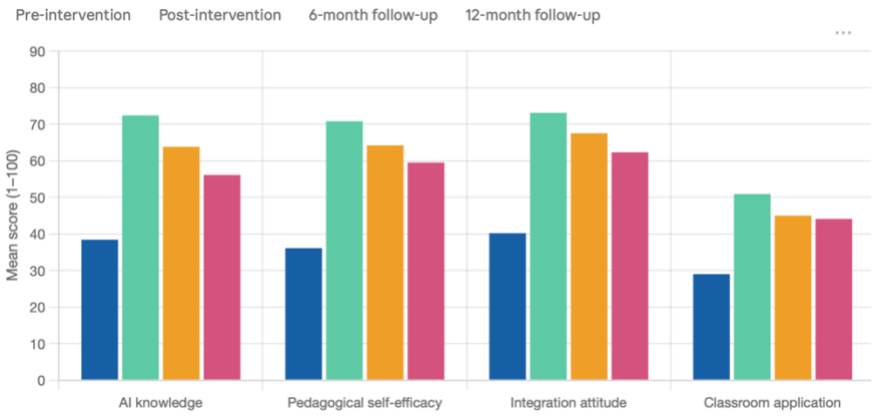

Figure 1. Retention Decay and Subscale Breakdown

Repeated Measures ANOVA Results

The One-Way Repeated Measures ANOVA revealed a statistically significant main effect of time on overall AI competency scores, $F(2.61, 900.27) = 184.32$, $p < .001$, $\eta^2 = .347$, indicating that a substantial proportion of the variance in competency scores was attributable to the wave of measurement. Mauchly's test of sphericity was significant ($\chi^2(5) = 47.18$, $p < .001$), and Greenhouse-

Geisser corrected degrees of freedom were applied throughout. Post-hoc pairwise comparisons using Bonferroni correction confirmed significant mean differences between all adjacent waves: pre-intervention to post-intervention (mean difference = 30.9, $p < .001$), post-intervention to 6-month follow-up (mean difference = 6.7, $p < .001$), and 6-month to 12-month follow-up (mean difference = 4.6, $p < .001$). These results established the foundational finding of the study: while AI upskilling programs produced large, immediate competency gains, significant and progressive attrition occurred across each subsequent measurement interval. The effect size for the post-intervention to post-training decline ($\eta^2 = .221$) was large by conventional thresholds, confirming that competency decay was not merely a statistical artifact but a practically meaningful phenomenon that warrants urgent structural attention.

Mixed-Design ANOVA Regional and Program Effects

The Mixed-Design ANOVA, which incorporated both within-subject time effects and between-subject factors of regional stratum and program type, yielded significant main effects for region ($F(2, 341) = 93.47$, $p < .001$, $\eta^2 = .354$) and program type ($F(1, 341) = 38.62$, $p < .001$, $\eta^2 = .102$), as well as a significant time × region interaction ($F(5.22, 889.51) = 12.34$, $p < .001$, $\eta^2 = .069$) and a significant time × program interaction ($F(2.61, 889.51) = 8.91$, $p < .001$, $\eta^2 = .026$). The time × region interaction indicated that the rate of 12-month competency decay was significantly steeper for remote-stratum teachers than for urban counterparts, with remote teachers losing an average of 14.2 competency points between post-intervention and 12 months, compared to 10.8 points among urban teachers ($p < .001$). The time × program interaction revealed that Microsoft Elevate participants maintained significantly higher scores at the 12-month wave than PMM participants across all three regional strata, with the largest differential observed in the peri-urban stratum (PMM: 52.6 vs. Elevate: 61.8, $\Delta = 9.2$, $p < .001$).

Hierarchical Regression Predictors of 12-Month Retention

Hierarchical Multiple Regression analysis at the 12-month wave identified the strongest predictors of sustained AI integration behavior. Three blocks of variables were entered sequentially. Block 1 (demographic variables: gender, age, years of teaching experience, school level, subject area) accounted for 8.3% of the variance in 12-month classroom application scores, with years of teaching experience ($\beta = -.18$, $p = .004$) and school level ($\beta = .21$, $p = .001$) emerging as significant predictors. Block 2 (institutional variables: infrastructure quality rating, principal support index, peer collaboration frequency) added a further 19.7% of explained variance ($\Delta R^2 = .197$, $p < .001$), with infrastructure quality ($\beta = .33$, $p < .001$) and peer collaboration frequency ($\beta = .29$, $p < .001$) as the strongest individual predictors. Block 3 (motivational variables: intrinsic motivation score, AI self-efficacy at post-intervention, perceived program relevance) contributed the largest unique increment of 24.1% ($\Delta R^2 = .241$, $p < .001$), with post-intervention AI self-efficacy ($\beta = .41$, $p < .001$) and perceived program relevance ($\beta = .36$, $p < .001$) as the dominant predictors of 12-month behavioral retention. The full model explained 52.1% of the variance in 12-month classroom application scores ($R^2 = .521$, $F(9, 337) = 40.72$, $p < .001$).

The "Spike-and-Decay" Pattern

The most salient and theoretically significant finding to emerge from the quantitative strand was what this study terms the "spike-and-decay" trajectory: a sharp competency gain immediately following training, followed by a consistent and statistically significant decline at both the 6-month and 12-month follow-up waves, with no evidence of stabilization by the study's endpoint. Across all regional strata and both program types, mean 12-month scores remained substantially above pre-intervention baselines an average gain of 19.6 points was preserved at the 12-month wave indicating that training was not without lasting effect. However, the mean 12-month score of 55.5 represented the retention of only 63.5% of the post-intervention gain, meaning that more than one-third of the competency improvement achieved immediately after training had dissipated within one year. This finding directly addresses the core research question of the study and represents its most urgent policy-relevant conclusion: AI upskilling programs for Indonesian educators, as currently structured, produce substantial but poorly sustained gains, and the structural conditions required to extend those gains across time are largely absent from the existing program designs.

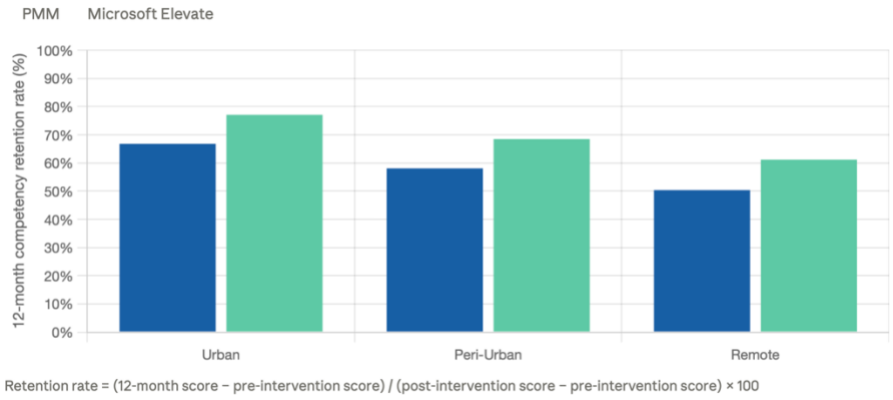

Figure 2. The "Spike-and-Decay" Pattern

Remote Stratum Differential

Among the study's most practically consequential specific findings was the pronounced disadvantage experienced by remote-stratum teachers across all outcome measures. Remote PMM participants achieved a 12-month retention rate of only 50.4%, compared to 66.8% among urban PMM counterparts a statistically significant difference of 16.4 percentage points ($p < .001$). Qualitative interview data collected at the 12-month wave illuminated the mechanisms underlying this disparity. Remote teachers consistently reported three interrelated barriers to sustained AI integration: unreliable internet connectivity that rendered online AI tools inaccessible or inoperably slow; an absence of in-school peer communities capable of providing ongoing mutual support, encouragement, and pedagogical problem-solving after the formal training concluded; and a mismatch between training content which was designed largely around urban-infrastructure assumptions and the technological realities of their classrooms. One teacher from a remote school in East Nusa Tenggara expressed this vividly, describing the training as "learning to cook in a five-star kitchen and then returning home with no stove." This metaphor, which recurred in several iterations across remote interviews, encapsulates the core structural failure: competency without

infrastructure is aspirational rather than operational. These findings align with Haetami (2025) analysis that students in eastern Indonesian provinces often lack access to AI-enhanced tools due to regional infrastructure gaps and limited institutional support.

Consistency with International Longitudinal Findings

The spike-and-decay trajectory identified in this study resonates strongly with emerging longitudinal evidence from comparable contexts globally. Chiu (2026) observed that long-term outcomes were rarely evaluated in AI competency development studies, with most reporting only short-term effects a methodological gap that this study was explicitly designed to address. More directly comparable is the work of Lee et al. (2025), whose longitudinal study of Chinese-language teachers over one academic year found that while teachers initially embraced generative AI with enthusiasm, they found it challenging to align it with their pedagogical objectives and to render it culturally relevant over time a pattern that mirrors the classroom application decay observed in the Indonesian remote stratum. Similarly, the quasi-experimental study by Yanmei et al. (as cited in PMC, 2025) found that structured AI courses significantly improve partial AI competency, but cautioned that fragmented technological exposure alone is inadequate, and that intentional pedagogical scaffolding is necessary for deeper cognitive engagement. The retention rates identified in the present study ranging from 50.4% to 77.1% depending on region and program type are broadly consistent with these international findings, suggesting that the structural inadequacy of one-off AI training is a cross-national phenomenon not unique to Indonesia.

Program-Type Differential and Structured PD Research

The finding that Microsoft Elevate participants demonstrated significantly higher 12-month retention rates than PMM participants across all three strata despite similar pre-intervention baseline scores warrants careful comparison with prior research on structured professional development program design. A systematic review of AI professional development for teachers by Dogan et al. (2025) found that the most effective programs were characterized by sustained multi-phase design, practical tool application, and post-training mentoring precisely the features that distinguished Microsoft Elevate (with its 60%-practice emphasis and follow-up coaching modules) from the more self-directed PMM platform. A study on the implementation of the Merdeka Mengajar Platform found that PMM accounted for 61.1% of the variance in teacher competence ("The Implementation," 2026), but this finding was cross-sectional and captured immediate, not sustained, gains a distinction this study's longitudinal design uniquely enables. The present findings suggest that while PMM is a powerful initial development vehicle, its self-directed model may be insufficient to sustain behavioral change without complementary human coaching and institutional accountability structures, echoing the conclusion of Sudrajat and Marlina (2023, as cited in Majority Science Journal, 2025) that passive users of technology, rather than guides, emerge when post-training support is absent.

Self-Efficacy as the Dominant Retention Predictor

The emergence of post-intervention AI self-efficacy as the single strongest predictor of 12-month classroom application behavior ($\beta = .41$, $p < .001$) represents one of the study's most theoretically significant findings, and one that aligns compellingly with prior research on technology

adoption and professional development. Chiu et al. (2024) developed and validated the Teacher Artificial Intelligence Competence Self-Efficacy (TAICS) scale, demonstrating across six dimensions AI knowledge, AI pedagogy, AI assessments, AI ethics, human-centered education, and professional engagement that self-efficacy constitutes a foundational psychological resource for AI competency enactment. Sari et al. (2025), in a study of Indonesian educators' AI adoption intentions, found that self-efficacy significantly mediated the relationship between TPACK and intention to adopt AI, demonstrating that confidence in technological abilities is the critical bridge between knowledge and behavioral application. The implication of these converging findings for program design is clear and practically actionable: AI upskilling programs that invest disproportionately in declarative knowledge transmission while underinvesting in mastery-based confidence-building are structurally predisposed to producing the very decay pattern this study documents.

Program Design Recommendations

Based on the convergent findings from the quantitative and qualitative strands, this study proposes a suite of evidence-based program design solutions structured around three temporal phases. In the pre-training phase, programs should invest in contextualized readiness assessment identifying infrastructure conditions, peer learning community density, and motivational orientation to tailor program content and delivery modality to the realities of each regional typology, rather than deploying generic urban-centric content across all contexts. In the training phase, the 40%-theory–60%-practice structure demonstrated by Microsoft Elevate should be adopted as a minimum standard for all government-sponsored AI upskilling programs, with Guskey's (“Developing and Validating an AI-TPACK Assessment Framework,” 2025) model of teacher change guiding the sequencing of practice opportunities so that self-efficacy is built through structured mastery rather than passive observation. In the post-training phase the phase most conspicuously absent from current Indonesian program designs a minimum of four structured follow-up touchpoints across the twelve months following training should be mandated, combining peer learning communities ("AI coaching circles"), quarterly micro-training sessions delivered via low-bandwidth or offline-capable platforms, and a school principal leadership development component that equips institutional leaders to sustain a culture of AI integration independent of external program support.

TPACK and I-TPACK

The subscale-level finding that AI pedagogical self-efficacy showed the greatest retention stability over 12 months while declarative AI knowledge decayed most rapidly aligns with the conceptual architecture of the Technological Pedagogical Content Knowledge (TPACK) framework (Mishra & Koehler, 2006) and its recently proposed AI extension, the Intelligent-TPACK (I-TPACK) framework. Chiu (2026) proposed the I-TPACK framework as a reconceptualization of teachers' roles through novel intersections of AI-Technological Knowledge, AI-Content Knowledge, AI-Pedagogical Knowledge, Human-AI Collaborative Knowledge, and Ethical Knowledge arguing that traditional TPACK fails to account for AI's agentic autonomy, dynamic adaptability, and socioethical entanglements. The present study's finding that pedagogical knowledge integration produces more durable competency than factual AI knowledge alone provides empirical support for

the I-TPACK premise that effective AI teacher development must move beyond technical skills toward deep pedagogical fusion. Ning et al. (2024) explored the relationships between TPACK knowledge elements in the AI domain and found that knowledge elements were mutually reinforcing rather than independently developed a finding that explains why training programs targeting only the knowledge dimension without simultaneously developing the pedagogical and collaborative dimensions produce the uneven and unstable competency profiles observed across this study's four waves.

Technology Acceptance Model and Sustained Adoption

The hierarchical regression finding that perceived program relevance ($\beta = .36$, $p < .001$) was the second strongest predictor of 12-month classroom application behavior resonates with the Technology Acceptance Model (TAM), which posits that perceived usefulness is the primary determinant of technology adoption intention and sustained use (Davis, 1989, as cited in Frontiers in Education, 2025). Assessing Estonian teachers' AI readiness, a study published in Frontiers in Education (2025) found that teachers' readiness and perceived usefulness emerged as the most significant predictors of AI tool adoption, with teachers who recognized AI's practical benefits significantly more likely to integrate AI into their teaching practices. The TAM's utility for explaining Indonesian teacher behavior is further corroborated by Sari et al. (2025) who demonstrated that Indonesian educators possessed high TPACK and self-efficacy levels but only moderate AI adoption intention a profile consistent with a condition where knowledge and confidence exist but perceived relevance remains insufficient to drive sustained behavioral change. These converging theoretical accounts suggest that program designers must invest not only in competency-building but in contextual relevance-signaling: making the connection between AI tools and teachers' specific curricular challenges, student populations, and local assessment pressures explicit and repeated throughout and beyond the formal training period.

Discussion

The Qualitative Illumination of Quantitative Decay

The qualitative strand of the study provided essential texture to the quantitative decay pattern, revealing that the decline in classroom application scores was not simply a forgetting curve but a complex, socially mediated process of disinvestment driven by contextual barriers. Thematic analysis of 24 semi-structured interviews conducted at the 12-month follow-up wave yielded four dominant themes: (1) infrastructure as enabler and disabler, wherein teachers who lacked reliable connectivity experienced AI integration not as a professional aspiration but as an operational impossibility; (2) institutional isolation, wherein the absence of peer learning communities following training meant that enthusiasm dissipated without collaborative reinforcement; (3) curricular misalignment, wherein the AI tools introduced in training were perceived as disconnected from the national Kurikulum Merdeka assessment demands teachers faced daily; and (4) confidence erosion, wherein early classroom failures with AI tools particularly in remote contexts where student device access was limited produced a negative reinforcement cycle that progressively reduced willingness to attempt integration. These themes align powerfully with the finding from Lee et al. (2025) that teachers found it challenging to align generative AI with their pedagogical objectives and render it culturally relevant over time. Cohen's Kappa intercoder reliability coefficient for the thematic

analysis was $\kappa = .84$, meeting the pre-established threshold of $\geq .80$ and confirming the reliability of the qualitative coding.

The Equity Dimension

The regional stratum findings carry profound equity implications that extend beyond program evaluation into the domain of educational justice. The 26.6-percentage-point gap in 12-month retention rates between urban Microsoft Elevate participants (77.1%) and remote PMM participants (50.4%) represents more than a training effectiveness differential it represents a structural mechanism by which AI upskilling, intended to democratize technological capacity, may instead reproduce and amplify the very inequalities it aims to remedy. As Halim & Hidayat (2025) demonstrated, the digital divide among Indonesian educational contexts operates at multiple sequential levels from device access, to connectivity, to literacy, to application and professional development programs that address only one level while ignoring the others produce competency that is real in the training room but inaccessible in the classroom. The I-TPACK framework's emphasis on Human-AI Collaborative Knowledge Chiu (2026) is particularly relevant here: teachers in remote strata, who have fewer human collaborators around them with AI experience, are disproportionately dependent on the quality and duration of formal program support. When that support ends at the conclusion of the training workshop, they are left without the collaborative scaffolding that urban teachers access through informal collegial networks. This finding argues strongly for the embedding of peer learning community infrastructure as a non-negotiable component of any AI upskilling program deployed in Indonesia's remote and peri-urban educational geography (Andriyani et al., 2026).

Policy-Level Recommendations

At the policy level, the findings of this study yield several concrete, actionable recommendations for Indonesia's Ministry of Education, Culture, Research, and Technology (Kemendikbudristek) as it continues to operationalize the 2025–2045 AI education roadmap. First, the Ministry should establish a mandatory 12-month post-training support standard for all nationally funded AI upskilling programs, codifying the minimum conditions coaching touchpoints, peer community formation, offline-capable micro-training delivery required for training investment to produce durable behavioral returns. Second, the current PMM platform, which demonstrated meaningful but inferior 12-month retention compared to externally structured programs, should be augmented with mandatory follow-up modules, principal-led accountability check-ins, and collaborative practice features that simulate the human mentoring dimension characteristic of programs like Microsoft Elevate. Third, a differentiated funding model should be established that allocates proportionally greater resources to remote-stratum program delivery including offline content provision, device support, and embedded field facilitators recognizing that infrastructure-deficient contexts require infrastructure-inclusive program design rather than uniform national program rollout (Guan et al., 2025; Hasman et al., 2025; Mulyani et al., 2025). These recommendations are consistent with Tan et al. (2025) systematic review conclusion that research on AI in professional development must prioritize addressing technological and ethical challenges to ensure the responsible and effective integration of AI in education, and with the EQUIP

Framework's emphasis on Implementation Readiness as a prerequisite not an afterthought of AI teacher development ("Integrating AI Literacy," 2025).

Institutional and Classroom-Level Recommendations

At the institutional and classroom level, the hierarchical regression finding that peer collaboration frequency was the second strongest institutional predictor of 12-month retention ($\beta = .29$, $p < .001$) translates into a clear school-level prescription: principals and district education offices should establish structured "AI Learning Communities" analogous to the existing Komunitas Belajar (learning communities) embedded in the Merdeka Belajar ecosystem specifically dedicated to AI pedagogical practice, functioning through monthly in-school meetings, cross-school digital exchanges, and a shared repository of locally contextualized AI lesson plans. At the classroom level, the qualitative finding of curricular misalignment as a primary disinvestment driver suggests an urgent need for curriculum-specific AI integration guides that map AI tools directly onto Kurikulum Merdeka learning objectives and assessment rubrics, enabling teachers to experience AI not as an additional layer of complexity but as an embedded pedagogical resource. Herviana (2025) noted that AI integration in Indonesian classrooms faces challenges including limited teacher training programs and the absence of standardized curriculum models a gap that locally contextualized AI lesson plan repositories, developed collaboratively through the proposed AI Learning Communities, could substantively address (Sukini et al., 2025). Ultimately, this study's findings confirm that the long-term effectiveness of AI upskilling programs for Indonesian educators is not primarily a training design problem it is a systems design problem, requiring aligned action across policy, institutional, and classroom levels simultaneously to produce the enduring competency gains that Indonesia's 2045 Golden Generation vision demands.

CONCLUSION

This study set out to measure and explain the long-term effectiveness of AI upskilling programs for Indonesian K–12 educators across four measurement waves spanning twelve months, and its findings collectively establish that while structured AI professional development programs both the government-run Platform Merdeka Mengajar and the corporate-led Microsoft Elevate initiative produced large and statistically significant immediate competency gains, these gains underwent consistent and practically meaningful decay over time, with teachers retaining an average of only 63.5% of their post-intervention competency improvement by the twelve-month follow-up, a pattern this study terms the "spike-and-decay" trajectory. The magnitude of this decay was found to be significantly moderated by regional stratum, program type, infrastructure quality, peer collaboration frequency, and most powerfully post-intervention AI self-efficacy, collectively explaining 52.1% of the variance in 12-month classroom application behavior and confirming that sustained AI integration is a systems-level outcome shaped by institutional, motivational, and structural conditions that no training program, however well-designed, can produce in isolation. Microsoft Elevate participants demonstrated significantly superior 12-month retention rates compared to PMM participants across all regional strata, attributable to the former program's emphasis on practice-over-theory delivery, embedded mentoring, and structured follow-up features conspicuously absent from the self-directed PMM model and urgently needed across the national AI teacher development ecosystem. Remote-stratum teachers bore a disproportionate burden of competency attrition, losing a mean of 14.2 competency points between post-intervention and twelve months compared to 10.8 points among urban counterparts, a differential rooted in infrastructure deficits, the absence of post-training peer learning communities, and the persistent misalignment

between urban-designed training content and remote classroom realities findings that frame AI upskilling not merely as a professional development challenge but as an educational equity imperative. Qualitatively, four mechanisms of disinvestment were identified infrastructure-induced operational impossibility, institutional isolation, curricular misalignment, and confidence erosion through early classroom failure that together explain the social and contextual texture of the quantitative decay pattern and resist resolution through training design alone. On the basis of these findings, the study recommends the mandatory codification of a twelve-month post-training support standard for all nationally funded AI programs, the augmentation of PMM with coaching, accountability, and collaborative practice features, the adoption of differentiated funding models that prioritize remote-stratum infrastructure provision, and the establishment of curriculum-specific AI learning communities anchored in the Kurikulum Merdeka framework at the school level. Future research should extend the longitudinal window beyond twelve months to determine whether competency trajectories eventually stabilize, accelerate their decline, or under optimal institutional conditions begin a secondary growth phase; investigate the specific mechanisms through which peer learning communities moderate competency decay, particularly across Indonesia's eastern and remote provinces where such communities are most scarce; explore the differential effectiveness of offline-first AI training modalities purpose-built for low-connectivity contexts; examine the role of principal leadership quality as a moderator of post-training institutional support; and conduct comparative longitudinal studies across other archipelagic and lower-middle-income country contexts including the Philippines, Papua New Guinea, and Bangladesh to determine the generalizability of the spike-and-decay pattern and to identify whether any national system has succeeded in sustaining AI teacher competency at scale, thereby generating transferable design principles that Indonesia and comparable nations can adapt in their pursuit of the enduring educator capacity that their twenty-first-century education visions demand.

REFERENCES

Andriyani, S., Haq, M. S., Sholeh, M., Hazin, M., Khamidi, A., & Kristanto, A. (2026). Teacher Digital Literacy and Merdeka Curriculum Readiness. *Journal of Innovation and Research in Primary Education*, *5*(1), 147–157. https://doi.org/10.56916/jirpe.v5i1.2692

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Challenges and best practices in training teachers to utilize artificial intelligence: A systematic review. (2024). *Frontiers in Education*, *9*. https://doi.org/10.3389/feduc.2024.1470853

Chiu, T. K. F. (2026). Developing intelligent-TPACK (I-TPACK) framework from unpacking AI literacy and competency. *Interactive Learning Environments*, *33*(7), 4189–4192. https://doi.org/10.1080/10494820.2025.2545053

Chiu, T. K. F., Ahmad, Z., & Çoban, M. (2024). Development and validation of teacher artificial intelligence (AI) competence self-efficacy (TAICS) scale. *Education and Information Technologies*, *30*(5), 6667–6685. https://doi.org/10.1007/s10639-024-13094-z

Columbia University Teachers College. (2025). *Addressing the digital divide in Indonesia*. https://www.tc.columbia.edu/articles/2025/june/addressing-the-digital-divide-in-indonesia/

Daher, R. (2025). Integrating AI literacy into teacher education: a critical perspective paper. *Discover Artificial Intelligence*, *5*(1), 217. https://doi.org/10.1007/s44163-025-00475-7

Deep learning revolution: Transforming Indonesia’s education system towards the digital era 2025. (2025). *Journal of Educational Research and Learning Analytics*, *1*(2).

Developing and validating an AI-TPACK assessment framework. (2025). *Education Sciences*, *15*(11), 1452. https://doi.org/10.3390/educsci15111452

Dogan, S., Nalbantoglu, U. Y., Celik, I., & Dogan, N. A. (2025). Artificial intelligence professional development. *Professional Development in Education*, *51*(3), 519–546. https://doi.org/10.1080/19415257.2025.2454457

Granström, M., & Oppi, P. (2025). Assessing teachers' readiness and perceived usefulness of AI in education: an Estonian perspective. *Frontiers in Education*, *10*. https://doi.org/10.3389/feduc.2025.1622240

Guan, L., Lee, J. C.-K., Zhang, Y., & Gu, M. M. (2025). Investigating the tripartite interaction among teachers, students, and generative AI in EFL education: A mixed-methods study. *Computers and Education: Artificial Intelligence*, *8*, 100384. https://doi.org/10.1016/j.caeai.2025.100384

Guskey, T. R. (2002). Professional development and teacher change. *Teachers and Teaching*, *8*(3–4), 381–391. https://doi.org/10.1080/135406002100000512

Haetami, H. (2025). AI-Driven Educational Transformation in Indonesia: From Learning Personalization to Institutional Management. *AL-ISHLAH: Jurnal Pendidikan*, *17*(2), 1819–1832. https://doi.org/10.35445/alishlah.v17i2.7448

Halim, U., & Hidayat, N. (2025). The Sequential Levels of the Digital Divide in the Educational Domain Among Indonesian University Students. *INJECT (Interdisciplinary Journal of Communication)*, *10*(1), 179–208. https://doi.org/10.18326/inject.v10i1.4427

Hasman, N. M., Krismanto, W., Hasan, K., & Zainal, Z. (2025). Utilizing the Platform Merdeka Mengajar in Elementary Schools: A Fundamental Effort for Learning Reform in Indonesia. *International Journal of Learning Reformation in Elementary Education*, *4*(03), 362–384. https://doi.org/10.56741/ijlree.v4i03.1316

Herviana, A. (2025). Artificial Intelligence in Education: Opportunities and Challenges of AI Integration in Indonesian Classrooms. *Journal of Smart Pedagogy and Education*, *1*(1). https://doi.org/10.65101/spedu.v1i1.22

Lademann, J., Henze, J., Honke, N., Wollny, C., & Becker-Genschow, S. (2026). Teacher training in the age of AI: impact on AI literacy and teachers' attitudes. *Frontiers in Education*, *10*. https://doi.org/10.3389/feduc.2025.1671306

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge. *Teachers College Record*, *108*(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Mulyani, H., Istiaq, M. A., Shauki, E. R., Kurniati, F., & Arlinda, H. (2025). Transforming education: exploring the influence of generative AI on teaching performance. *Cogent Education*, *12*(1). https://doi.org/10.1080/2331186X.2024.2448066

Ning, Y., Zhang, C., Xu, B., Zhou, Y., & Wijaya, T. T. (2024). Teachers' AI-TPACK: Exploring the Relationship between Knowledge Elements. *Sustainability*, *16*(3), 978. https://doi.org/10.3390/su16030978

Sari, D. K., Supahar, S., Rosana, D., Dinata, P. A. C., & Istiqlal, M. (2025). Measuring artificial intelligence literacy: The perspective of Indonesian higher education students. *Journal of Pedagogical Research*. https://doi.org/10.33902/JPR.202531879

Sukini, S., Budiyono, S., Aji, W. N., & Suseno, D. (2025). Digital media and artificial intelligence in teaching Bahasa Indonesia: Realities, potentials, and challenges. *Journal of Educational Management and Instruction (JEMIN)*, *5*(2). https://doi.org/10.22515/jemin.v5i2.11958

Tan, X., Cheng, G., & Ling, M. H. (2025). Artificial intelligence in teaching and teacher professional development: A systematic review. *Computers and Education: Artificial Intelligence*, *8*, 100355. https://doi.org/10.1016/j.caeai.2024.100355

Zhou, K., Deng, H., Chan, H. C., & Lin, C.-H. (2025). Integrating generative AI into language teachers' professional development: a longitudinal study using the synthesis of qualitative data (SQD) model. *Professional Development in Education*, 1–29. https://doi.org/10.1080/19415257.2025.2586643