E-ISSN: 3025-0994
Halaman 4314-4323
Volume 4 Nomor 4 Tahun 2026

Beyond Teacher Certification: A CIPP Evaluation of Curriculum Implementation in an Indonesian Elementary School

Risty Okta Nawangwulan[1*], Abdul Rasid Tuharea[2], Moy Phin Batmomolin[3], Izaak Hendrik Wenno[4], Emma Rumahlewang[5]

[1,2,3,4,5] Master's Program in Educational Management, Postgraduate Program, Universitas Pattimura, Ambon, Indonesia
e-mail: ristynawangwulan30@guru.sd.belajar.id

Abstract

This study evaluated the enactment of the Merdeka Curriculum in a public elementary school in Ambon, Indonesia, using the context, input, process, and product (CIPP) framework. A qualitative, decision-oriented case evaluation was undertaken on 7–8 May 2026. Evidence reported in the original study comprised interviews with the principal and teachers, classroom observations, school-document review, and photographs of facilities. Findings revealed a persistent implementation gap across the four evaluation dimensions. In the context dimension, approximately 85% of pupils were reported to come from lower- to middle-income households, alongside limited parental engagement. In the input dimension, eight of ten teachers held professional certification, and basic teaching materials were available, but instructional plans and digital equipment were not consistently used to support active learning. In the process dimension, approximately 80% of observed or reported instruction was described as teacher-centred, formative assessment was inconsistent, and principal-led classroom supervision occurred once annually. In the product dimension, approximately 70% of pupils reportedly exceeded the school's learning-objective attainment criteria, despite teachers' concerns about literacy, numeracy, attendance, and discipline. The main contribution is an evidence-bounded diagnosis of the discontinuity between formal teacher qualifications and enacted pedagogy. The findings support prioritising instructional coaching, formative assessment, targeted foundational-skills support, and school–family collaboration. Given the two-day, single-site design and incomplete quantitative denominators, the reported proportions are descriptive rather than generalisable estimates.

Keywords: CIPP Evaluation; Curriculum Implementation; Elementary Education; Instructional Leadership; Merdeka Curriculum

INTRODUCTION

Curriculum reform is realised in classrooms rather than in policy documents. The quality of elementary education therefore depends not only on teacher qualifications, curricula, equipment, and textbooks, but also on the opportunities that children actually encounter to read, reason, discuss, receive feedback, and demonstrate what they know. Indonesia's Merdeka Curriculum promotes adaptive teaching, meaningful learning, and attention to foundational competencies. Yet the transition from formal adoption to routine instructional practice is not automatic, particularly when schools face socioeconomic barriers and limited pedagogical support (Asmahasanah et al., 2023; Sutinah et al., 2024; Ningsi et al., 2024).

A central challenge in evaluating implementation is that readiness indicators can give a misleadingly positive picture when they are treated as substitutes for learning processes. Teacher certification, lesson-plan completion, and possession of digital devices document potentially valuable inputs; they do not, by themselves, demonstrate differentiated teaching, classroom discourse, diagnostic assessment, or improved student learning. Reviews of teacher digital competence similarly distinguish access to digital resources from the capacity

to use those resources for meaningful student activity (Heine et al., 2023; Claro et al., 2024). This distinction is especially important in schools where investments in facilities coexist with conventional instructional routines.

The context–input–process–product (CIPP) evaluation model provides an appropriate decision-oriented lens for examining this problem. Context evaluation identifies needs and constraints, input evaluation assesses strategies and resources, process evaluation documents implementation, and product evaluation considers results and intended or unintended consequences (Stufflebeam, 1967; Stufflebeam & Zhang, 2017). These dimensions are analytically connected: the presence of capable teachers or resources should be interpreted against local needs, observed instructional enactment, and the quality of learning evidence. Recent reviews confirm the model's continued value for examining complex educational initiatives, while also identifying the need for stronger outcome measures and longitudinal observation (Riawan et al., 2025; Wulan et al., 2026; Edriati et al., 2026).

Research on Merdeka Curriculum implementation has documented recurring challenges in teachers' interpretation of learning outcomes, project-based instruction, assessment, and professional development (Sitorus, 2024; Wardana, 2024; Sutinah et al., 2024). However, broad statements that the curriculum is 'implemented' or 'not implemented' can obscure where a particular school requires the most urgent corrective action. A school-level evaluation that connects context, resources, classroom behaviour, and student outcomes can provide a more actionable account. Equally, the international relevance of such an account lies not in assuming that one school represents an entire country, but in identifying a plausible implementation mechanism that other resource-constrained settings can examine.

This study revisits an original school evaluation undertaken in Ambon City. Its distinctive analytical problem is an apparent mismatch: a relatively substantial share of teachers possessed formal certification, yet the evidence suggested continued reliance on teacher-centred instruction, weak formative assessment, and limited instructional supervision. Rather than treating these patterns as a causal sequence that can be proven from a short visit, the study examines them as interrelated signals of an implementation gap. This reframing also avoids describing locally judged learning difficulties as results from standardised assessments when no such instruments were reported.

The CIPP approach also invites evaluators to distinguish judging a programme from improving it. Early evaluation theory framed the need for credible evidence about the decisions that schools and programme leaders actually face (Stufflebeam, 1966, 1967). Subsequent educational applications have used the framework in settings ranging from language programmes to vocational education and national curriculum reviews (Sopha & Nanni, 2019; Warju, 2016; Turmuzi et al., 2022). Their common relevance to the present case is not that the four categories operate as a standardised scoring instrument, but that they organise different kinds of evidence around distinct management questions. In particular, the product dimension must be interpreted in relation to locally specified goals, assessment quality, and the conditions under which learning opportunities were delivered.

The Indonesian elementary-school setting raises an additional interpretive issue. In contexts with uneven household resources, a curriculum designed to support learner agency may demand more deliberate scaffolding, not simply a declaration that teaching should be student-centred. Teachers may require examples of high-quality questioning, manageable group tasks, multimodal resources, and assessment checks suited to students with heterogeneous prior knowledge. Conversely, a routine that looks traditional on a brief visit may serve a particular explanatory purpose. Judgement should therefore focus on patterns across lessons and evidence of students' thinking, rather than attaching a negative label to all teacher explanations. A useful evaluation must separate curriculum intentions, observed enactment, and documented learning responses.

Three questions guide the analysis: (1) What contextual needs and input conditions characterised the school? (2) How did instructional practices and supervision align with the Merdeka Curriculum's student-centred aspirations? (3) What outcomes were reported, and what decisions are warranted by the strength of the available evidence? The article contributes

a transparent CIPP synthesis, an explicit account of evidentiary limitations, and a sequenced improvement agenda that distinguishes actions justified by observed concerns from effects that remain to be tested.

METHOD

The study used a qualitative, single-site evaluative case design structured by the CIPP model. Its purpose was to support interpretation and school-level improvement decisions rather than to estimate the prevalence of implementation problems across schools. The unit of analysis was one public elementary school in Ambon City, Indonesia. The institution was purposively selected because it was considered relevant to the socioeconomic and instructional challenges under examination; the original account does not establish the school's representativeness through a comparative sampling frame.

Fieldwork took place on 7 and 8 May 2026. The original report identified the principal and teachers as interview sources. It also described direct classroom observation, review of lesson plans, student records and the school profile, and photographic documentation of facilities. The total number of interviewees, classes or lessons observed, interview prompts, observation-duration schedule, and sampling rules were not reported in the source material. Consequently, these details are not reconstructed here. The statement that eight of ten teachers held certification concerns staff composition and should not be read as confirmation that all ten teachers were interviewed.

Evidence was organised deductively according to the CIPP dimensions. Context covered student socioeconomic conditions, family engagement, and learning needs; input covered teacher certification, curriculum documentation, textbooks and physical or digital infrastructure; process covered lesson routines, pedagogical approaches, formative assessment, student participation, and supervision; product covered achievement relative to the school's Learning Objective Achievement Criteria (Kriteria Ketercapaian Tujuan Pembelajaran, KKTP), foundational skills, attendance, discipline, and recorded achievements. Within each dimension, the report juxtaposed accounts from interviews, observation, and documents wherever described in the original evaluation. This was a structured re-synthesis of reported findings, not a new round of coding, interviews, or statistical analysis.

Four methodological safeguards governed interpretation. First, numerical summaries retained the original qualifiers (for example, 'approximately 85%' and 'approximately 70%') because denominator definitions and calculation procedures were unavailable. Second, teachers' judgements about poor literacy and numeracy were classified as stakeholder assessments, not standardised test scores. Third, explanations linking inadequate supervision or teacher-centred lessons to lower achievement were treated as hypotheses rather than demonstrated causes. Fourth, the article distinguished directly reported conditions from proposed improvement steps. These safeguards are essential because two days of observation cannot establish stable teaching patterns, trends in student character, or intervention effects.

A further issue concerns the unit of quantification. The ratio eight of ten certified teachers has a clear teacher denominator in the source, whereas the statements of approximately 85% disadvantaged pupils, approximately 80% teacher-centred instruction, and approximately 70% KKTP attainment are not accompanied by counts, observation units, or disaggregation rules. The present reanalysis did not treat these estimates as parallel indicators on a common scale. They answer different questions about enrolment, pedagogy, and assessment, and their precision cannot exceed that of the original records. A future reproducible evaluation would specify the number of pupils assessed, grade levels, subjects, minimum achievement thresholds, observed teaching segments, and the number of eligible lessons for each computation.

The case also has clear boundary conditions. Field information was collected during a short period and may not represent ordinary instructional activity across terms, weeks or grades. The study contains no intervention, counterfactual, repeated measurement, or independent verification of changes in student discipline. As a result, the analysis reports

associations and possible mechanisms, not effect estimates. The proposed improvement indicators later in the article are constructed as decision-support recommendations; they are not additional empirical observations. These distinctions allow the manuscript to contribute a practical evaluation argument without overstating what the source data can sustain.

The analysis followed the original report's CIPP categorisation and cross-source comparison, but detailed verbatim transcripts, coding records, observation rubrics, anonymised student-level attainment data, and formal triangulation logs were not supplied with the manuscript. No inferential tests were used or warranted. The percentages are presented to preserve the source record, while interpretation gives greater weight to converging qualitative observations and to the operational implications of identified discrepancies. No claim about research ethics approval or consent documentation is made because these records were not available in the supplied material.

RESULTS AND DISCUSSION

The evaluation suggested that the central problem was not a complete absence of educational inputs, but a weak connection between those inputs, everyday instruction, and evidence of student learning. Table 1 consolidates the reported findings and identifies the boundaries of their interpretation. It is a descriptive synthesis of the source study, not an independently validated indicator dashboard.

Table 1. CIPP evidence profile and limits of inference

CIPP domain	Reported evidence	Interpretive caution
Context	Approximately 85% of pupils were from lower- to middle-income families; limited parental engagement, inconsistent attendance, and reading delays were described.	Household classification, pupil denominator, and measures of parent involvement were not provided.
Input	Eight of ten teachers held certification; textbooks and basic furniture were available; projectors and an interactive board existed but were infrequently used.	Certification indicates formal status, not observed instructional quality; equipment-use frequency was not systematically measured.
Process	Lessons relied mainly on lecture and questioning; approximately 80% of instruction was characterised as teacher-centred; formative assessment varied; annual principal supervision was reported.	The 80% estimate lacks a defined observation denominator; annual supervision was reported, not independently audited.
Product	Approximately 70% of pupils reportedly exceeded KKTP; teachers expressed concerns about literacy, numeracy, attendance and conduct; no competition awards were reported in three years.	Achievement denominators, assessment comparability, baseline trend and objective skill tests were not supplied.

Context: household conditions and learning needs

The school was located beside a provincial road linking the urban area and the airport, providing access but potentially exposing classrooms to road noise. Interviews and school records were reported to indicate that approximately 85% of students came from lower- to middle-income households. Fishing, farming, and insecure employment were common among parents. Staff also associated irregular attendance and limited home support with concerns about pupil motivation, reading fluency, and discipline. These observations identify a need for supportive responses tailored to family circumstances, but the available data do not justify characterising economic status as the cause of individual pupils' learning difficulties.

This distinction matters for both equity and programme planning. A school can respond constructively to inconsistent attendance or low reading proficiency without attributing deficits to families. International syntheses find positive, although variable, associations between parental involvement and achievement, and they suggest that the nature of engagement matters more than assuming that all forms of homework assistance are equally effective (Kim, 2022; Cosso et al., 2022; Wang & Wei, 2024). Here the most defensible CIPP decision is to identify specific barriers through non-stigmatising family consultation, monitor attendance consistently, and make school-based reading support available regardless of household resources.

For contextual evaluation to become actionable, the school needs a shared definition of which learning barriers it can influence. For example, information about who has access to books, a quiet place to read, or reliable school transport would be more directly linked to intervention design than a broad socioeconomic classification by itself. None of those variables was documented in the original case; they are suggested subjects for follow-up assessment, not additional findings. The same principle applies to attendance: separating illness, travel, household work and motivation-related absence would prevent a single descriptive label from masking different needs. Family engagement should be reciprocal and feasible, with communication channels adapted to actual parent availability rather than treating low attendance at conventional school meetings as a lack of concern for children's learning.

Input: qualified staff and underused instructional assets

The school employed ten teachers, of whom eight held teaching certification. Classrooms contained basic furniture and whiteboards, textbooks were considered sufficient, and an interactive digital board was installed in the library. Projectors were reportedly stored centrally and moved into classrooms when needed. Part of the library served as a teachers' room, and there was no dedicated laboratory. Curriculum documents had been prepared, but staff indicated that these materials often functioned as administrative evidence rather than day-to-day instructional guides. The resulting input profile was therefore mixed: personnel credentials and core materials were relatively strong, but arrangements for pedagogical use were weaker.

A distinction between the availability and enactment of resources sharpens the interpretation. Professional certification is an important formal credential, but it does not directly measure questioning quality, task design, feedback, or differentiation during the observed lessons. Likewise, ownership of a digital board is not evidence of digital pedagogy. Recent reviews show that teachers' digital competence includes selecting, adapting, and deploying resources for student learning, not merely knowing that equipment exists (Heine et al., 2023; Claro et al., 2024). Studies of Merdeka Curriculum implementation also identify uncertainty about learning plans and assessment as an obstacle to curriculum enactment (Sutinah et al., 2024; Asmahasanah et al., 2023). The actionable input priority is thus not automatically additional procurement; it is supported use of existing resources and translation of curriculum plans into teachable routines.

A school-level inventory of resources and practices could distinguish at least three stages: whether a resource is available, whether it is accessible to teachers when needed, and whether its use generates observable student engagement or learning evidence. The source confirms the first stage for some devices and raises concerns about the second and third. For instance, storing shared projectors in an office may complicate access, but it does not show that the equipment would improve a lesson if moved to every classroom. A more defensible improvement cycle would identify relevant learning objectives, model a high-quality lesson using existing tools, observe whether pupils perform useful tasks, and retain the technology only when its pedagogical value is evident. Such analysis also protects limited school budgets from equipment-led solutions unsupported by an instructional rationale.

Process: instruction, formative assessment and supervision

Observed lesson openings reportedly included prayer, activation of prior knowledge and attendance checks. Classroom interaction was not uniformly passive: pupils answered questions and teachers maintained some dialogue. Nevertheless, the source characterised

approximately 80% of teaching as teacher-centred, with lecture and question–answer sequences predominating and little project work or structured group discussion. Formative assessment was applied inconsistently, and principal-led classroom supervision was reportedly undertaken once per year. Because the record does not specify the observation schedule behind the 80% estimate, it should be treated as a descriptive signal requiring further verification rather than a precise school-wide rate.

The strongest CIPP diagnosis arises from the mismatch between the curriculum's learning orientation and the prevailing classroom routine. Student-centred instruction need not mean eliminating teacher explanations; it requires giving learners appropriately scaffolded opportunities to apply, discuss, investigate, and use feedback. The formative-assessment evidence is especially consequential because assessment processes can make learning needs visible before end-of-unit results. A meta-analysis of 118 K–12 studies reported a positive mean effect of formative assessment on achievement, although this external evidence cannot predict the size of any improvement at the Ambon school (Yao et al., 2024). The practical implication is to specify observable instructional routines—such as lesson-level checks for understanding, timely feedback, and short follow-up tasks—rather than treating the presence of lesson plans as implementation success.

Infrequent supervision adds a potential organisational explanation but does not establish causality. Reviews of school leadership identify teacher collaboration, professional learning, and self-efficacy as pathways through which leadership may influence teaching practice (Ninković & Knežević Florić, 2024). A study of instructional leadership also reported a mediating role for teacher professional learning (Dorukbaşi & Cansoy, 2024). At this site, an annual observation is unlikely to create a frequent feedback cycle; however, the study did not measure the quality of that observation, informal coaching, or other professional learning that may have occurred. The preferred improvement is therefore a documented, supportive supervision cycle paired with peer planning rather than an unsupported claim that the principal caused weak instructional outcomes.

A process-monitoring protocol suitable for this school could record the lesson objective, the ratio of teacher exposition to pupil activity, the cognitive demand of pupil tasks, the evidence used to check understanding, and the feedback that follows. These are proposed observation fields and were not recorded systematically in the original report. Their advantage is that they convert the broad category of 'teacher-centred' instruction into behaviours that can be discussed constructively. Coaches could, for example, distinguish a carefully sequenced explanation followed by substantial independent practice from a lesson in which pupils receive little opportunity to apply concepts. Similarly, a brief exit task or oral probing can yield formative information without imposing an elaborate administrative burden. What matters is whether the next instructional decision responds to what pupils have demonstrated.

Supervision also requires attention to the conditions under which teachers can adopt change. Scheduling more observations without providing useful feedback, planning time, and examples of practice could simply increase compliance work. Research with Indonesian teachers has associated principal instructional and distributed leadership with professional development, suggesting the relevance of a collaborative rather than inspection-only approach (Amzat et al., 2022). At this school, practical options include common review of anonymised pupil work, paired lesson planning, shared reflection after observation, and recording one manageable follow-up action per teacher. These proposals should be evaluated for feasibility and acceptability before being scaled. The source data cannot establish the quality of current professional relationships, so no judgement about teachers' willingness to improve is warranted.

Product: attainment data and unmeasured foundational competencies

School-reported attainment data indicated that approximately 70% of students scored above the school's KKTP threshold, leaving roughly 30% below it. At the same time, teachers described literacy and numeracy as weak and expressed concern that some older pupils experienced difficulty with fluent reading. Reports also noted inconsistent attendance, classroom conduct requiring attention, and a lack of recorded academic or extracurricular

awards at subdistrict or city level over the previous three years. These are meaningful concerns, but they represent different types of evidence: school-specific threshold attainment, staff perceptions of competence, attendance or behavioural descriptions, and competition participation or awards.

The juxtaposition of 70% KKTP attainment with concerns about literacy and numeracy should not be treated as a statistical contradiction. A criterion-referenced classroom threshold and a broad judgement of foundational skills can describe different competencies, standards, assessment tasks, or cohorts. Without student-level records, subject- or grade-specific denominators, assessment instruments, and comparison data, the two observations cannot be reconciled quantitatively. Nor do competition awards provide a valid proxy for whole-school learning quality. The product component therefore yields a more cautious conclusion: existing indicators were insufficient to demonstrate robust attainment of foundational competencies, and a better-aligned monitoring system is needed before school-level effectiveness can be assessed confidently.

A more interpretable product evaluation would disaggregate KKTP attainment by grade, subject, time period and assessment format. It would also include short, consistent indicators of decoding and reading comprehension, basic number sense, and progression in pupils' work over a term. The present source supplies none of these distributions, so it cannot identify whether lower reported attainment is concentrated in particular cohorts or competencies. For this reason, the appropriate response is a baseline assessment and a plan for repeated measurement, not a retrospective claim of curriculum failure. Discrepancies between classroom scores and teachers' global judgements may also prompt useful professional dialogue about how attainment thresholds are set, what tasks assess, and whether the work demands comprehension or memorisation.

Interpretation of student conduct warrants similar restraint. Attendance can be monitored using clear administrative indicators, whereas changes in 'character' require specified behaviours, comparable observation windows, and safeguards against subjective judgements about children. The absence of competition awards is not sufficient to infer an absence of academic growth, especially when opportunities to compete, student selection, and programme priorities are unknown. A balanced product account would therefore prioritise meaningful learning, attendance, engagement, and wellbeing rather than treating competition outcomes as definitive institutional performance. These distinctions are relevant internationally because assessment validity and equity concerns arise wherever readily available school indicators are used to stand in for broader educational purposes.

Cross-domain interpretation and contribution

Taken together, the four CIPP dimensions reveal an implementation discontinuity. Contextual needs were substantial; some essential inputs were available; classroom enactment did not consistently reflect the intended curriculum; and product information was too uneven to establish how fully students had benefited. The observed configuration is compatible with a resource-to-practice bottleneck: formal preparedness has limited educational value if everyday pedagogical routines, assessment feedback, and professional support are not organised around student learning. It is not, however, evidence that certified teachers are ineffective as a group or that one factor caused poor outcomes.

This interpretation extends the descriptive value of the original CIPP report in two ways. First, it replaces a simple judgement of whether each component was 'good' or 'poor' with an account of their alignment, identifying transitions where improvement effort may generate more useful information. Second, it treats evidentiary quality as part of evaluation itself. Contemporary CIPP syntheses emphasise evidence-informed decisions while noting the need to strengthen outcome measurement and follow-up (Riawan et al., 2025; Wulan et al., 2026; Edriati et al., 2026). In the present case, that means prioritising a stable observation protocol and disaggregated student data before making broad claims about the effectiveness of curriculum reform.

The case also resonates with other Indonesian accounts of teachers' difficulty translating Merdeka Curriculum concepts into classroom practice (Sitorus, 2024; Ningsi et al.,

2024; Wardana, 2024). The contribution is analytic rather than representative: the school makes visible a problem that may plausibly occur elsewhere, while the transferability of any intervention depends on local leadership, staffing, resources, and family needs. International audiences can use this case to formulate comparative questions about implementation coherence, not to infer regional or national prevalence from a single site.

The improvement logic is best understood as a sequence of testable propositions. If teachers receive specific coaching and time to plan, then lesson designs may include more opportunities for student reasoning and regular checks for understanding; if these practices are implemented consistently and pupils receive targeted support, their foundational skills may improve. Each link requires evidence, and none can be assumed from certification data alone. In this formulation, CIPP is not merely four headings used after fieldwork: context identifies whose needs must be addressed; input specifies what support is feasible; process measures whether practice actually changes; and product tests whether intended learning benefits emerge. A future cycle can then revise the intervention where the sequence fails rather than attributing all shortcomings to an undifferentiated shortage of resources.

Evidence-linked school improvement priorities

CIPP evaluation is most useful when findings inform decisions. Table 2 translates the concerns identified in the original case into a proposed sequence of activities. These entries are recommendations, not interventions that were implemented or evaluated during the May 2026 fieldwork. Timeframes are illustrative planning horizons, and their feasibility should be agreed with the school.

Table 2. Proposed improvement actions and verification indicators

Proposed priority	Practical action	Evidence to collect
First 30 days: verify the baseline	Confirm pupil totals and KKTP records by grade; sample basic reading/numeracy skills; log attendance and document observation coverage.	Anonymised assessment summary, denominator definitions, attendance register, observation schedule.
Within one term: strengthen instruction	Run short peer-planning and coaching cycles focused on active tasks, checks for understanding, and feedback.	Revised teaching plans, annotated student work, repeat observation rubric, coaching logs.
Within one term: activate resources	Schedule shared use of projectors and the interactive board where they improve a specific learning objective.	Equipment-use calendar and short lesson reflections tied to learning activities.
Across the school year: engage families	Use accessible two-way communication and practical home-reading support; identify attendance barriers without blaming families.	Contact records, attendance trends and feedback from parents or carers.
End of term: evaluate change	Review learning indicators and teaching evidence against the baseline; revise the improvement plan.	Grade-level learning comparison, supervision records, documented decisions.

Limitations and directions for further research

The principal limitation is the short, two-day observation period in one purposively chosen school. The report does not state the number of interviews, observed lessons, or pupils underpinning the percentages; no full instrument set, transcripts, assessment data or observation rubric was provided for independent verification. The 80% teacher-centred estimate may be sensitive to which lessons were visited, while the 85% household and 70%

attainment figures cannot be audited without denominators. Stakeholder judgements about literacy, numeracy and student character may reflect genuine concerns but are not equivalent to validated measures or longitudinal change. The design cannot establish causal effects or support statistical generalisation.

Follow-up studies could combine a longer period of structured lesson observations with anonymised grade-level assessment records, periodic attendance monitoring, and interviews with families and students. Repeated observations before and after a clearly documented coaching intervention would help distinguish short-term classroom variation from sustained instructional changes. A multi-school comparison could examine which features of leadership support, teacher learning and local context explain differences in curriculum enactment. Such work should include explicit ethics procedures, clear sampling rules and a documented analysis trail.

CONCLUSION

The CIPP evaluation identified a gap between relatively adequate formal inputs and their inconsistent translation into learning practice at one Ambon elementary school. Teacher certification and available materials coexisted with largely teacher-centred instruction, uneven formative assessment and limited formal supervision. Reported attainment against KKTP did not resolve concerns about foundational skills because the measures were insufficiently documented. School leaders should prioritise verified learning baselines, recurring instructional coaching, targeted literacy and numeracy support, purposeful use of existing technology, and practical engagement with families. These recommendations require prospective monitoring; the two-day case cannot demonstrate their effectiveness or justify generalisation.

ACKNOWLEDGEMENTS

The authors thank the school principal and teachers for facilitating access, participating in interviews and sharing school information, and acknowledge the guidance of the Educational Evaluation and Supervision course in the Master's Program in Educational Management, Universitas Pattimura. No funding information was reported in the source manuscript.

REFERENCES

Amzat, I. H., Yanti, P. G., & Suswandari. (2022). Estimating The Effect Of Principal Instructional And Distributed Leadership On Professional Development Of Teachers In Jakarta, Indonesia. SAGE Open. Https://Doi.Org/10.1177/21582440221109585

Asmahasanah, S., Chairunnissa, I. C., & Hakim, N. (2023). Navigating Merdeka Curriculum In First Grade: Teacher Challenges And Strategies. Journal Of Integrated Elementary Education, 3(2), 137–149. Https://Doi.Org/10.21580/Jieed.V3i2.17592

Claro, M., Castro-Grau, C., Ochoa, J. M., Hinostroza, J. E., & Cabello, P. (2024). Systematic Review Of Quantitative Research On Digital Competences Of In-Service School Teachers. Computers & Education, 215, 105030. Https://Doi.Org/10.1016/J.Compedu.2024.105030

Cosso, J., Von Suchodoletz, A., & Yoshikawa, H. (2022). Effects Of Parental Involvement Programs On Young Children's Academic And Social-Emotional Outcomes: A Meta-Analysis. Journal Of Family Psychology, 36(8), 1329–1339. Https://Doi.Org/10.1037/Fam0000992

Dorukbaşi, E., & Cansoy, R. (2024). Examining The Mediating Role Of Teacher Professional Learning Between Perceived Instructional Leadership And Teacher Instructional Practices. European Journal Of Education, 59(3), E12672. Https://Doi.Org/10.1111/Ejed.12672

Edriati, S., Ambiyar, Yasmi, F., & Husnita, L. (2026). Dominasi Model CIPP Dalam Evaluasi Program Pendidikan Indonesia: Studi Literatur Sistematis Terhadap Model-Model Evaluasi. Jurnal Evaluasi Pendidikan, 17(1), 32–50. Https://Doi.Org/10.21009/Jep.V17i1.65711

Heine, S., Krepf, M., & König, J. (2023). Digital Resources As An Aspect Of Teacher Professional Digital Competence: One Term, Different Definitions – A Systematic Review. Education And Information Technologies, 28, 3711–3738. Https://Doi.Org/10.1007/S10639-022-11321-Z

Kim, S. (2022). Fifty Years Of Parental Involvement And Achievement Research: A Second-Order Meta-Analysis. Educational Research Review, 37, 100463. Https://Doi.Org/10.1016/J.Edurev.2022.100463

Ningsi, A., Sukiman, Agustina, A., Hardiyana, M. R., & Nirmala, S. U. (2024). Identifikasi Tantangan Dan Strategi Dalam Implementasi Kurikulum Merdeka Pada Tingkat Sekolah Dasar. Ideguru: Jurnal Karya Ilmiah Guru, 9(2), 678–682. Https://Doi.Org/10.51169/Ideguru.V9i2.877

Ninković, S., & Knežević Florić, O. (2024). School Leadership And Teaching Practice: A Systematic Review Of Studies Of The Indirect Effects. Journal Of Educational Administration, 62(4), 357–372. Https://Doi.Org/10.1108/JEA-08-2023-0190

Riawan, I. M. O., Citariani, N. M., & Lokasanti, I. K. (2025). Efektivitas Model Evaluasi CIPP Dalam Program Pendidikan: Kajian Literatur Sistematis. Jurnal Evaluasi Pendidikan, 16(2), 108–115. Https://Doi.Org/10.21009/Jep.V16i2.58457

Sitorus, F. R. (2024). Tantangan Dalam Implementasi Kurikulum Merdeka Di Sekolah Dasar. Jurnal BELAINDIKA (Pembelajaran Dan Inovasi Pendidikan), 6(1), 16–23. Https://Doi.Org/10.52005/Belaindika.V6i1.155

Sopha, S., & Nanni, A. (2019). The CIPP Model: Applications In Language Program Evaluation. Journal Of Asia TEFL, 16(4), 1360–1367. Https://Doi.Org/10.18823/Asiatefl.2019.16.4.19.1360

Stufflebeam, D. L. (1966). A Depth Study Of The Evaluation Requirement. Theory Into Practice, 5(3), 121–133.

Stufflebeam, D. L. (1967). The Use And Abuse Of Evaluation In Title III. Theory Into Practice, 6(3), 126–133.

Stufflebeam, D. L., & Zhang, G. (2017). The CIPP Evaluation Model: How To Evaluate For Improvement And Accountability. The Guilford Press.

Sutinah, C., Riyadi, A. R., Muftianti, A., Wulandari, M. A., & Ruqoyyah, S. (2024). Navigating Change: An Analysis Of Elementary School Teachers' Readiness And Implementation Challenges With The Merdeka Curriculum. AL-ISHLAH: Jurnal Pendidikan, 16(2), 1718–1729. Https://Doi.Org/10.35445/Alishlah.V16i2.5142

Turmuzi, M., Ratnaya, I. G., Al Idrus, S. W., Paraniti, A. A. I., & Nugraha, I. N. B. S. (2022). Literature Review: Evaluasi Keterlaksanaan Kurikulum 2013 Menggunakan Model Evaluasi CIPP (Context, Input, Process, Dan Product). Jurnal Basicedu, 6(4), 7220–7232. Https://Doi.Org/10.31004/Basicedu.V6i4.3428

Wang, X., & Wei, Y. (2024). The Influence Of Parental Involvement On Students' Math Performance: A Meta-Analysis. Frontiers In Psychology, 15, 1463359. Https://Doi.Org/10.3389/Fpsyg.2024.1463359

Wardana, M. D. K. (2024). Classroom Teachers' Problems And Strategies In Implementing The Merdeka Curriculum. Academia Open, 9(1). Https://Doi.Org/10.21070/Acopen.9.2024.9399

Warju. (2016). Educational Program Evaluation Using CIPP Model. Innovation Of Vocational Technology Education, 12(1), 36–42. Https://Doi.Org/10.17509/Invotec.V12i1.4502

Wulan, M. P., Listyasari, W. D., Santosa, H., & Huda, S. A. (2026). A Systematic Literature Review On The CIPP Evaluation Model In Education: Examining Its Trends, Applications, And Research Gaps. Journal Of Educational Sciences, 10(3), 307–320. Https://Doi.Org/10.31258/Jes.10.3.P.307-320

Yao, Y., Amos, M., Snider, K., & Brown, T. (2024). The Impact Of Formative Assessment On K–12 Learning: A Meta-Analysis. Educational Research And Evaluation, 29(7–8), 452–475. Https://Doi.Org/10.1080/13803611.2024.2363831