# BETWEEN POLICY AND PRACTICE: A QUALITATIVE CASE STUDY OF ASSESSMENT LITERACY IN AN INDONESIAN PRIMARY EFL CLASSROOM UNDER THE MERDEKA CURRICULUM

**Ananda Suraiya Ramdani[1], Nur Aisyah Zulkifli[2]**
22590824668@students.uin-suska.ac.id[1], nur'aisyah.zulkifli@uin-suska.ac.id[2]
**Faculty Of Tarbiyah And Teacher Training**

***ABSTRACT***

*When Indonesia launched the Merdeka Curriculum in 2022, it asked primary school English teachers to do something genuinely difficult: replace the familiar rhythms of summative testing with ongoing, dialogic, learner-centred assessment in classes of thirty-five students, with two forty-minute periods per week, and without meaningful professional development support. This study asks what that ask actually looks like from inside a classroom. Drawing on a qualitative single-case study of one experienced primary EFL teacher (Teacher A) in an urban Indonesian elementary school, it investigates how she understands, enacts, and adapts assessment practice under the Merdeka Curriculum. Data were generated through three semi-structured interviews, eight non-participant classroom observations, and systematic document analysis over a six-week period, and analysed using Braun and Clarke's (2022) reflexive thematic analysis framework. Three themes emerged: a structural divergence between Teacher A's formative assessment convictions and the summative logic embedded in school reporting; significant socio-contextual constraints large classes, absent rubric resources, and developmentally complex affective demands that shape her assessment design; and a repertoire of pragmatic adaptive strategies through which she improvises formative intent within systemic limitations. The findings extend existing frameworks of teacher assessment literacy (TAL) by proposing a third dimension assessment improvisation that captures how experienced teachers perform assessment competence under resource constraint. Practically, the study provides granular evidence that the Merdeka Curriculum's formative agenda cannot be realised without structural reform of school-level reporting systems and sustained, context-specific professional development for primary EFL practitioners.*

***Keywords:*** *Classroom-Based Assessment, Assessment Literacy, Merdeka Curriculum, Young Learner EFL, Primary Education, Indonesia, Qualitative Case Study.*

## INTRODUCTION

There is a quiet contradiction at the heart of English language teaching in Indonesian primary schools. On paper, the Merdeka Curriculum (Kurikulum Merdeka, 2022) charts an ambitious course: away from the discrete-point tests and end-of-term examinations that have long defined classroom evaluation, and toward an assessment culture that is formative, diagnostic, and genuinely centred on the learner's developing competence. On the ground, the teachers asked to enact this shift are working in conditions that were never designed to support it. They manage classes of thirty to forty young learners in rooms with a whiteboard and fixed rows of desks. They teach English for eighty minutes a week. They receive rubric templates and professional development from the same institution that still requires them to submit a numerical grade for every child, every semester.

This tension between a progressive curriculum philosophy and the institutional architecture that surrounds it is not unique to Indonesia. But the Indonesian primary EFL context gives it a particular urgency. The Merdeka Curriculum has been in active national implementation only since 2022, meaning that the window for empirically informing its rollout is narrow.

This study enters that space through the practice of one teacher. By following Teacher an experienced primary EFL practitioner at an urban elementary school that adopted the

Merdeka Curriculum in its first year across four weeks of lessons, interviews, and document review, it asks what formative assessment actually looks like when it meets the real conditions of Indonesian primary EFL teaching. The aim is not to evaluate whether Teacher A succeeds or fails against an ideal standard, but to understand the practical reasoning, the compromises, and the inventions that her work requires.

**a. The Problem: A Policy Mandate Without Infrastructure**

The Merdeka Curriculum's approach to assessment is formally grounded in the Profil Pelajar Pancasila the Profile of Pancasila Students a holistic, competency-based framework that requires teachers to evaluate learning processes rather than just products, and to use assessment diagnostically rather than merely for grading. The curriculum documentation distinguishes explicitly between Asesmen Formatif (ongoing, feedback-oriented assessment) and Asesmen Sumatif (summative, reporting-oriented assessment), and positions the former as the primary engine of learning (Kemendikbudristek, 2022).

Yet surveys of Indonesian EFL teachers conducted before and during the Merdeka transition consistently reveal a picture of unresolved ambiguity. Sulistyo et al. (2020) found that the majority of their sample could articulate formative assessment principles but seldom enacted them consistently; Wahyuni and Umam (2021) documented similar patterns across primary and secondary settings. Raharjo et al. (2023) conducted an early evaluation of the Merdeka Curriculum itself and observed that primary school teachers remained confused about the operational distinction between Asesmen Formatif and Asesmen Sumatif as defined in the new framework. These are survey-level findings: they tell us that the problem is widespread but not what it looks and feels like in a specific classroom, for a specific teacher, on a specific Tuesday morning.

That gap between policy-level description and classroom-level reality is precisely what a qualitative single-case study is positioned to fill.

**b. Teacher Assessment Literacy: The Theoretical Frame**

The concept of teacher assessment literacy (TAL) provides the theoretical anchor for this study. Over the past decade, TAL has been reconceptualised from a primarily technical competence knowing how to construct a valid test—to a complex, socially situated practice shaped by context, values, and institutional pressure (Xu & Brown, 2016; DeLuca et al., 2021; Lam, 2019). DeLuca et al.'s (2021) Approaches to Classroom Assessment Inventory (ACAI) framework is particularly relevant here: it identifies four intersecting assessment orientations (measurement, standards-based, student-centred, and interpretive) and demonstrates empirically that teachers navigate among them dynamically, shifting orientation in response to changing contextual demands

Coombe et al. (2020) identified insufficient pre-service training and heavy teaching loads as structural barriers to TAL development in the Gulf context. Gan et al. (2021) demonstrated that primary EFL teachers in China a context with material and institutional parallels to Indonesia reverted systematically to summative practices under reporting pressure even when they held strong conceptual commitments to formative assessment; they termed this "assessment reversion." The present study takes "assessment reversion" as a starting framework but asks whether Teacher A's practice is better described by a different concept—one that captures not retreat but adaptation.

The young learner dimension adds a further layer. Rixon (2020) and Edelenbos and Johnstone (2021) argue that primary school teachers function as the first assessment architects in a child's second language trajectory that the assessment practices children experience at ages nine to eleven shape their language anxiety, their intrinsic motivation, and their orientation toward formal language learning over years to come.

**c. Research Gap and Contribution**

Three features of the existing literature define the gap this study addresses. First, Indonesian EFL assessment research is concentrated at the secondary and tertiary levels (Serafini et al., 2021; Sukyadi, 2022), with primary school settings receiving disproportionately little attention. Second, the methodological mainstream is quantitative and survey-based; qualitative studies that use sustained classroom observation to examine the gap between teachers' espoused assessment values and their observed practice are rare (Saukah & Cahyono, 2021). Third, empirical research on the Merdeka Curriculum is still nascent (Raharjo et al., 2023), and almost none of it attends to the lived experience of primary EFL teachers enacting its assessment demands.

This study addresses all three dimensions simultaneously that may have broader relevance for researchers working in constrained educational settings beyond Indonesia.

**d. Research Questions**

This study is guided by three research questions:

1. How does a primary school EFL teacher navigate the tension between formative assessment ideals and summative institutional mandates under the Merdeka Curriculum?
2. What socio-contextual constraints shape this teacher's assessment design for young EFL learners?
3. What adaptive strategies does the teacher employ to sustain formative assessment intent within those constraints?

## METODE PENELITIAN

**1. Research Design**

This study employs a qualitative single-case study design following Yin's (2018) framework. The case is conceptualised as instrumental (Stake, 1995): Teacher A's practice is examined not as a biographical subject in its own right but as a window onto the systemic conditions that shape primary EFL assessment in Indonesia. Yin's model was preferred over Merriam's (2009) constructivist variant because of its explicit attention to the relationship between theoretical propositions and empirical evidence a logic well suited to a study that aims not only to describe but to extend existing TAL theory.

The single-case design requires justification. Critics of case study research correctly note that findings from one participant cannot be statistically generalised. This study makes no such claim. Rather, it pursues what Geertz (1973) calls "thick description" the contextually dense account of a particular situated practice that enables readers operating in comparable settings to make their own transferability judgements (Lincoln & Guba, 1985). The phenomenon under investigation the navigation of formative-summative tensions by a primary EFL teacher in the Merdeka Curriculum era is inherently local and context-embedded; the depth of analysis it demands is achievable only through intensive, sustained engagement with a bounded unit of analysis.

**2. Participant and Setting**

Teacher A is a female in-service EFL teacher with eleven years of primary school classroom experience. She holds a Bachelor's degree in English Language Education (S.Pd.) from a state university and has completed professional development training under both the 2013 Curriculum (K-13) and the Merdeka Curriculum frameworks. At the time of data collection, she was responsible for English instruction across four classes at Years 4 and 5 (ages nine to eleven), with class sizes ranging from thirty-two to thirty-seven students.

Teacher A was selected through purposive sampling on three criteria: sustained primary school EFL experience (minimum five years), placement in a school that had

adopted the Merdeka Curriculum in its inaugural year of implementation, and willingness to participate in extended fieldwork including classroom observation. These criteria were designed to maximise the informativeness of the case (Patton, 2015) to ensure that the participant could speak from direct, recent experience of the phenomena under investigation.

The setting is a state primary school in an urban district of a major Indonesian provincial capital, serving a predominantly lower-to-middle-income student population. The school adopted the Merdeka Curriculum at the start of the 2022–2023 academic year. English is taught as a compulsory subject from Year 4, with two forty-minute instructional periods per week. The physical environment is modest: standard fixed-row classrooms with a whiteboard, limited printed supplementary resources, and restricted access to digital devices. This material context is not incidental; it is the ground on which Teacher A's assessment practice unfolds, and it shapes what is and is not feasible for her to do.

## 3. Data Collection

Data were generated through three complementary methods deployed in sequence over six weeks, following the methodological triangulation logic established by Lincoln and Guba (1985).

Semi-structured interviews. Three interview sessions were conducted with Teacher A, each lasting approximately fifty to seventy minutes. The protocol was developed iteratively: the first interview established Teacher A's conceptual understanding of formative and summative assessment; the second, conducted mid-observation, focused on specific recent lessons and her real-time assessment decision-making; the third revisited emerging themes and invited Teacher A to reflect on provisional interpretations. All sessions were audio-recorded with consent, transcribed verbatim, and subsequently returned to Teacher A for member-checking. Her corrections and clarifications were incorporated into the final dataset.

Non-participant classroom observations. Eight observation sessions were conducted across the six-week period, covering all four of Teacher A's assigned classes (two Year 4, two Year 5). The researcher maintained structured field notes using an observation protocol adapted from Leung and Lewkowicz's (2020) CBA Observation Framework, tracking: the type and frequency of assessment events; the nature and recipients of feedback; observable student responses; and moments of visible tension between formative intent and summative procedure. Reflective memos were written within twenty-four hours of each session to preserve contextual detail and researcher reflexivity, and to flag emerging interpretive patterns for discussion in subsequent interviews.

Document analysis. A corpus of institutional and pedagogical documents was collected: Teacher A's lesson plans (Modul Ajar) for all observed units; her self-designed assessment rubrics; three sets of teacher-constructed written tests; student work samples (n=15 per class, selected through stratified purposive sampling across performance bands); and the school's official semester assessment reporting templates. Analysis followed Bowen's (2009) document analysis protocol, attending to both manifest content what the documents explicitly state and latent content what they assume, prioritise, or exclude.

## 4. Data Analysis

All data were analysed using Braun and Clarke's (2022) reflexive thematic analysis (RTA) framework, which positions themes as actively constructed through sustained interpretive engagement rather than as pre-existing entities within data. Analysis proceeded through six phases: familiarisation with the complete dataset through repeated reading and listening; systematic generation of initial codes across interviews, field notes, and documents; clustering of codes into candidate themes; iterative theme review and

refinement in dialogue with the theoretical literature; theme naming and definition; and report production.

Coding was conducted manually using marginal annotation and thematic mapping to maintain close contact with the primary data. The analytical process was documented in a reflexive research journal maintained throughout data collection and analysis; this journal records interpretive decisions, emergent uncertainties, and the researcher's engagement with potential biases particularly the risk of over-identifying with Teacher A's perspective given the collaborative nature of extended fieldwork.

**5. Trustworthiness**

Trustworthiness was established through four strategies aligned with Lincoln and Guba's (1985) criteria.

1. Credibility: Member-checking was conducted at two points after transcription of each interview, and after the preliminary thematic map was developed. Teacher A's confirmatory and corrective responses were incorporated into the analysis, and any persistent interpretive disagreements are noted in the findings.
2. Transferability: Thick description of the research context, participant, and school setting is provided throughout the methodology and findings sections to enable readers to assess the degree of contextual resonance with their own settings.
3. Dependability: All methodological decisions instrument development, sampling rationale, analytical procedures are documented here and in the research journal. Interview sessions were audio-recorded and transcribed verbatim.
4. Confirmability: Reflexive memoing throughout the study surfaced and managed potential interpretive biases. Emerging interpretations were cross-verified against all three data sources before inclusion in findings.

## RESULT AND DISCUSSION

Three themes emerged from the thematic analysis of the triangulated dataset. Each theme is presented with illustrative evidence from interviews, observations, and documents, followed by discussion that positions the findings in dialogue with the theoretical literature. The themes are interconnected they describe different facets of a single, coherent predicament but for analytical clarity they are presented sequentially.

**1. The Formative-Summative Fault Line**

The most persistent pattern across all data sources was the structural tension between Teacher A's articulated commitment to formative assessment and the summative logic embedded in the school's reporting architecture. This was not a gap between what Teacher A knew and what she did: she understood formative assessment well, she valued it, and she practised it. The problem was that the institutional infrastructure surrounding her practice had not been redesigned to accommodate it.

In interview, Teacher A described the dilemma with characteristic directness:

*"I know that assessment is supposed to be for learning, not just of learning. The Merdeka Curriculum says we should observe students, give feedback, adjust the lesson. I agree with that completely. But at the end of the semester, the school needs a number. Parents need a number. So I give feedback all lesson, and then I still have to give a score for the report. They don't always match."*

What is striking here is not the tension itself which is well-documented in the assessment literature but Teacher A's clarity about it. She has not confused formative and summative assessment; she operates both systems, simultaneously and in parallel. Observation data confirmed this duality. In six of eight observed lessons, Teacher A engaged in what the field notes describe as "in-flight assessment events": oral questioning

sequences that tracked comprehension in real time, peer correction activities that functioned as diagnostic checks, and brief written comments on student work that were clearly intended to inform revision rather than assign a grade. Yet the same lessons ended with Teacher A recording student participation scores in the class register using a frequency-based tally that compressed complex qualitative judgements into a single numeral.

Document analysis made the systemic dimension visible. Teacher A's Modul Ajar for the Year 4 unit "Describing People" explicitly referenced Asesmen Diagnostik as the lesson's opening move and specified formative check-ins at three points. The school's official assessment reporting template, however, offered only three columns: numerical scores for Knowledge, Skills, and Attitude. There was no field for qualitative feedback, developmental notes, or process observations. The curriculum spoke formative; the reporting system spoke summative; and Teacher A was the translator between them.

This finding extends Gan et al.'s (2021) concept of "assessment reversion" the retreat to summative practices under institutional pressure in an important direction. Teacher A was not reverting; she was bifurcating. She maintained a formative practice for pedagogical purposes and a summative practice for bureaucratic ones, running them in parallel rather than allowing one to displace the other. This is a more costly and more sustainable response than reversion, and it suggests that the problem lies not in Teacher A's assessment literacy but in the systemic incoherence between the Merdeka Curriculum's formative epistemology and the school's summative accountability infrastructure.

**2. Three Constraints on Young Learner Assessment Design**

A second theme concerns the socio-contextual factors that limit Teacher A's capacity to design assessment appropriate to her students' developmental and linguistic profiles. Three constraint clusters were identified across the dataset.

Class size and time poverty. With thirty-two to thirty-seven students and eighty minutes of English instruction per week, Teacher A's capacity for individual assessment was structurally constrained before she had made a single pedagogical decision. The mathematics of oral assessment are straightforward: with thirty-five students and two minutes per child, an individual speaking assessment consumes the entire weekly allocation for one class. In practice, this meant that Teacher A consistently used group-based speaking formats and holistic impressionistic scoring rather than rubric-referenced individual evaluation. In interview, she was precise about the tradeoff:

*"To test speaking properly, I need to hear each child individually. With thirty-five students and eighty minutes per week, that's impossible. If I give each student two minutes, the class is finished. So I do group speaking—but then how do I score the individual child? I can't be certain who contributed what."*

This is an assessment design problem with no clean solution at current resourcing levels

Absent rubric infrastructure. Teacher A's self-designed assessment rubrics were predominantly binary or three-point constructs (Correct / Partially Correct / Incorrect) that lacked the performance-level descriptors recommended for young learner EFL assessment (Rixon, 2020; Edelenbos & Johnstone, 2021). In interview, she was candid about why:

*"I made these rubrics myself, based on what I know from my degree years ago. Nobody has given me training on how to write better ones. I've looked online but the examples are usually for high school or university."*

The absence of ministry-supplied or school-provided rubric templates for primary EFL assessment is not a minor administrative gap. The Merdeka Curriculum endorses formative, process-oriented assessment; the support infrastructure for implementing it has not materialised at the primary school level.

Affective complexity. The third constraint is perhaps the most theoretically interesting and the most under-examined in the TAL literature. Teacher A's classroom observations revealed a consistently high degree of affective sensitivity in her assessment practice: she avoided public individual error correction, favoured whole-class feedback, and delivered evaluative comments in what the field notes describe as an "encouragement-heavy register." In interview, she explained the reasoning explicitly:

*"These children are nine or ten years old. If I mark them wrong in front of everyone, or give a low score on a test, they feel ashamed. In Indonesian culture, shame is very powerful especially in front of friends. I have to be very careful. Sometimes I soften the feedback in a way that means the assessment data is not fully accurate, but the child's confidence is protected."*

This testimony surfaces a genuine tension between assessment validity and affective safety. Dewaele and MacIntyre's (2022) research on language anxiety in young learners provides strong theoretical support for Teacher A's orientation: fear- or shame-based assessment environments systematically suppress authentic language performance, meaning that technically "valid" instruments deployed in affectively unsafe conditions can produce profoundly invalid data. Teacher A's softening of feedback is not a failure of assessment rigour; it is a contextually intelligent response to the psychological dynamics of young learner EFL. The problem and this is a theoretical contribution of the present study—is that mainstream TAL frameworks do not yet account for this dimension. Assessment literacy in young learner contexts requires an affective competence that existing frameworks have not systematically theorised.

**3. Assessment Improvisation as Professional Practice**

The first two themes describe what constrains Teacher A. The third describes what she does about it. Far from a passive subject of systemic pressures, Teacher A had developed a repertoire of adaptive strategies that preserved formative assessment intent within the structural limits imposed upon her. These strategies were not explicitly trained; they were improvised through years of practical reasoning. Together, they constitute what this study proposes to call assessment improvisation: the adaptive, constraint-responsive performance of assessment literacy in under-resourced contexts.

Shadow documentation. Unable to maintain the formal student portfolios that the Merdeka Curriculum recommends, Teacher A had developed a parallel system of her own: a personal notebook in which she recorded brief qualitative observations of individual student performance across lessons observations that informed her summative judgements but were invisible in the school's reporting infrastructure. This practice closely parallels what Inbar-Lourie and Donitsa-Schmidt (2020) call "shadow assessment" unofficial, teacher-held assessment intelligence that supplements and corrects the distortions of official measurement. The notebook was the real assessment record; the grade sheet was the institutional artefact.

Embedded assessment. Teacher A consistently embedded assessment within learning activities rather than designating separate testing occasions. Vocabulary knowledge was checked through game formats; listening comprehension through physical response tasks; spoken production through pair-work conversations that students experienced as communicative practice. Observation notes record students laughing during a vocabulary game that Teacher A was simultaneously using to track who could and could not access key lexical items for the upcoming unit. The students were not aware they were being assessed. The assessment was, in this sense, ecologically valid in ways that formal testing procedures cannot achieve: it captured language behaviour in conditions of genuine communicative engagement rather than test anxiety.

Strategic dual compliance. Perhaps most revealing was what Teacher A explicitly described as her "two-system" approach:

*"I have two systems and they don't communicate. My notebook tells me the real picture. The school report shows the simplified picture. I manage both but it takes energy, and I worry I am the only teacher who does this. If I leave, the system stays broken."*

These three strategies extend existing TAL frameworks in a way that has theoretical purchase beyond the Indonesian context. DeLuca et al. (2021) distinguish between assessment knowledge (formal understanding of assessment principles) and assessment practice (contextually enacted behaviour). Teacher A's case suggests that in under-resourced settings, a third dimension is needed: assessment improvisation, defined as the capacity to construct and sustain assessment practice that is pedagogically principled despite the absence of the structural conditions that would normally support it. This is not the same as assessment knowledge applied to constrained practice; it is a qualitatively different competence, developed through experience and reflective practice rather than training, and it carries a sustainability cost that individual improvisation cannot resolve.

## CONCLUSION

This study set out to understand what formative assessment looks like for a primary EFL teacher navigating the Merdeka Curriculum inside an under-resourced Indonesian elementary school. The answer is: remarkably sophisticated, and structurally unsupported. Teacher A possesses a well-developed, practice-constructed assessment literacy that aligns closely with the formative philosophy of the curriculum she is asked to implement.

The systemic finding is equally clear: the Merdeka Curriculum's formative assessment mandate is currently being sustained by individual teacher ingenuity in the absence of the institutional infrastructure reformed reporting systems, targeted professional development, ministry-supplied young learner rubric resources that would make it sustainable at scale. Teacher A's "two systems" are not a sign of her limitation; they are a sign of the system's.

The study makes two theoretical contributions. First, it extends the concept of assessment reversion (Gan et al., 2021) by identifying bifurcation as an alternative response to institutional pressure: rather than retreating from formative practice, Teacher A ran formative and summative systems in parallel, at personal cost. Second, it proposes the concept of assessment improvisation the adaptive, constraint-responsive performance of TAL in under-resourced contexts as a necessary addition to existing frameworks. DeLuca et al.'s (2021) two-dimensional model of assessment knowledge and practice needs a third dimension if it is to describe what competent teachers actually do when the conditions for competent practice are absent.

For Indonesian policymakers, the evidence is clear: the Merdeka Curriculum's formative assessment mandate will not be realised without commensurate reform of school-level reporting architectures. The immediate, high-leverage intervention is the replacement or augmentation of purely numerical reporting templates with qualitative descriptor frameworks a change that the curriculum's own documentation recommends but that has rarely been enacted at school level. Without this structural change, teachers will continue to maintain two systems, and the formative one will remain invisible to the institution.

For school principals and district supervisors, the establishment of assessment-focused professional learning communities structured around young learner EFL rubric design, embedded assessment strategies, and the affective dimensions of primary school evaluation would provide the collaborative, contextually situated development environment that supports TAL growth.

For researchers, the "assessment improvisation" concept proposed here merits theoretical development and empirical testing through comparative case studies across Indonesian provinces, school types, and curriculum implementation phases. Longitudinal designs that track individual teacher assessment practice across a full academic year capturing seasonal pressures such as end-of-semester reporting would provide evidence about whether and how improvised assessment strategies are sustainable over time. Research that explicitly includes student perspectives on primary EFL assessment, particularly its affective dimensions, would complement the teacher-centred focus of this study and contribute to a more complete picture of young learner assessment ecology in Indonesia.

## REFERENCES


Bowen, G. A. (2009). Document analysis as a qualitative research method. Qualitative Research Journal, 9(2), 27–40. https://doi.org/10.3316/QRJ0902027

Braun, V., & Clarke, V. (2022). Thematic analysis: A practical guide. SAGE Publications.

Coombe, C., Troudi, S., & Al-Hamly, M. (2020). Foreign and second language teacher assessment literacy: Issues, challenges, and recommendations. In D. Tsagari & R. Banerjee (Eds.), Handbook of second language assessment (pp. 55–70). De Gruyter Mouton.

Creswell, J. W., & Poth, C. N. (2018). Qualitative inquiry and research design: Choosing among five approaches (4th ed.). SAGE Publications.

DeLuca, C., Coombs, A., Macgregor, S., & Luhanga, U. (2021). Approaches to classroom assessment inventory: A new instrument to support teacher assessment literacy. Educational Assessment, 24(4), 255–272. https://doi.org/10.1080/10627197.2019.1670056

Dewaele, J.-M., & MacIntyre, P. D. (2022). The predictive power of multicultural personality traits, tolerance of ambiguity and self-confidence on foreign language enjoyment and anxiety. European Journal of Applied Linguistics, 10(1), 7–34.

Edelenbos, P., & Johnstone, R. (2021). Preparing young learners for lifelong language learning: Assessment perspectives. Language Learning Journal, 49(3), 255–268.

Gan, Z., Leung, C., & Yu, G. (2021). Teacher assessment literacy in practice: A study of primary EFL teachers in China. Language Testing, 38(2), 243–266. https://doi.org/10.1177/0265532220954108

Geertz, C. (1973). The interpretation of cultures. Basic Books.

Hamied, F. A. (2020). English in multicultural and multilingual Indonesian education. In A. Kirkpatrick & A. J. Liddicoat (Eds.), The Routledge international handbook on language education policy in Asia (pp. 261–273). Routledge.

Hopwood, N. (2021). Expertise, learning and agency in practice: A sociomaterial analysis. Teaching and Teacher Education, 108, 103504.

Inbar-Lourie, O., & Donitsa-Schmidt, S. (2020). Exploring the assessment literacy waters: School teachers and university teachers swimming together. Language Assessment Quarterly, 17(2), 186–203.

Kemendikbudristek. (2022). Panduan pembelajaran dan asesmen pendidikan anak usia dini, pendidikan dasar, dan pendidikan menengah. Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi.

Kremmel, B., & Harding, L. (2020). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups. Language Assessment Quarterly, 17(1), 100–120.

Lam, R. (2019). Teacher assessment literacy: Surveying knowledge, conceptions and practices of classroom-based writing assessment in Hong Kong. System, 81, 78–89.

Leung, C., & Lewkowicz, J. (2020). English language teaching and the practitioner-researcher project. TESOL Quarterly, 54(1), 5–30.

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. SAGE Publications.

Merriam, S. B. (2009). Qualitative research: A guide to design and implementation. Jossey-Bass.

Patton, M. Q. (2015). Qualitative research and evaluation methods (4th ed.). SAGE Publications.
Raharjo, T. J., Mulyono, S. E., & Nurtanto, M. (2023). Evaluation of Merdeka Curriculum implementation in Indonesian primary schools. Journal of Education and Learning, 17(2), 301–312.
Rixon, S. (2020). British Council survey of policy and practice in primary English language teaching worldwide. British Council.
Saukah, A., & Cahyono, B. Y. (2021). Assessment practices of EFL teachers in Indonesian schools: A national survey. TEFLIN Journal, 32(1), 118–140.
Scarino, A. (2013). Language assessment literacy as self-awareness. Language Testing, 30(3), 309–327.
Serafini, F., Kachorsky, D., & Aguilera, E. (2021). Multimodal literacies in the context of assessment. Journal of Adolescent & Adult Literacy, 64(4), 401–410.
Stake, R. E. (1995). The art of case study research. SAGE Publications.
Sukyadi, D. (2022). Assessment literacy in Indonesian higher education. Indonesian Journal of Applied Linguistics, 12(1), 118–131.
Sulistyo, G. H., Basthomi, Y., Cahyono, B. Y., & Prayogo, J. A. (2020). Exploring teachers' assessment literacy within Indonesian curriculum reform. TEFLIN Journal, 31(1), 95–117.
Wahyuni, S., & Umam, M. K. (2021). Classroom-based assessment implementation in Indonesian primary and secondary EFL settings. Journal of Language Teaching and Research, 12(3), 396–405.
Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. Teaching and Teacher Education, 58, 149–162.
Yin, R. K. (2018). Case study research and applications: Design and methods (6th ed.). SAGE Publications.