Jurnal Inovasi Pendidikan
dan Sosial Humaniora

SAH PUBLISHER

Vol. 1, No. 1, August 2026

Assessment Without Scores through Narrative and Formative Assessment Paradigms

Anindya Nurul Kusuma Dewi[1*], **Ayun Hapsari**[2]

[1]Universitas Sugeng Hartono

[2]Universitas Terbuka

**Corresponding author. Email: aninnkd@ush.ac.id*

Abstract. Scores and grades dominate educational assessment because they compress complex learning into portable, comparable symbols, yet an influential body of scholarship argues that this compression damages learning itself. This paper critically reviews two paradigms that loosen the grip of the score: narrative assessment, which represents learning as an evidenced account of progress over time, and formative assessment, which subordinates judgement to the improvement of learning while it is still underway. Using a critical narrative review method, the paper selects landmark theoretical works, systematic and narrative reviews, meta-analyses, and official curriculum documents, and reads them against one another in order to test what each paradigm can and cannot do. The review finds that scores perform social and psychometric functions that narratives cannot replace at scale, including selection, certification, aggregation, and system monitoring, while narratives preserve exactly what scores erase: process, context, identity, disposition, and the next step a learner should take. Evidence on learning effects supports well designed formative practice but is more modest and more contested than early claims suggested, and the defining experiments on comments and grades show that appending a grade can cancel the benefit of formative comment. Score-free assessment therefore fails when it is treated as a slogan and succeeds only where purpose is separated from certification, where shared criteria and moderation discipline narrative judgement, where workload is honestly resourced, and where reporting to parents and systems is redesigned rather than merely renumbered. The paper's novelty lies in joining validity theory, grading history, and curriculum policy, including Indonesia's Kurikulum Merdeka, into a single set of conditions under which assessment without scores is defensible at scale.

Keywords: assessment without scores; narrative assessment; formative assessment; feedback; grading; Kurikulum Merdeka

Received July 2026 / Revised August 2026 / Accepted August 2026

This work is licensed under a Creative Commons Attribution 4.0 International License.

INTRODUCTION

Few artefacts of schooling are as ordinary, and as consequential, as the score. A single mark condenses weeks of effort, a teacher's judgement, and a curriculum's expectations into a symbol that can be averaged, ranked, reported, and compared across classrooms, schools, and nations. Reviewing a century of grading research, Brookhart and her colleagues show that grading has always carried several meanings at once, mixing achievement with effort, behaviour, and improvement, and that its value lies less in precision than in its role as the most common currency through which schools communicate [1]. That currency, however, was not born neutral. Schneider and Hutt trace

how the A to F marking scheme in the United States began as an intimate and local form of communication between teacher and student, and was progressively standardised so that judgements could travel between institutions, trading nuance and local knowledge for uniformity and portability [2]. The history matters for the present debate, because it shows that the score is not a natural unit of learning. It is an administrative technology, designed for movement across systems, and its dominance reflects institutional needs as much as pedagogical ones.

Against this technology stands a persistent counter tradition. In 1998 Black and Wiliam published their landmark review of classroom assessment and concluded that formative assessment, understood as assessment whose evidence is actually used to adapt teaching and learning, produced learning gains that were large by the standards of educational interventions [3]. The claim travelled around the world and was repeated so often that it hardened into a slogan. Later and more careful syntheses complicated the picture. Kingston and Nash, screening more than three hundred studies and retaining only thirteen that met their methodological criteria, found a weighted mean effect of 0.20, substantial enough to matter but far below the 0.40 to 0.70 range associated with the earlier claim [4]. Bennett, reviewing the same field, warned that formative assessment was not yet a well defined set of practices, that its evidence base rested partly on sources that were untraceable, flawed, or dated, and that the policy enthusiasm had outrun the scholarship [5]. Any honest review of assessment without scores must therefore begin by resisting two temptations: the temptation to treat scores as self evidently corrupting, and the temptation to treat their alternatives as self evidently liberating.

The second paradigm examined here, narrative assessment, makes an even stronger claim. Where formative assessment changes the use of judgement, narrative assessment changes its form. In the early childhood tradition associated with Margaret Carr, assessment is written as a Learning Story: a structured, evidenced narrative that notices learning, recognises its significance, and responds by planning what might come next [6]. Such stories reject the checklist logic of fragmented outcomes and instead treat learning as situated, dispositional, and continuous with a child's identity and community. Their intellectual roots reach into narrative inquiry more broadly, where experience itself is understood as storied and where the reconstruction of experience in narrative form is treated as a legitimate mode of educational knowledge [7]. Between them, the formative and narrative paradigms raise the question this paper addresses. What, precisely, is lost when the score is removed, what is gained, what does the evidence actually support, and under what conditions could assessment without scores be defended not as an experiment in one classroom but as policy at scale?

The paper proceeds as follows. The Methods section states the review approach and its limits. The Result and Discussion section then works through the social and psychometric work that scores perform, the knowledge that narratives preserve, the evidence on learning effects, the risks and costs of score-free designs, the validity question, hybrid designs that combine the paradigms, and the Indonesian policy context of Kurikulum Merdeka, before stating the conditions under which score-free assessment is defensible at scale.

METHODS

This paper is a critical narrative review. It does not report a systematic review and it makes no claim to exhaustive coverage, to a reproducible search protocol, or to

quantitative synthesis. Instead it follows the logic of the scholarly review essay: a purposive selection of sources that are theoretically central, empirically weighty, or officially normative for the question at hand, read closely and set in critical dialogue with one another. Selection proceeded by three routes. First, the review began from the recognised landmark works of the field, including the formative assessment syntheses of Black and Wiliam, the theoretical statements of Sadler, the feedback models of Hattie and Timperley, and the narrative assessment programme of Carr, because any account of these paradigms that bypassed them would not be credible. Second, it used backward and forward citation tracing from those works to locate both their intellectual sources, such as validity theory, and their critics, such as the meta-analytic and conceptual correctives that followed the first wave of enthusiasm. Third, it admitted official curriculum and policy documents where the argument turns on what systems actually require, including the New Zealand early childhood curriculum under which Learning Stories developed, the OECD review of evaluation and assessment frameworks, and the Indonesian guidance for Kurikulum Merdeka.

Three inclusion judgements governed the selection. Sources had to address the form, function, or consequences of assessment rather than instruction in general. Empirical claims about learning effects were admitted from peer reviewed meta-analyses and controlled studies in preference to advocacy literature, and where a famous claim had been corrected by later work, both the original and the correction were retained so that the reader can see the dispute rather than a settled fiction. Theoretical and philosophical sources were admitted where they changed what counts as evidence, since the dispute between scores and stories is partly a dispute about the nature of educational knowledge. The resulting corpus is dominated by books and journals from major international publishers, supplemented by a small number of official documents and association journals, and the full list with verification details accompanies this paper as an evidence file.

The method has known limits, which are stated rather than concealed. A narrative review is vulnerable to selection bias, because the reviewer chooses the conversation; it cannot estimate an average effect; and its conclusions are arguments, not measurements. Its compensating strength is that it can do what a meta-analysis cannot, namely examine concepts, trace how a practice acquires meaning inside a policy system, and test whether the claims made for a paradigm survive contact with validity theory and with the administrative realities of schooling. Where this paper makes empirical claims, they are tied to the strongest available syntheses and are hedged to the strength of those syntheses. Where it makes normative claims, they are presented as conditions and judgements, open to dispute.

RESULT AND DISCUSSION

3.1 What Scores Do That Narratives Cannot

The case for the score begins with functions that are easy to overlook because they are institutional rather than pedagogical. Scores aggregate. They allow a system to describe the attainment of a cohort, compare schools, monitor equity between groups, and detect whether standards are rising or falling, and no narrative corpus, however rich, can be averaged across ten thousand learners without first being converted back into categories or numbers. Scores select. Where places in the next stage of education or in employment are scarce, some defensible and impersonal ordering device is required, and a narrative

read by an admissions tutor who knows the applicant's school carries obvious risks of favouritism that a common examination is designed to suppress. Scores certify. A qualification is a public promise that a stranger can rely on, and its credibility depends on the judgement being detachable from the relationship in which it was formed. Scores also discipline judgement itself, because a common scale forces teachers to confront discrepancies between their standards and those of colleagues, and moderation around marks is one of the few reliable mechanisms by which a profession calibrates itself.

Psychometrically, the score is a technology for making an inference defensible. Kane's argument based approach to validation insists that a score interpretation rests on a chain of inferences, from scoring to generalisation to extrapolation to implication, each of which requires its own warrant and backing [8]. Numbers make that chain inspectable. Reliability can be estimated, measurement error can be quantified, bias can be investigated statistically across groups, and equating can place performances from different occasions on a common scale. Messick's unified account of validity, with its insistence that the social consequences of assessment belong inside validity itself, was developed for scored tests, and its apparatus of evidence presupposes performances that have been rendered comparable [9]. None of this shows that scores are valid by nature; both writers are clear that validity belongs to interpretations and uses, not to instruments. It shows instead that narrative assessment inherits a genuine problem. A story is singular, contextual, and authored, and the qualities that make it pedagogically powerful are the same qualities that make it hard to audit, hard to compare, and hard to defend when a disappointed family challenges it.

The social functions deserve equal weight, because they explain why reform efforts that ignore them tend to collapse. Broadfoot situates assessment inside the politics of the modern state, where it operates as a technology of control and legitimation as much as a tool of learning, supplying the apparently objective evidence by which systems justify the unequal distribution of educational opportunity [10]. Stobart similarly describes how assessment came to serve accountability regimes whose demand for comparable numbers reshaped curriculum and teaching, so that what is measured displaces what is valued [11]. Parents, too, are not merely obstacles to reform. For most families a grade is the only legible signal they receive about whether a child is thriving, and its removal without a credible replacement reads as concealment rather than liberation. The conclusion of this subsection is therefore symmetrical to the critique that follows. Scores are not a distortion imposed on an otherwise innocent process. They do real work, social, psychometric, and communicative, and a score-free paradigm must either perform that work differently or decline it explicitly and accept the consequences.

3.2 What Narratives Reveal That Scores Erase

If the score's virtue is compression, its vice is the same. Compression deletes everything that does not fit the scale, and what it deletes is often the learning itself. A mark of seven out of ten does not show which idea the learner has almost grasped, which misconception is organising their errors, what strategy they chose, how they responded to difficulty, or what they should do next. Shepard's influential account of assessment in a learning culture argues that the dominant measurement paradigm, inherited from behaviourist and psychometric traditions, sits uneasily with contemporary accounts of learning as constructive, situated, and social, and that classroom assessment must be rethought so that it reveals thinking rather than merely tallying its products [12]. Gipps made the parallel argument a few years earlier, proposing a move beyond testing toward a broader theory of

educational assessment in which teacher judgement, observation, and a range of evidence over time replace the single occasion examination as the paradigm case [13]. Drummond, writing about the assessment of young children, insisted that assessment is essentially provisional, partial, tentative, exploratory, and inevitably incomplete, a characterisation that narrative forms can acknowledge openly while a precise looking number tends to conceal [14].

Narrative assessment operationalises these claims. Carr's Learning Stories were developed in New Zealand early childhood settings as an assessment practice congruent with a curriculum organised around wellbeing, belonging, contribution, communication, and exploration rather than around atomised skill targets [6]. A Learning Story documents an episode of learning, interprets it against those broad outcomes and against the dispositions of a learner who is ready, willing, and able, and proposes a next step, often in language addressed to the child and shared with the family. Carr and Lee later extended the account, showing how such stories construct learner identities over time, so that the record of assessment becomes a resource a child can revisit in order to see themselves as a learner with a history and a trajectory [15]. The institutional setting matters. Learning Stories grew up inside Te Whariki, the New Zealand early childhood curriculum first published in 1996 and updated in 2017, whose woven mat metaphor and reduced set of twenty learning outcomes deliberately left space for local curriculum making and for assessment that honours the child's family, language, and culture [16]. Narrative assessment here is not a technique bolted onto a standard curriculum. It is the assessment form that a particular curriculum philosophy requires.

Three kinds of knowledge survive in narrative that vanish in scores. The first is process: how a learner approached a task, what they tried, where they hesitated, and how they recovered. The second is context: the relationships, interests, prior experiences, and cultural resources that made the performance possible, which are not noise around a true score but part of what the performance means. The third is disposition and identity: persistence, curiosity, willingness to collaborate, and the learner's own sense of becoming someone who can do this kind of thing. For young children, for learners with complex additional needs, for creative and practical domains, and for any learning whose value lies in judgement exercised under uncertainty, these are not decorative extras. They are the substance. Narrative also changes the audience and the temporality of assessment. A story written to be read by the child and the family, and added to across a term, makes assessment a continuing conversation rather than a periodic verdict, and it invites the learner into the assessment rather than positioning them as its object.

3.3 The Formative Paradigm and the Evidence on Learning Effects

Formative assessment is the older and more extensively researched of the two paradigms, and its mechanism is worth stating precisely, because imprecision about the mechanism is what allowed inflated claims to circulate. Sadler's classic analysis holds that formative assessment depends on feedback, and that feedback is information about the gap between a learner's actual level and a reference level, information that is formative only when it is used to alter the gap [17]. Sadler added three conditions that remain the sternest test of any formative practice: the learner must possess a concept of the standard or goal being aimed for, must be able to compare their current performance with that standard, and must be able to act to close the gap. On this account most of what schools call feedback fails, not because it is unkind but because the learner has no usable concept of quality and therefore cannot self monitor. Hattie and Timperley systematised the feedback literature into a

model built on three questions, where am I going, how am I going, and where to next, and on four levels at which feedback operates, the task, the process, self regulation, and the self, finding that feedback directed at the self as praise or blame is the least powerful and that feedback which builds self regulation is the most consequential [18]. Shute's review of formative feedback reached convergent practical conclusions, that effective formative feedback is nonevaluative, supportive, timely, and specific, focused on the task rather than the learner, and delivered in manageable units [19].

Set against this theoretical coherence, the quantitative evidence is real but modest, and it is heterogeneous in ways that matter for policy. The feedback intervention meta-analysis of Kluger and DeNisi, spanning 607 effect sizes, found an average effect of 0.41 on performance, but also found that more than one third of feedback interventions decreased performance, typically where feedback drew attention away from the task and toward the self [20]. Kingston and Nash, applying strict inclusion criteria to the classroom formative assessment literature, found only thirteen studies adequate to the question and a weighted mean effect of 0.20 [4]. Bennett's critical review explains part of the discrepancy: the term formative assessment covers such different practices, from commercial item banks to Socratic questioning, that averaging across them conceals more than it reveals, and several of the sources underpinning the most quoted effect sizes do not bear the weight placed on them [5]. Black and Wiliam themselves, in their later theoretical paper, moved away from effect size rhetoric toward a framework of five strategies, clarifying learning intentions, engineering discussion that elicits evidence, providing feedback that moves learners forward, activating students as resources for one another, and activating students as owners of their own learning, with the teacher's minute by minute decision making at the centre [21]. Wiliam subsequently defined assessment for learning by function rather than by instrument: a process is formative to the extent that evidence about achievement is elicited, interpreted, and used by teachers, learners, or their peers to make decisions about next steps that are likely to be better than the decisions that would have been made without it [22].

Two further bodies of work refine what the paradigm demands in practice. Torrance and Pryor, from classroom action research, distinguished convergent formative assessment, which checks whether a prescribed outcome has been reached, from divergent formative assessment, which seeks to discover what a learner understands and might do next, and they showed how accountability pressures push teachers toward the convergent form even when their intentions are divergent [23]. Harlen and James, earlier still, warned that formative and summative purposes differ in their requirements for reliability, in the reference base of their judgements, and in the focus of the information they use, so that summative judgements cannot honestly be manufactured by adding up formative ones, and that confusing the two corrupts both [24]. Nicol and Macfarlane-Dick repositioned feedback as a support for self regulated learning, setting out seven principles of good feedback practice that emphasise dialogue, self assessment, and closing the loop between comment and subsequent work [25]. Clark extended this line by arguing that formative assessment reaches its potential only when it is understood as assessment for self regulated learning, with learners generating internal feedback rather than merely receiving external comment [26]. Andrade and Brookhart have recently described classroom assessment as the co-regulation of learning, in which goals, monitoring, and adjustment are shared between teacher and students rather than imposed by either [27]. The pattern across these sources is consistent. The learning benefits attributed to

formative assessment are not produced by the absence of scores. They are produced by a demanding set of practices about goals, evidence, dialogue, and learner agency, which score-free reporting may support but cannot substitute for.

One experiment deserves separate mention because it speaks directly to the removal of marks. Butler compared the effects of comments alone, grades alone, and grades together with comments on students' interest and performance, and found that comments alone sustained and improved both, that grades undermined them, and that the combination behaved like grades alone, as though the grade cancelled the comment [28]. The study is small and old, and it should not be asked to carry more than it can. But its mechanism is plausible and repeatedly echoed in practice: a symbol that invites comparison with peers and judgement of the self draws attention away from the information that would improve the work. It is the single most cited empirical reason for believing that removing the score is not a neutral reporting choice but an intervention in how learners attend.

3.4 Risks, Costs, and Failure Modes

A critical review must be as rigorous about the risks of score-free assessment as about the failures of grading, because reforms fail in practice more often than they fail in principle. The first risk is workload. A defensible narrative is slow to produce. It requires observation, selection of evidence, interpretation against a framework, drafting in language a family can read, and filing so that the record accumulates rather than evaporates. Multiplied across a class of thirty children and several domains, narrative assessment can consume the hours that formative teaching itself requires, and a policy that mandates rich narrative reporting without withdrawing something else will be quietly converted by exhausted teachers into formulaic text, at which point the narrative carries no more information than the grade it replaced. Earl's account of assessment as learning, which places the student at the centre of connecting assessment to learning, is explicit that such practices depend on classroom conditions, time, and teacher learning that systems must deliberately build [29]. Carless, studying award winning university teachers, similarly found that productive assessment practice rests on trust, dialogue, and feedback processes that are designed into courses rather than appended to them, and that institutional demands for auditable marks repeatedly pull against them [30].

The second risk is bias. A score from a common instrument is not free of bias, but its bias can be investigated; a narrative judgement is formed inside a relationship, and the warmth, expectations, and cultural assumptions of that relationship enter the text. Teachers write more generously about children they know well and whose families they understand, they describe the same behaviour differently depending on the child's background, and they reach for the vocabulary available in their setting, which may flatter some learners and flatten others. Messick's consequential aspect of validity applies with full force here, because an assessment whose systematic distortions track social advantage is not merely imprecise but unjust [9]. Narrative traditions from communities whose knowledge is carried in story offer a partial answer, since they show that rigour in narrative judgement is possible, but they achieve it through communal scrutiny, shared protocols, and long apprenticeship, not through the assumption that writing prose is naturally fair.

The third risk is comparability and the fourth is demand. Without common reference points, a parent cannot tell whether a warmly written report describes a child who is thriving or a child whose school has lowered its expectations, and a receiving teacher or institution cannot plan. Systems respond to this vacuum predictably: they reintroduce

numbers through the back door, in the form of coded levels, traffic light categories, or standardised descriptors that are grades in prose costume, while adding the labour of writing on top. Boud and Molloy document a version of this dynamic in higher and professional education, where feedback processes are distorted by audit requirements and by the expectation that every judgement be defensible as a mark [31]. Sambell, McDowell, and Montgomery describe assessment for learning reforms repeatedly bending back toward summative norms because students, quite rationally, prioritise whatever counts toward the final classification [32]. The general lesson is that demand for numbers is not a failure of understanding that better communication will cure. It is a rational response by families and institutions to genuine uncertainty about standards and to real competition for scarce opportunities. Any score-free design that does not provide an alternative answer to the question how is my child doing compared with what should be expected will be defeated by that question.

3.5 Validity for Judgements That Are Not Numbers

The deepest objection to narrative assessment is not practical but epistemic: how would anyone know whether a narrative judgement is any good? The resources for an answer exist, although they require validity theory to be used more imaginatively than it usually is in policy debate. Messick's unified conception already insists that validity is a property of inferences and actions, not of instruments, and that it applies to any mode of assessment, so nothing in the framework restricts it to tests [9]. Kane's argument based approach is even more portable, because it asks the assessor to state the interpretation and use being proposed, to lay out the chain of inferences that connects an observed performance to a decision, and then to assemble evidence for the inferences that are most doubtful [8]. For narrative assessment the doubtful inferences are usually generalisation, since a story reports selected episodes and must warrant a claim about what a learner characteristically does, and extrapolation, since a claim about learning in one setting must support expectations in another. Naming those inferences converts a vague anxiety about subjectivity into a research and moderation agenda: sample more episodes, sample them across contexts, and test whether independent readers reach compatible interpretations.

Moss pressed this line furthest, arguing that conventional validity theory fits classroom assessment poorly, because it presupposes standardised conditions and a single integrated interpretation, whereas classroom knowledge is built incrementally, from multiple perspectives, in the service of local decisions [33]. Drawing on sociocultural theory and hermeneutics, she proposed that validity for classroom assessment should also be judged by the quality of the interpretive conversation: whether evidence is read against rival interpretations, whether the learner's own perspective is included, and whether the resulting understanding supports better teaching. This is not a lowering of standards. It is a recognition that the standard appropriate to a decision depends on the decision, and that the stakes, reversibility, and proximity of classroom decisions differ from those of certification. Where a narrative judgement feeds a high stakes and irreversible decision, the psychometric demands return in full, and narrative evidence alone is unlikely to meet them. Where it feeds next week's teaching, a well evidenced, moderated, and contestable story may be more valid, in the only sense that matters, than a precise score on a construct nobody can define. Validity theory thus yields the paper's central analytic conclusion: score-free assessment is not a single thing to be accepted or rejected, but a family of practices whose defensibility varies with the inference and the stakes attached to it.

3.6 Hybrid Designs: Separating Purposes Rather Than Averaging Them

Most real systems that have moved away from scores have not abolished numbers. They have reassigned them. The design principle that emerges from the literature is separation of purposes in time, in audience, and in instrument. During learning, assessment is formative and narrative: evidence is gathered, feedback is given without marks, learners assess themselves and one another against shared criteria, and records are written to be read by the learner and the family. At defined transition points, where selection or certification genuinely requires it, a separate summative process makes the necessary judgement, drawing on a body of evidence that may include narrative records but is conducted under its own rules of moderation and comparability. Harlen and James proposed exactly this architecture, arguing that the two purposes should be linked by procedure rather than by arithmetic, with summative judgements formed from a review of evidence at the point of reporting instead of accumulated from marks awarded along the way [24]. The OECD review of evaluation and assessment frameworks reaches a compatible conclusion at system level, recommending that countries seek synergies between classroom formative practice, school evaluation, and system monitoring while protecting each from the distortions of the others, and warning that using the same instrument for improvement and for accountability reliably corrupts the improvement function [34].

Several practical devices recur in hybrid designs. Criterion referenced descriptors, written as descriptions of what work looks like at successive levels of quality, give narrative judgement a shared spine without reducing it to a number. Moderation meetings, in which teachers read one another's narratives and evidence and must defend their interpretations, build the common standards that comparability requires and double as professional learning. Exemplars, annotated collections of real work and real stories, make tacit standards visible to teachers, students, and parents. Digital portfolios allow narrative records to accumulate with their evidence attached, so that a claim in prose can be traced back to the work that supports it. Reporting can be split into two documents, a narrative account addressed to the family about progress, strengths, and next steps, and a concise standards statement that locates the learner against curriculum expectations in plain categories. Each device costs time and each can degenerate into compliance, but together they answer the legitimate objections of the previous sections without reinstating the mark as the unit of classroom life. The hybrid is therefore not a compromise born of timidity. It is the form the evidence points to, because it refuses the false choice between information that helps learning and information that a system can trust.

3.7 Kurikulum Merdeka and the Indonesian Opening

Indonesia offers an instructive test of these arguments because its current curriculum reform moves, on paper, in the direction this paper has been describing. Kurikulum Merdeka, introduced as an option for schools from 2022 and progressively generalised, restructures assessment into three functions. Diagnostic assessment is conducted at the beginning of a learning period to identify learners' competencies, strengths, and weaknesses so that teaching can be planned accordingly. Formative assessment is conducted during learning to monitor progress and provide feedback for the improvement of both teaching and learning. Summative assessment is conducted at the end of a period to evaluate achievement of learning objectives. The official guidance issued by the Ministry of Education, Culture, Research, and Technology for learning and assessment under Kurikulum Merdeka states the formative principle with unusual clarity: the results

of formative assessment are not used to determine report card grades, promotion, or graduation, but to improve the learning process [35]. Teaching at the right level, differentiated according to diagnostic evidence, is the pedagogical companion of that principle, and the report card under the curriculum combines numeric or descriptive summative results with descriptive accounts of the learner's highest and lowest achievements and of character development through the Pancasila learner profile projects.

Read against the literature reviewed here, the Indonesian design is a hybrid of the kind Section 3.6 describes, and its strengths and vulnerabilities are predictable. Its strength is that the separation of formative from summative purposes is stated in official guidance rather than left to individual schools, which protects teachers who give feedback without marks from the accusation that they are failing to assess. Its vulnerability is the familiar one. Where school leadership, district monitoring, parental expectation, and entry to the next level of schooling continue to reward a single aggregate number, formative assessment is liable to be performed as a sequence of small summative tests, exactly the conflation Harlen and James warned against [24] and the convergent narrowing Torrance and Pryor documented [23]. A second vulnerability is capacity and workload: diagnostic and formative practice of the quality Sadler's conditions require depends on teachers holding a clear concept of the standard and being able to compare current performance with it, which is a matter of professional learning and of time, not of regulation [17]. A third is equity of narrative quality, since descriptive reporting will be richer in well resourced schools unless moderation and exemplars are supplied at system level. Indonesia therefore illustrates the general argument of this paper in concrete form. Policy can legitimate assessment without scores, and Kurikulum Merdeka does so more explicitly than many wealthier systems, but legitimacy at the level of guidance converts into practice only where the surrounding incentives, workload settlement, and moderation infrastructure are aligned with it.

3.8 Conditions Under Which Score-Free Assessment Is Defensible at Scale

The review supports neither abolitionism nor complacency, and it converges on a set of conditions. Score-free assessment is defensible at scale, first, where the purpose of the assessment is learning and the decision it feeds is proximal and reversible, so that the cost of an imperfect judgement is a better next lesson rather than a foreclosed future. Second, it requires shared criteria and exemplars that make quality visible before work begins, since without a public concept of the standard neither feedback nor narrative can do its work. Third, it requires moderation as routine professional infrastructure, with teachers reading and challenging one another's judgements and evidence, because comparability in a narrative system is produced socially or not at all. Fourth, it requires an honest workload settlement, in which narrative assessment replaces other reporting and accountability demands rather than being added to them, with class sizes and timetabled time that make observation and writing possible. Fifth, it requires reporting designed with families rather than for them, giving parents both the story and a plain statement of where a learner stands against expectations, so that trust does not depend on pretending the comparative question away. Sixth, it requires that selection and certification, where they are genuinely needed, be handled by separate and openly defended processes, so that classroom assessment is not silently required to do that work and corrupted by it. Seventh, it requires system monitoring to be conducted by instruments designed for that purpose, such as sampled national assessment, rather than by aggregating classroom judgements, which is the surest route to destroying their formative value. Where these conditions are absent, removing

scores does not liberate learning. It removes information, shifts advantage toward families best able to read an opaque system, and leaves teachers carrying an impossible double burden. Where they are present, the evidence and the theory together suggest that assessment without scores can be not merely humane but more valid for the purposes it claims to serve.

CONCLUSION

Assessment without scores is defensible neither as a slogan nor as a wholesale replacement for the measurement functions that modern education systems genuinely need, but as a disciplined reallocation of judgement: scores and standardised instruments should carry the work of selection, certification, aggregation, and system monitoring for which their portability and auditability are indispensable, while narrative and formative assessment should carry the work of learning itself, because they alone preserve the process, context, disposition, and next step that a number erases, because the evidence for formative practice, though more modest than its early advocates claimed, supports it when goals are shared, feedback is usable, and learners are made agents of their own improvement, and because appending a grade can cancel the very comment meant to help. Kurikulum Merdeka shows that a national curriculum can state this separation officially, declaring that formative results serve the improvement of learning rather than the determination of grades, promotion, or graduation, yet the same case shows that policy language succeeds only where shared criteria, moderation, exemplars, family facing reporting, separate certification processes, and an honest settlement of teacher workload are built beneath it; where they are not, score-free assessment collapses into prose grades, hidden inequity, and teacher exhaustion, and where they are, it offers the most credible available route to an assessment culture that tells learners the truth about their learning in a form they can act on.

REFERENCES

[1] S. M. Brookhart, T. R. Guskey, A. J. Bowers, S. McMillan, J. K. Smith, L. F. Smith, M. Stevens, and M. Welsh, "A century of grading research: Meaning and value in the most common educational measure," Review of Educational Research, vol. 86, no. 4, pp. 803-848, 2016.

[2] J. Schneider and E. Hutt, "Making the grade: A history of the A-F marking scheme," Journal of Curriculum Studies, vol. 46, no. 2, pp. 201-224, 2014.

[3] P. Black and D. Wiliam, "Assessment and classroom learning," Assessment in Education: Principles, Policy and Practice, vol. 5, no. 1, pp. 7-74, 1998.

[4] N. Kingston and B. Nash, "Formative assessment: A meta-analysis and a call for research," Educational Measurement: Issues and Practice, vol. 30, no. 4, pp. 28-37, 2011.

[5] R. E. Bennett, "Formative assessment: A critical review," Assessment in Education: Principles, Policy and Practice, vol. 18, no. 1, pp. 5-25, 2011.

[6] M. Carr, Assessment in Early Childhood Settings: Learning Stories. London: SAGE, 2001.

[7] F. M. Connelly and D. J. Clandinin, "Stories of experience and narrative inquiry," Educational Researcher, vol. 19, no. 5, pp. 2-14, 1990.

[8] M. T. Kane, "Validating the interpretations and uses of test scores," Journal of Educational Measurement, vol. 50, no. 1, pp. 1-73, 2013.

[9] S. Messick, "Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning," American Psychologist, vol. 50, no. 9, pp. 741-749, 1995.

[10] P. Broadfoot, An Introduction to Assessment. London: Continuum, 2007.

[11] G. Stobart, Testing Times: The Uses and Abuses of Assessment. London: Routledge, 2008.

[12] L. A. Shepard, "The role of assessment in a learning culture," Educational Researcher, vol. 29, no. 7, pp. 4-14, 2000.

[13] C. V. Gipps, Beyond Testing: Towards a Theory of Educational Assessment. London: Falmer Press, 1994.

[14] M. J. Drummond, Assessing Children's Learning, Classic ed. London: Routledge, 2012.

[15] M. Carr and W. Lee, Learning Stories: Constructing Learner Identities in Early Education. London: SAGE, 2012.

[16] Ministry of Education, Te Whariki: He whariki matauranga mo nga mokopuna o Aotearoa: Early Childhood Curriculum. Wellington: Ministry of Education, New Zealand, 2017.

[17] D. R. Sadler, "Formative assessment and the design of instructional systems," Instructional Science, vol. 18, no. 2, pp. 119-144, 1989.

[18] J. Hattie and H. Timperley, "The power of feedback," Review of Educational Research, vol. 77, no. 1, pp. 81-112, 2007.

[19] V. J. Shute, "Focus on formative feedback," Review of Educational Research, vol. 78, no. 1, pp. 153-189, 2008.

[20] A. N. Kluger and A. DeNisi, "The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory," Psychological Bulletin, vol. 119, no. 2, pp. 254-284, 1996.

[21] P. Black and D. Wiliam, "Developing the theory of formative assessment," Educational Assessment, Evaluation and Accountability, vol. 21, no. 1, pp. 5-31, 2009.

[22] D. Wiliam, "What is assessment for learning?" Studies in Educational Evaluation, vol. 37, no. 1, pp. 3-14, 2011.

[23] H. Torrance and J. Pryor, "Developing formative assessment in the classroom: Using action research to explore and modify theory," British Educational Research Journal, vol. 27, no. 5, pp. 615-631, 2001.

[24] W. Harlen and M. James, "Assessment and learning: Differences and relationships between formative and summative assessment," Assessment in Education: Principles, Policy and Practice, vol. 4, no. 3, pp. 365-379, 1997.

[25] D. J. Nicol and D. Macfarlane-Dick, "Formative assessment and self-regulated learning: A model and seven principles of good feedback practice," Studies in Higher Education, vol. 31, no. 2, pp. 199-218, 2006.

[26] I. Clark, "Formative assessment: Assessment is for self-regulated learning," Educational Psychology Review, vol. 24, no. 2, pp. 205-249, 2012.

[27] H. L. Andrade and S. M. Brookhart, "Classroom assessment as the co-regulation of learning," Assessment in Education: Principles, Policy and Practice, vol. 27, no. 4, pp. 350-372, 2020.

[28] R. Butler, "Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and performance," British Journal of Educational Psychology, vol. 58, no. 1, pp. 1-14, 1988.

[29] L. M. Earl, Assessment as Learning: Using Classroom Assessment to Maximize Student Learning, 2nd ed. Thousand Oaks: Corwin, 2013.

[30] D. Carless, Excellence in University Assessment: Learning from Award-Winning Practice. London: Routledge, 2015.

[31] D. Boud and E. Molloy, Eds., Feedback in Higher and Professional Education: Understanding It and Doing It Well. London: Routledge, 2013.

[32] K. Sambell, L. McDowell, and C. Montgomery, Assessment for Learning in Higher Education. London: Routledge, 2013.

[33] P. A. Moss, "Reconceptualizing validity for classroom assessment," Educational Measurement: Issues and Practice, vol. 22, no. 4, pp. 13-25, 2003.

[34] OECD, Synergies for Better Learning: An International Perspective on Evaluation and Assessment. Paris: OECD Publishing, 2013.

[35] Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi, Panduan Pembelajaran dan Asesmen: Pendidikan Anak Usia Dini, Pendidikan Dasar, dan Pendidikan Menengah. Jakarta: Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi, 2022.