

National Assessment and School Improvement in Indonesia: From Accountability to Diagnostic Use

Muhammad Amar Adly¹ Tri Alykson²

¹Universitas Islam Negeri Sumatera Utara, Medan, Indonesia

²Universitas Islam Negeri Sumatera Utara, Medan, Indonesia

Email: Amaradly73@yahoo.com¹ trialyksont@gmail.com²

Article History	Abstract
Received : 08-10-2026 Revised : 09-05-2026 Published : 10-01-2026	Indonesia replaced its high-stakes school-leaving examination with a national assessment designed on deliberately different principles: it samples learners rather than examining all of them, reports results at the level of the school rather than the individual, measures literacy and numeracy competence rather than curricular coverage, and surveys character and the learning environment alongside cognitive outcomes. The design is diagnostic by construction, yet the uses to which its results are put are determined downstream by local authorities and schools, and those uses may reintroduce the incentives the reform was intended to remove. This article aims to analyse the design logic of the national assessment, to identify the mechanisms through which a diagnostic instrument may acquire accountability stakes, and to assess what conditions are required for its results to inform improvement. The study is a qualitative policy analysis combining document analysis of the regulatory framework with critical synthesis of the assessment and accountability literature; no field data were collected. The analysis finds that the design successfully removes the individual-stakes mechanism that drove examination-oriented teaching, that the residual risk lies in institutional stakes attached to publicly reported school results, and that the capacity to interpret and act upon diagnostic data is the binding constraint on improvement. It recommends protection of the low-stakes character of reporting, investment in school-level data interpretation capability, and explicit separation of diagnostic reporting from any subsequent selection instrument.
Keywords: Assessment policy; Diagnostic assessment; Educational accountability; Minimum competency; School improvement.	

I. INTRODUCTION

The evaluation of learning outcomes is a statutory function of the Indonesian education system, exercised by government to monitor quality at national level.¹ For two decades that function was discharged through a national examination taken by every learner at the end of each level, carrying consequences for individual certification and, informally, for school reputation. The instrument was progressively criticised on grounds familiar from international experience: it narrowed teaching toward examined content, generated a substantial private tutoring market oriented toward examination technique, and produced information of limited diagnostic value because it reported aggregate scores without indicating what learners could or could not do.

The national assessment introduced in its place was designed on different principles.² It samples learners within each school rather than examining the full cohort; it reports at the level of the school and the region rather than the individual; it measures

¹Law Number 20 of 2003 concerning the National Education System, Articles 57–59.

²Regulation of the Minister of Education, Culture, Research and Technology Number 17 of 2021 concerning the National Assessment.

minimum competency in literacy and numeracy, understood as the capacity to apply reading and mathematical reasoning to unfamiliar material, rather than coverage of the syllabus; and it incorporates a character survey and a learning-environment survey alongside the cognitive instruments. Results are returned to schools through a reporting platform intended to support planning.

The design is diagnostic by construction, and its intellectual foundation is well established. The formative assessment literature demonstrates that assessment improves learning when it generates information usable by teachers and learners to close the gap between current and intended performance, and that this function is impaired when the same instrument carries consequences for those being assessed.³ The accountability literature establishes the corresponding negative proposition, namely that a quantitative indicator used to allocate rewards and sanctions will be subject to distortion and corruption pressures proportionate to the stakes attached.⁴

The policy problem is therefore not the design of the instrument, which is sound, but the governance of its use. Stakes are not a property of an assessment; they are attached to it by the decisions of those who receive its results. Where local authorities rank schools on publicly reported measures, where school leadership is evaluated against changes in those measures, or where a subsequent selection instrument is aligned with the assessment's content, the diagnostic design is preserved on paper while the behavioural consequences of a high-stakes regime reappear.

Indonesian scholarship on the national assessment has largely examined teacher and school readiness, and international scholarship on low-stakes diagnostic systems has developed largely in contexts with different governance structures. What has not been analysed is the specific mechanism by which stakes may attach downstream in the Indonesian configuration, and what safeguards would prevent it. The objectives of this study are accordingly to analyse the design logic of the assessment, to identify those mechanisms, and to specify the conditions under which its results can inform improvement. Three research questions follow: what does the design achieve relative to the instrument it replaced; through what mechanisms may stakes attach to a formally low-stakes instrument; and what capacity is required for diagnostic results to change practice?

II. RESEARCH METHODS

This study is a qualitative policy analysis based on document analysis and critical synthesis. Primary materials comprise the National Education System Law provisions on evaluation, the ministerial regulation governing the national assessment, and the regulatory instruments establishing the current curriculum framework within which the assessment operates. Secondary materials comprise the formative assessment literature, the test-based accountability literature, and scholarship on the measurement properties

³Paul Black and Dylan Wiliam, "Assessment and Classroom Learning," *Assessment in Education: Principles, Policy & Practice* 5, no. 1 (1998): 7–74.

⁴Donald T. Campbell, "Assessing the Impact of Planned Social Change," *Evaluation and Program Planning* 2, no. 1 (1979): 67–90; Daniel Koretz, *Measuring Up: What Educational Testing Really Tells Us* (Cambridge, MA: Harvard University Press, 2008), 235–259.

of low-stakes instruments. No field data were collected, and the study reports no measurement of assessment outcomes.

Analysis proceeded in three stages. First, the design features of the assessment were reconstructed from the regulatory instruments and compared, feature by feature, with those of the examination it replaced, in order to identify what the reform changed. Second, the mechanisms by which stakes may attach to results were derived from the accountability literature and mapped onto the Indonesian governance configuration, which distributes authority over schools between central ministry, provincial and district governments, and school leadership. Third, the requirements for diagnostic use were specified from the formative assessment literature and assessed against available institutional arrangements. The limitation of the design is that it identifies mechanisms and risks rather than measuring their operation; empirical verification would require study of how local authorities actually use reported results.

III. RESULTS AND DISCUSSION

A. What the Reform Changed

The first finding concerns the specific mechanism the reform removed. Under the previous instrument, every learner sat the examination and received an individual result that affected certification and progression. This design creates stakes at the level of the individual learner, and it is this feature, rather than the existence of a national test as such, that generates examination-oriented teaching: where a learner's future depends on performance, rational teachers and families allocate effort to the examined format.

Sampling and school-level reporting remove that mechanism entirely. A learner who does not know whether they will be sampled, and who receives no individual result, has no instrumental reason to prepare for the format. The reform therefore addressed the generating condition rather than its symptoms, which distinguishes it from reforms that retain individual examination while attempting to broaden its content.

The second design change concerns what is measured. An assessment of minimum competency in literacy and numeracy, constructed around the application of reading and mathematical reasoning to unfamiliar material, is substantially resistant to preparation by content coverage, because there is no defined body of content to cover. This property is important and underappreciated: competence-based instruments are harder to teach to than curriculum-coverage instruments, which means that improvement in results is more likely to reflect improvement in learning. The inclusion of learning-environment and character surveys extends the information base to conditions that schools can act upon, such as safety and instructional climate, which is precisely what a diagnostic system requires.

B. How Stakes May Reattach

The second finding identifies three mechanisms by which stakes may attach downstream. Table 1 summarises them.

Table 1. Mechanisms of Stakes Attachment to a Low-Stakes Instrument

Mechanism	Operation	Predicted distortion
Public ranking of schools	Reported results compared across schools	Preparation for the assessment format; roster effects
Evaluation of school leadership	Results used in appraisal or appointment	Pressure on teachers; selective attention to sampled grades
Local government targets	Regional aggregates set as performance targets	Administrative pressure transmitted to schools
Alignment of a selection test	A later selection instrument mirrors the format	Restoration of individual stakes by proxy
Commercial preparation market	Providers market preparation materials	Instructional time diverted to format familiarisation

Source: Analysis of policy design and accountability literature, 2026.

The first three mechanisms operate through the governance chain: results reported publicly become comparative, comparison becomes ranking, and ranking becomes a basis for administrative judgement even where no regulation provides for it. The accountability literature indicates that this progression is not a misuse requiring bad faith but the ordinary behaviour of administrative systems presented with comparable numbers.⁵ The predicted distortions follow: attention concentrates on the sampled grade levels, instructional time is allocated to format familiarisation, and the information value of the results degrades precisely as their administrative salience rises.

The fourth mechanism deserves separate emphasis because it would be the most consequential. Proposals to introduce a standardised test for secondary school leavers, whether for certification or for university selection, have recurred in Indonesian policy debate. If such an instrument were aligned in format with the national assessment, individual stakes would be restored by proxy, and the diagnostic instrument would become a rehearsal for the selective one. The design safeguard is straightforward and should be stated explicitly: any selection instrument must be constructed independently, and its relationship to the diagnostic assessment must be one of deliberate separation rather than of alignment.

C. Interpretive Capacity as the Binding Constraint

The third finding concerns use. Diagnostic information changes practice only where its recipients can interpret it and act upon it. The formative assessment literature is explicit that the effective element is not the data but the instructional response, and that the response requires the recipient to understand what the result indicates about learner thinking.⁶ A school receiving a report indicating that a proportion of learners has not reached a minimum competency threshold in numeracy possesses information about a shortfall, not about its causes, and the inferential step from one to the other requires professional capability.

Two consequences follow. Reporting platforms should present results in terms of what learners can and cannot do rather than in terms of scores and comparisons, since the former supports instructional inference while the latter supports ranking. And

⁵Laura S. Hamilton, Brian M. Stecher, and Stephen P. Klein, eds., *Making Sense of Test-Based Accountability in Education* (Santa Monica: RAND Corporation, 2002), 79–100.

⁶Black and Wiliam, “Assessment and Classroom Learning,” 20–30.

investment should be directed to school-level capability in interpreting data and designing responses, organised as sustained collaborative work rather than as briefing on the platform's operation. The distinction matters: training in how to read a dashboard is not training in how to infer an instructional problem from a pattern of results.

A further consideration concerns measurement properties. Low-stakes instruments carry a known limitation: where learners have no incentive to exert effort, results may understate competence, and the degree of understatement may vary systematically across schools and regions. This does not invalidate the design, since the comparison of a school with itself over time remains informative, but it does argue against fine-grained comparison across schools, which is an additional reason to resist the ranking mechanism identified above.

D. Recommendations

Three recommendations follow. First, the low-stakes character of the instrument should be protected by policy rather than assumed, through explicit provision that results may not be used for ranking, for the appraisal of individual school leaders, or as regional performance targets. A diagnostic design without a use restriction is an incomplete policy, because the design governs what the instrument measures while the restriction governs what it does.

Second, reporting should be framed in competence terms, describing what proportions of learners can perform which operations, and platforms should be evaluated by whether they support instructional inference rather than by their analytical sophistication.

Third, any future selection instrument should be developed independently, with an explicit statement that it is not aligned with the diagnostic assessment. This is the single most important safeguard, because alignment would reintroduce individual stakes and thereby undo the central achievement of the reform.

E. Comparative Reference: What Distinguishes Monitoring from Accountability

Comparative experience clarifies the distinction on which the Indonesian design depends. Some systems operate national instruments explicitly as monitoring devices: they sample learners, report at system and regional level, publish no school-level comparison, and attach no consequence to results. Others operate census-based instruments whose results are published by school and used in inspection, funding, or school leadership decisions. The two designs measure similar constructs and produce dissimilar behaviour, and the difference lies wholly in reporting and use rather than in the assessment instrument itself.

The documented behavioural consequences of the second design are consistent across jurisdictions: narrowing of the taught curriculum toward assessed subjects, reallocation of teaching attention toward learners near a reporting threshold, and score inflation in which measured performance rises without corresponding improvement in

the underlying competence.⁷ The last of these is the most damaging because it is invisible without an independent check: a system may record sustained improvement on its own instrument while learners' actual capability is unchanged, and the policy conclusions drawn from such data are correspondingly unreliable.

The Indonesian design sits closer to the first model, but it occupies an intermediate position in one respect: results are returned to schools and are, in practice, visible to regional authorities. This creates the conditions for comparison without the protections of a published-ranking regime, which typically includes methodological safeguards such as confidence intervals and contextual value-added adjustment. The intermediate position is therefore not automatically safer, and its safety depends on the use restrictions recommended above being adopted explicitly rather than assumed.

F. Limitations

Two limitations qualify the analysis. The study identifies mechanisms by which stakes may attach and predicts the distortions that would follow; it does not establish that those mechanisms are currently operating, which would require evidence of how regional authorities actually use reported results. Second, the comparative inferences drawn are from systems with different governance arrangements, and the transmission of administrative pressure through Indonesia's distributed authority over schools may operate differently from the more centralised inspection systems from which much of the accountability literature derives.

IV. CONCLUSION

This study asked what the national assessment reform changed, how stakes may reattach to a formally low-stakes instrument, and what capacity is required for diagnostic use. It found that the reform removed the individual-stakes mechanism that generated examination-oriented teaching, through sampling and school-level reporting, and that its competence-based construction makes results relatively resistant to preparation by content coverage. It found that residual risk lies in five downstream mechanisms, of which public ranking and the possible alignment of a future selection instrument are the most consequential. And it found that the binding constraint on improvement is the capability of schools to infer instructional problems from diagnostic patterns, which is a professional capacity rather than a technological one.

The theoretical implication is that the stakes of an assessment are a property of its governance rather than of its design, so that assessment reform is incomplete unless it regulates use alongside construction. The practical recommendations are to prohibit ranking and appraisal uses explicitly, to report in competence terms that support instructional inference, and to keep any selection instrument deliberately independent of the diagnostic assessment. Subsequent research should examine how local authorities actually use reported results, and should test whether school-level investment in data interpretation produces measurable changes in instructional response.

⁷Daniel Koretz, *Measuring Up: What Educational Testing Really Tells Us* (Cambridge, MA: Harvard University Press, 2008), 236–252.

REFERENCES

- Black, Paul, and Dylan Wiliam. "Assessment and Classroom Learning." *Assessment in Education: Principles, Policy & Practice* 5, no. 1 (1998): 7–74.
- Campbell, Donald T. "Assessing the Impact of Planned Social Change." *Evaluation and Program Planning* 2, no. 1 (1979): 67–90.
- Hamilton, Laura S., Brian M. Stecher, and Stephen P. Klein, eds. *Making Sense of Test-Based Accountability in Education*. Santa Monica: RAND Corporation, 2002.
- Koretz, Daniel. *Measuring Up: What Educational Testing Really Tells Us*. Cambridge, MA: Harvard University Press, 2008.

Regulations and Policy Documents

- Law Number 20 of 2003 concerning the National Education System.
- Regulation of the Minister of Education, Culture, Research and Technology Number 17 of 2021 concerning the National Assessment.
- Regulation of the Minister of Education, Culture, Research and Technology Number 12 of 2024 concerning the Curriculum for Early Childhood, Basic and Secondary Education.