

Assessment Reform after the National Examination: Purpose Alignment and Capability Gaps in Indonesian Schools

David Nwanna Dumbiri¹ Ingah Maulana²

¹Department of Vocational and Technical Education, Faculty of Education, University of Benin, Benin City, Nigeria

²Majelis Ulama Indonesia Provinsi Sumatera Utara

Email: ¹david.dumbiri@uniben.edu, ²ingah4001253005@uinsu.ac.id ,

Info Articles	Abstract
Article History Received : 05-09-2026 Revised : 29-09-2026 Published : 05-10-2026 Keywords: <i>assessment policy, classroom assessment, educational accountability, National Assessment, purpose validity</i>	The discontinuation of Indonesia's national examination and its replacement by a sampled, low-stakes National Assessment represents the most significant assessment reform in decades, yet schools continue to prepare students for an instrument that produces no individual results while the classroom capability the reform presupposes remains undeveloped. This study aims to examine the purposes the reconstructed framework assigns to its components, to identify the conditions under which system-level diagnostic information changes school practice, and to formulate the reconstruction required for the reform to achieve its aims. The research applies qualitative document analysis with a normative orientation, examining the National Education System Law, national education standards, and ministerial regulations on the National Assessment, assessment standards, process standards, and prior examination arrangements, analysed through the purpose-validity framework alongside scholarship on high-stakes testing, assessment for learning, and system evaluation design. The findings show that purposes are separated more coherently than under the previous arrangement but that the selection function is left largely ungoverned, producing a proliferation of ad hoc instruments; that informal accountability stakes have migrated to the diagnostic instrument, reproducing the distortion the reform was meant to remove; and that the vacated classroom space requires diagnostic assessment capability never developed under the examination regime.

Info Articles	Abstract
	The article recommends collaborative moderation as the primary implementation investment, explicit governance of selection, deliberate protection of the diagnostic instrument, and redefinition of supervision toward interpretive support.

I. INTRODUCTION

The discontinuation of the national examination and its replacement by a system-level assessment constitutes the most significant change in Indonesian assessment policy since the establishment of the national education standards. The national examination had functioned for decades as the principal instrument of accountability, the determinant of progression, and in practice if not in law the definition of what schooling was for. Its replacement, the National Assessment, is built on a different logic: it samples students rather than testing all of them, reports at the level of the school and the system rather than the individual, and measures reading literacy, numeracy, character, and learning environment rather than subject syllabus coverage.¹

The change is defensible on well-established grounds. High-stakes examinations that determine individual futures generate measurement distortion of a kind documented across many systems: instruction narrows to tested content, teaching converges on test format, and the indicator becomes progressively less informative about the underlying construct as pressure on it increases.² Removing individual stakes removes the mechanism that drives this distortion. Whether the new arrangement produces the improvement in learning that motivated it, however, depends on questions that the policy change does not by itself answer: what schools do with the information the assessment generates, what replaces the examination as the instrument of classroom assessment, and whether the accountability pressure that the national examination carried has genuinely dissipated or merely relocated.

The ideal condition is specified across several instruments. The National Education System Law requires that evaluation be conducted to monitor the process, progress, and

¹ Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Nomor 17 Tahun 2021 tentang Asesmen Nasional, Pasal 2-4.

² Donald T. Campbell, "Assessing the Impact of Planned Social Change," *Evaluation and Program Planning* 2, no. 1 (1979): 67-90; Daniel Koretz, *The Testing Charade: Pretending to Make Schools Better* (Chicago: University of Chicago Press, 2017), 38-62.

improvement of learning outcomes on a continuing basis.³ The assessment standard establishes that assessment of learners is carried out to ascertain achievement and to serve as a basis for improving the learning process, and that its principles include validity, comparability, transparency, and educational purpose.⁴ The National Assessment regulation frames the instrument explicitly as a diagnostic tool for system and school improvement rather than as a measure of individual attainment.⁵

The actual condition displays three recurring patterns. First, schools have in numerous cases begun preparing students for the National Assessment through practice items and drilling, notwithstanding that the sampling design and absence of individual stakes make such preparation largely pointless. Second, the assessment's diagnostic reports which are its principal contribution are underused, in part because school staff lack the capability to interpret them and in part because no institutional process requires or supports their use. Third, classroom assessment practice has changed less than the policy assumes: the descriptive, formative model the standards contemplate requires a set of teacher capabilities that were never developed under the examination regime, and their absence is not remedied by removing the examination.

Scholarly debate has addressed these questions along three lines. Measurement scholarship has documented the distorting effects of high-stakes testing and the conditions under which assessment information supports rather than displaces instructional judgement.⁶ Assessment-for-learning research has established the conditions under which classroom assessment improves learning, and the substantial gap between those conditions and ordinary practice.⁷ System-evaluation scholarship has examined how national assessment frameworks can be designed so that their components serve distinct purposes without undermining one another.⁸

³ Undang-Undang Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional, Pasal 58 ayat (1).

⁴ Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Nomor 21 Tahun 2022 tentang Standar Penilaian Pendidikan, Pasal 3 dan Pasal 5.

⁵ Permendikbudristek Nomor 17 Tahun 2021, Pasal 3 dan Pasal 5.

⁶ Sharon L. Nichols and David C. Berliner, *Collateral Damage: How High-Stakes Testing Corrupts America's Schools* (Cambridge: Harvard Education Press, 2007), 26–58; Daniel Koretz, *Measuring Up: What Educational Testing Really Tells Us* (Cambridge: Harvard University Press, 2008), 233–260.

⁷ Paul Black and Dylan Wiliam, "Assessment and Classroom Learning," *Assessment in Education: Principles, Policy and Practice* 5, no. 1 (1998): 7–74; Richard J. Stiggins, "Assessment Crisis: The Absence of Assessment FOR Learning," *Phi Delta Kappan* 83, no. 10 (2002): 758–765.

⁸ OECD, *Synergies for Better Learning: An International Perspective on Evaluation and Assessment* (Paris: OECD Publishing, 2013), 139–182.

The research gap this article addresses is the limited analysis of the Indonesian transition as a problem of purpose alignment across an assessment system rather than as the replacement of one instrument by another. Indonesian scholarship has extensively discussed teacher and student perceptions of the new arrangement and has reported on preparation practices, but has rarely asked whether the system as a whole now assigns coherent purposes to its components, or what institutional conditions would be required for the diagnostic information generated to change classroom practice. The novelty of the article lies in applying the distinction between assessment purposes to the reconstructed Indonesian system, and in arguing that the reform's principal vulnerability is not resistance to change but the absence of a classroom assessment capability to occupy the space the examination vacated.

Three research questions guide the analysis. First, what purposes does the reconstructed Indonesian assessment framework assign to its components, and are those purposes mutually consistent? Second, what conditions determine whether system-level diagnostic information changes school and classroom practice? Third, what reconstruction of policy, capability, and institutional process would enable the reform to achieve its stated aims?

II. LITERATURE REVIEW AND THEORETICAL FRAMEWORK

Assessment theory begins with the proposition that an assessment has no validity in the abstract: validity is a property of an interpretation and use, not of an instrument.⁹ The same test may yield valid inferences about system performance and invalid inferences about individual attainment, or the reverse. Newton's analysis distinguishes judgement-level, decision-level, and purpose-level descriptions of assessment, and shows that confusion among these levels is a persistent source of policy error. This framework is directly applicable to the Indonesian transition, in which an instrument designed for system diagnosis is at risk of being treated by schools as a measure of their standing.

The literature on high-stakes testing supplies the second theoretical component. Campbell's law states that the more a quantitative indicator is used for social decision-making, the more it will distort and corrupt the processes it is intended to monitor.¹⁰ Koretz's elaboration identifies the mechanisms: reallocation of instructional time toward tested content, coaching on test-specific features, and in some cases outright manipulation,

⁹ Paul E. Newton, "Clarifying the Purposes of Educational Assessment," *Assessment in Education: Principles, Policy and Practice* 14, no. 2 (2007): 149–170.

¹⁰ Campbell, "Assessing the Impact of Planned Social Change," 85.

all of which inflate scores without corresponding gains in the underlying capability.¹¹ The theoretical significance for the present analysis is that score inflation is a function of stakes rather than of test format, which implies that removing stakes is a more effective remedy than improving items and equally that stakes attaching informally to a low-stakes instrument will reproduce the distortion.

The third component is the assessment-for-learning tradition. Sadler's account identifies the conditions under which feedback improves performance: the learner must hold a concept of the standard, be able to compare their current performance with it, and be able to act to close the gap.¹² Black and Wiliam's review established the magnitude of the effect and its dependence on the use made of the information rather than on its collection.¹³ Wiliam's later work specifies the classroom practices involved and, importantly for policy, demonstrates that these practices are learnable but require sustained support over a period of years rather than exposure in a workshop.¹⁴

A fourth component concerns system evaluation design. Comparative analysis of evaluation and assessment frameworks identifies the central design problem as achieving complementarity: each component student assessment, teacher appraisal, school evaluation, system monitoring must serve a defined purpose, and the framework must prevent any component from being pressed into service for purposes it cannot validly support.¹⁵ The characteristic failure is the use of student assessment data for teacher or school accountability at a level of aggregation that the data cannot sustain, which reintroduces the stakes that distort measurement.

The theoretical gap addressed here lies in the application of these frameworks to a system undergoing rapid transition. Most of the literature analyses mature systems with established practices; comparatively little addresses the interim period during which an accountability instrument has been removed and its replacement has not yet acquired institutional meaning. That interim is precisely where Indonesian schools now stand, and it has distinctive properties: expectations formed under the previous regime persist, the new instrument's purpose is imperfectly understood, and the classroom capability that the new arrangement presupposes has not yet been built.

¹¹ Koretz, *The Testing Charade*, 38–62.

¹² D. Royce Sadler, "Formative Assessment and the Design of Instructional Systems," *Instructional Science* 18, no. 2 (1989): 119–144.

¹³ Black and Wiliam, "Assessment and Classroom Learning," 7–74.

¹⁴ Dylan Wiliam, *Embedded Formative Assessment* (Bloomington: Solution Tree Press, 2011), 45–72.

¹⁵ OECD, *Synergies for Better Learning*, 139–182.

III. RESEARCH METHOD

This study employs qualitative document analysis with a normative orientation, combining doctrinal examination of assessment regulation with conceptual analysis drawn from assessment theory. Document analysis is a systematic procedure for reviewing and interpreting documentary material in order to elicit meaning and develop understanding, appropriate where the object of inquiry is the design of a policy framework.¹⁶

Two approaches are combined. The statutory approach examines legislation and subordinate regulation governing evaluation and assessment, with attention to the relationship between instruments and to the purposes each assigns. The conceptual approach applies the purpose-validity framework to the assembled corpus, testing whether the purposes assigned are internally consistent and whether the inferences the framework invites are supportable by the instruments it specifies.

Legal and documentary materials are organised in three tiers. Primary materials comprise Law Number 20 of 2003 on the National Education System; Government Regulation Number 57 of 2021 as amended by Government Regulation Number 4 of 2022 on National Education Standards; Regulation of the Minister of Education, Culture, Research and Technology Number 17 of 2021 on the National Assessment; Regulation Number 21 of 2022 on Assessment Standards; Regulation Number 16 of 2022 on Process Standards; and the earlier regulation governing examinations conducted by educational units, which establishes the position from which the transition began. Secondary materials comprise peer-reviewed scholarship in educational measurement, assessment for learning, and system evaluation. Supporting materials comprise international comparative reports on assessment frameworks and on learning outcomes.¹⁷

Analysis proceeded in four stages. The regulatory corpus was coded for provisions assigning purposes to assessment components. The assigned purposes were then classified using the purpose-validity framework, distinguishing certification, selection, accountability, diagnosis, and instructional improvement. Points of inconsistency and of purpose overload were identified. Finally, conditions for the effective use of diagnostic information were derived from the research literature and tested against the institutional arrangements the corpus establishes. The analysis is interpretive and normative;

¹⁶ Glenn A. Bowen, "Document Analysis as a Qualitative Research Method," *Qualitative Research Journal* 9, no. 2 (2009): 27–40.

¹⁷ OECD, *PISA 2022 Results (Volume I): The State of Learning and Equity in Education* (Paris: OECD Publishing, 2023), 45–98.

propositions about school behaviour are derived from the design of the framework and from comparative evidence rather than from primary observation in Indonesian schools.

IV. RESULTS AND DISCUSSION

A. Purposes Assigned and the Problem of Overload

Analysis of the corpus identifies five distinct purposes assigned to assessment components: certification of individual attainment, selection into subsequent stages, accountability for institutional performance, diagnosis of system and school strengths, and improvement of classroom instruction. The reconstruction has redistributed these among components in a way that is substantially more coherent than the previous arrangement, in which the national examination carried certification, selection, accountability, and implicitly and unsuccessfully instructional improvement simultaneously.

Under the current framework, certification and the determination of graduation rest with the educational unit; the National Assessment carries system and school diagnosis; and classroom assessment carries instructional improvement.¹⁸ This separation is theoretically sound and reflects the design principle that components should not be asked to serve purposes their construction cannot support.

Two difficulties nonetheless arise. The first concerns selection, which the framework addresses least clearly. Selection into higher stages continues to operate, and where a nationally comparable measure is absent, selection instruments proliferate institutional entrance tests, achievement portfolios, and school-reported grades whose comparability across institutions is unestablished. The removal of a distorting instrument does not remove the underlying demand for selection, and a framework that does not address that demand will find it met by arrangements it does not govern. This is not an argument for restoring the examination; it is an argument for explicit attention to the selection function, which is currently the framework's least examined component.

The second concerns the informal migration of accountability to the National Assessment. Although the instrument is designed for diagnosis, its results are published at school level and enter the quality assurance apparatus. Where results carry reputational consequences for schools and their leaders, stakes exist irrespective of whether the regulation attaches them. Campbell's law does not distinguish formal from informal stakes; it responds to use. The observable phenomenon of schools drilling students for a sampled,

¹⁸ Permendikbudristek Nomor 21 Tahun 2022, Pasal 9 dan Pasal 10; Permendikbudristek Nomor 17 Tahun 2021, Pasal 3.

low-stakes instrument is direct evidence that informal stakes are operating, and it should be read as a warning rather than as a curiosity.

B. The Vacated Space: Classroom Assessment Capability

The most consequential finding of this analysis concerns what has replaced the national examination in the classroom. The examination, whatever its defects, supplied teachers with a clear specification of what to teach, a model of how attainment would be judged, and a shared reference point for standards. Its removal leaves a space that the new framework expects descriptive, formative assessment to occupy.

The capability required to occupy that space is substantial and was not developed under the previous regime. A teacher conducting formative assessment must be able to specify what competence looks like at successive levels, design tasks that elicit evidence of that competence, interpret student responses diagnostically rather than merely scoring them, and adjust instruction accordingly.¹⁹ These are demanding professional capabilities, and there is no reason to expect them to emerge spontaneously from the removal of an examination.

The evidence on developing these capabilities is consistent and sobering: change in assessment practice is slow, requires sustained collaborative work over a period measured in years rather than months, and depends heavily on teachers examining student work together against shared criteria.²⁰ Indonesian professional development provision has not been organised in this form. The implication is that the reform's central mechanism better classroom assessment producing better learning currently lacks the capability on which it depends, and that this capability gap, rather than resistance or misunderstanding, is the binding constraint.

A secondary consequence deserves note. In the absence of a shared external reference for standards, school-determined attainment judgements will vary across institutions in ways that are invisible to the system and consequential for students. Moderation the process by which teachers compare judgements against common exemplars to align standards is the established remedy, and it is largely absent from Indonesian practice. Its introduction would address both the comparability problem and, as a by-product, the capability problem, since moderation is among the most effective forms of assessment professional development available.

¹⁹ Wiliam, *Embedded Formative Assessment*, 45–72; Sadler, "Formative Assessment," 129–138.

²⁰ Linda Darling-Hammond, Maria E. Hyler, and Madelyn Gardner, *Effective Teacher Professional Development* (Palo Alto: Learning Policy Institute, 2017), 4–14.

C. Whether Diagnostic Information Changes Practice

The National Assessment generates school-level reports intended to inform improvement planning. Whether such information changes practice depends on conditions that comparative experience has identified with reasonable clarity.

The first condition is interpretive capability. A report showing that a school's students perform below expectation in reading literacy tells school staff what, not why. Moving from the former to the latter requires the capacity to generate hypotheses about causes, to examine local evidence bearing on those hypotheses, and to design a response. Where this capability is absent, the characteristic response to a diagnostic report is either dismissal or the adoption of a generic intervention unconnected to the specific pattern the data reveal.

The second is institutional process. Information changes practice where a process requires its examination, allocates time for that examination, involves the people whose practice would change, and revisits the question later. Where improvement planning is an annual documentary exercise completed by a small group for submission upward, the report will appear in the document and not in the classroom.

The third is the absence of threat. A substantial body of evidence indicates that performance information is used constructively where it is not experienced as a judgement on those receiving it, and defensively where it is.²¹ This creates a genuine tension with the transparency that publication serves, and it is the reason the design question of what is published, at what level of aggregation, and with what framing is not a presentational detail.

Table 1. Assessment Purposes and Their Institutional Requirements

Purpose	Component Carrying It	Requirement for Valid Use	Current Vulnerability
Certification of attainment	Educational unit assessment	Comparable standards across institutions	No moderation process; comparability unestablished
Selection into later stages	Institutional instruments; school grades	Fair, comparable, transparent criteria	Least addressed component; proliferation of ad hoc instruments

²¹ Koretz, *Measuring Up*, 233–260; OECD, *Synergies for Better Learning*, 139–182.

Purpose	Component Carrying It	Requirement for Valid Use	Current Vulnerability
System and school diagnosis	National Assessment	Low stakes; interpretive capability; improvement process	Informal stakes; limited capacity to interpret reports
Instructional improvement	Classroom formative assessment	Teacher capability in diagnostic assessment	Capability not developed under prior regime
Institutional accountability	Quality assurance apparatus	Indicators that do not corrupt the measures they use	Diagnostic instrument drawn into accountability use

Source: Author's analysis of the regulatory corpus and assessment literature, 2026.

D. Reading and Numeracy as Constructs, and What Follows for Teaching

The National Assessment measures reading literacy and numeracy conceived as the capacity to use texts and mathematical reasoning in varied contexts, rather than as the recall of curricular content. This choice of construct has an implication for teaching that the policy discussion has largely left implicit: these capabilities are developed across the curriculum and over years, and they are not the property of the language and mathematics subjects alone.

Reading literacy in the sense the assessment measures is developed when students read extended, demanding texts in history, science, and religious studies, and are required to interpret, infer, evaluate, and connect. Numeracy in the relevant sense is developed when quantitative reasoning is required in contexts outside mathematics lessons. A school that responds to a low literacy result by increasing the hours of language instruction has misread the construct; a school that responds by examining what students are asked to read across all subjects has read it correctly.

This has a further consequence for the drilling phenomenon noted above. Practising assessment items is not merely pointless given the sampling design; it is counterproductive with respect to the construct, since the capability being measured is developed through extended engagement with substantive material and not through exposure to item formats. The clearest statement a system could make about its intentions would be to say this

explicitly and repeatedly, and to ensure that supervision practice does not reward the opposite.

E. Comparative Perspectives on Low-Stakes System Monitoring

Comparative experience with low-stakes system monitoring suggests three lessons of relevance.

The first concerns stability of purpose. Systems that have maintained genuinely low-stakes monitoring have generally done so by maintaining an explicit separation between monitoring instruments and accountability instruments over a long period, and by resisting recurrent pressure to use monitoring data for accountability purposes as it becomes available. The pressure is persistent, because the data exist and the accountability demand is real; the discipline required is correspondingly substantial.

The second concerns the intermediate layer. Where system-level information has changed practice, it has generally done so through an intermediate body a district team, a school improvement partner, a network of schools with the capability to interpret and the mandate to support. Reports delivered directly to schools without such a layer are much less likely to be used, because the interpretive capability the reports presuppose is precisely what schools receiving poor results are least likely to possess. Indonesia's supervisory apparatus is the obvious candidate for this role, but its current orientation toward administrative compliance would require substantial redefinition.

The third concerns sampling and its communication. Sampled assessment is unfamiliar in a system accustomed to census testing, and the implications that individual results are not produced, that school-level estimates carry uncertainty, that year-to-year fluctuation may reflect sampling rather than change are not widely understood. Misunderstanding of this kind is a direct source of misuse: a school that treats a fluctuating estimate as a precise performance measure will react to noise.

F. Reconstruction: Aligning Capability with Design

The analysis supports four reconstruction proposals.

First, invest in classroom assessment capability as the reform's primary implementation task rather than as an accompaniment to it. The concrete form is collaborative moderation: teachers within and across schools examining student work against shared criteria, over an extended period, with facilitation. This simultaneously builds diagnostic capability, aligns standards across institutions, and addresses the comparability problem created by devolving certification.

Second, define the selection function explicitly. Leaving selection to be organised by receiving institutions produces a proliferation of instruments whose fairness and comparability are unexamined, and the resulting burden falls disproportionately on students without the resources to prepare for multiple processes. A clear national framework for what may be used in selection, and on what conditions, is a matter of equity rather than of administrative tidiness.

Third, protect the diagnostic instrument from accountability use. This requires attention to how results are published and framed, an explicit statement that preparation is neither expected nor useful, and most importantly supervision practice that does not inquire about assessment results in ways that create stakes the regulation disclaims.

Fourth, redefine the supervisory role toward interpretive support. Supervisors who can help a school read its diagnostic report, generate hypotheses, and design a response would convert the assessment from an information product into an improvement mechanism. This is a substantial change in the professional content of the supervisory role, and it would require its own programme of development.

G. Equity Effects of the Transition

The distributional consequences of the reform deserve separate treatment, because they run in opposite directions and are easily misread.

On one side, the removal of a high-stakes examination is straightforwardly equity-enhancing. Examination preparation was a market, and participation in that market was determined by household resources. Students whose families could purchase tutoring, preparation courses, and practice materials held an advantage that was unrelated to the quality of their schooling, and the examination's stakes converted that advantage into differential access to subsequent opportunity. Removing the examination removes the point at which purchased preparation converts into institutional advantage a substantial gain, and one that has attracted less attention than it deserves in Indonesian policy discussion.

On the other side, devolving certification to educational units without establishing comparability produces a different inequity. Where a school's assessment of attainment is not referenced to any common standard, the meaning of a given grade varies by institution, and the variation is not random: it correlates with institutional resources, with the professional capability of staff, and with local expectations about what constitutes acceptable performance. A student from a school with weak internal standards may receive a favourable record that misrepresents their preparation, which harms them at the point of

subsequent transition; a student from a demanding school may be disadvantaged by a record that is more honest than the comparison. Neither outcome is visible to the system, because nothing measures it.

The selection question compounds this. Where receiving institutions do not trust school records and the absence of comparability gives them reason not to they will construct their own instruments. Multiple institutional entrance processes impose costs in time, travel, fees, and information, and those costs fall most heavily on precisely the students the reform was intended to benefit. It is possible for a reform to remove one mechanism of advantage and install another, and careful attention to the selection function is what prevents this.

A third equity consideration concerns the diagnostic instrument itself. Publishing school-level results makes visible the substantial disparities between schools serving different populations. Visibility is valuable, and the case for transparency is strong. But visibility without a resourced improvement response converts diagnosis into labelling, and the schools labelled will be those serving the most disadvantaged communities, which are also the schools least able to convert a diagnosis into a remedy on their own. The equity value of the National Assessment therefore depends almost entirely on what the system does with the disparities it reveals an observation that applies with equal force to the assessment frameworks of many other countries and that has been learned repeatedly and expensively.

H. Limitations of the Analysis and Alternative Readings

Intellectual honesty requires acknowledging that the analysis presented here rests on documentary evidence and comparative inference, and that alternative readings are available.

A first alternative holds that the capability gap identified above is transitional and self-correcting: that teachers, released from examination pressure, will develop formative practice through ordinary professional experience over a period of years. This reading is not unreasonable, and it finds some support in the observation that professional practice does evolve without programmatic intervention. The counter-consideration is that the evidence on assessment practice specifically indicates slow change even under deliberate support, and very limited spontaneous change without it. The prudent policy position does not depend on resolving this dispute: investment in capability is low-risk if the optimistic reading is correct and essential if it is not.

A second alternative holds that the drilling phenomenon reflects a transitional misunderstanding that will dissipate as the instrument becomes familiar. This too is plausible. But the persistence of preparation behaviour depends on whether informal stakes persist, and the analysis above suggests that they will so long as results enter quality assurance processes with reputational consequence. Familiarity with the instrument does not remove an incentive.

A third alternative advanced in public debate rather than in scholarship holds that removing the examination has removed a useful discipline, and that standards will decline without an external reference. This position deserves engagement rather than dismissal, because it identifies something real: the examination did supply a shared standard, and nothing currently replaces it. The disagreement is about remedy. Restoring a high-stakes examination would restore the standard together with the distortion. Establishing moderation would supply the standard without the stakes, at greater cost in professional time and at slower pace. The choice between these is a genuine policy judgement, and the case for moderation rests on the evidence that score inflation under high-stakes conditions makes the examination's apparent standard substantially less real than it appears.

Finally, the analysis is limited in an important respect: it examines the design of the framework and not its enactment. What schools actually do with diagnostic reports, how certification judgements are in fact made, and whether the selection instruments now in use are more or less equitable than the examination they replaced are empirical questions that documentary analysis cannot answer. The propositions advanced here should be read as hypotheses with a defensible theoretical basis, and the research agenda proposed below is designed to test them.

V. CONCLUSION

This article has examined the reconstruction of Indonesian assessment policy following the discontinuation of the national examination, and has argued that the reform's principal vulnerability lies not in its design but in the capability and institutional process required to realise it.

In response to the first research question, the framework assigns purposes to components in a manner substantially more coherent than the previous arrangement, separating certification, diagnosis, and instructional improvement; but it addresses the selection function least clearly, and it does not adequately protect the diagnostic instrument from informal accountability use, as the observable phenomenon of preparation for a sampled low-stakes assessment demonstrates. In response to the second,

diagnostic information changes practice only where interpretive capability, an institutional process requiring its use, and freedom from perceived threat are present; none of these can be assumed in the current Indonesian arrangement. In response to the third, reconstruction requires sustained investment in classroom assessment capability through collaborative moderation, explicit governance of the selection function, deliberate protection of the diagnostic instrument from accountability drift, and redefinition of the supervisory role toward interpretive support.

The theoretical implication is that assessment reform should be analysed as a redistribution of purposes across a system rather than as the substitution of one instrument for another, and that the capability required by the receiving components is the critical variable. Removing a distorting instrument is a necessary condition for improvement and not a sufficient one; what occupies the vacated space determines the outcome.

The practical implications are addressed to policymakers, who should treat classroom assessment capability as the reform's principal implementation task and fund it accordingly; to school leaders, who determine whether diagnostic information is examined collectively or filed; and to supervisory authorities, whose inquiries communicate the system's real expectations more effectively than its regulations do.

Further research should pursue three directions: empirical study of how schools actually use National Assessment reports and what distinguishes those that use them well; evaluation of collaborative moderation as a means of building assessment capability in Indonesian conditions; and analysis of the selection instruments that have emerged since the examination's discontinuation, with attention to their comparability and their distributional effects.

REFERENCES

- Black, Paul, and Dylan Wiliam. "Assessment and Classroom Learning." *Assessment in Education: Principles, Policy and Practice* 5, no. 1 (1998): 7-74.
- Bowen, Glenn A. "Document Analysis as a Qualitative Research Method." *Qualitative Research Journal* 9, no. 2 (2009): 27-40.
- Campbell, Donald T. "Assessing the Impact of Planned Social Change." *Evaluation and Program Planning* 2, no. 1 (1979): 67-90.
- Darling-Hammond, Linda, Maria E. Hyler, and Madelyn Gardner. *Effective Teacher Professional Development*. Palo Alto: Learning Policy Institute, 2017.
- Fullan, Michael. *The New Meaning of Educational Change*. 5th ed. New York: Teachers College Press, 2016.

-
- Hattie, John. *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. London: Routledge, 2009.
- Koretz, Daniel. *Measuring Up: What Educational Testing Really Tells Us*. Cambridge: Harvard University Press, 2008.
- Koretz, Daniel. *The Testing Charade: Pretending to Make Schools Better*. Chicago: University of Chicago Press, 2017.
- Newton, Paul E. "Clarifying the Purposes of Educational Assessment." *Assessment in Education: Principles, Policy and Practice* 14, no. 2 (2007): 149–170.
- Nichols, Sharon L., and David C. Berliner. *Collateral Damage: How High-Stakes Testing Corrupts America's Schools*. Cambridge: Harvard Education Press, 2007.
- OECD. *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*. Paris: OECD Publishing, 2023.
- OECD. *Synergies for Better Learning: An International Perspective on Evaluation and Assessment*. Paris: OECD Publishing, 2013.
- Peraturan Menteri Pendidikan dan Kebudayaan Nomor 43 Tahun 2019 tentang Penyelenggaraan Ujian yang Diselenggarakan Satuan Pendidikan dan Ujian Nasional.
- Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Nomor 17 Tahun 2021 tentang Asesmen Nasional.
- Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Nomor 16 Tahun 2022 tentang Standar Proses.
- Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Nomor 21 Tahun 2022 tentang Standar Penilaian Pendidikan.
- Peraturan Pemerintah Nomor 57 Tahun 2021 tentang Standar Nasional Pendidikan sebagaimana diubah dengan Peraturan Pemerintah Nomor 4 Tahun 2022.
- Sadler, D. Royce. "Formative Assessment and the Design of Instructional Systems." *Instructional Science* 18, no. 2 (1989): 119–144.
- Stiggins, Richard J. "Assessment Crisis: The Absence of Assessment FOR Learning." *Phi Delta Kappan* 83, no. 10 (2002): 758–765.
- Tyack, David, and Larry Cuban. *Tinkering toward Utopia: A Century of Public School Reform*. Cambridge: Harvard University Press, 1995.
- Undang-Undang Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional.
- Wiliam, Dylan. *Embedded Formative Assessment*. Bloomington: Solution Tree Press, 2011.
- World Bank. *The State of Global Learning Poverty: 2022 Update*. Washington: World Bank, 2022.