IMPROVING THE MONITORING OF EDUCATION QUALITY IN GENERAL SECONDARY SCHOOLS IN LINE WITH THE REQUIREMENTS OF INTERNATIONAL ASSESSMENT PROGRAMMES (PISA, TIMSS)

Kambarova Nodiraxon Abdumuxtorovna
Director of General Secondary School No. 283,
Uchtepa District Tashkent City, Uzbekistan
kambarovanodiraxon@gmail.com

Abstract. International large-scale assessments have become one of the main reference points for judging the quality of school education, yet their influence on everyday school practice often remains limited to test preparation. This article argues that the real value of PISA and TIMSS for schools lies not in their league tables but in the assessment design principles behind them: precise construct definition, explicit cognitive demand, described proficiency scales, anchoring for comparability over time, matrix sampling, contextual questionnaires and cautious reporting of uncertainty. Drawing on assessment validity theory, research on data use in schools and selected international experience, the article translates these principles into requirements for school-level monitoring, identifies typical gaps between them and current practice, and proposes a three-tier monitoring architecture supported by school data teams. Safeguards against score inflation and a phased implementation roadmap are also outlined.

Keywords: education quality monitoring, international large-scale assessment, PISA, TIMSS, assessment validity, proficiency scale, data literacy, data teams, school improvement.

Over the past two decades, school systems around the world have moved from judging quality mainly through inspection of inputs and compliance with regulations

towards judging it through evidence of student learning. International large-scale assessments have accelerated this shift. When a country joins PISA or TIMSS, it gains a mirror in which its curriculum, teaching and assessment practices can be compared with those of systems that have achieved consistently high results. For Uzbekistan, which has taken part in several recent cycles of these studies, this mirror has prompted a serious national conversation about what students actually learn and how schools know whether they are succeeding.

However, a gap persists between national participation in international studies and the way schools monitor their own work. International surveys are sample-based, take place every three or four years, and report results for the system as a whole. A school principal cannot see his or her own students in a PISA report. Meanwhile, school-level monitoring frequently continues to rely on instruments designed for a different purpose: checking whether the content of the syllabus has been covered and recalled. The result is a kind of double accountability. Schools are expected to contribute to better international outcomes, but they measure themselves with tools that are not aligned with what those outcomes reflect.

A second risk is that schools respond to international assessment pressure in the least productive way, by drilling students on items that resemble PISA or TIMSS tasks. Research on high-stakes testing has repeatedly warned that when a quantitative indicator becomes a target, it tends to lose its value as an indicator. Narrow drilling may raise scores on familiar formats while leaving the underlying competences unchanged.

The central question of this article is therefore not how to prepare students for international tests, but how schools can borrow the methodological strengths of these tests to build monitoring systems that produce trustworthy, actionable information about learning. The aim is to formulate a set of design requirements for school-level monitoring derived from PISA and TIMSS, to diagnose the main gaps in current

practice, and to propose an organisational and methodological model for closing them.

The first is assessment validity theory. Following Messick, validity is understood not as a property of a test but as the degree to which evidence and theory support the interpretations and uses of test scores. From this perspective, the key question for any school monitoring instrument is whether the conclusions a principal draws from it, and the decisions taken on that basis, are justified. A test made up mostly of recall questions may be perfectly reliable, yet it cannot support the conclusion that students are able to apply mathematics to unfamiliar problems. Much of what makes PISA and TIMSS credible is the effort their designers put into building a validity argument: published frameworks, item review, field trials, scaling and documentation of limitations.

The second is the theory of constructive alignment, associated with Biggs, which holds that intended learning outcomes, teaching activities and assessment tasks must be consistent with one another. If the national curriculum declares the development of reasoning and problem solving as aims, but classroom and school assessment reward reproduction, then teaching will drift towards what is assessed. Monitoring is therefore not a neutral observer of teaching; it is one of the strongest signals of what the school actually values.

The third body of theory concerns data use in schools. Studies by Mandinach and Gummer on data literacy for educators, and by Schildkamp and Poortman on data teams, show that data improve teaching only when three conditions are met: the data are relevant to questions teachers genuinely care about, staff have the skills to interpret them, and there is a structured, collaborative routine through which interpretation leads to action and the effects of that action are checked. Without these conditions, even high-quality data end up in reports that nobody reads. The OECD review of evaluation and assessment frameworks reached a similar conclusion at the

system level: the various components of evaluation produce synergy only when they are coherent and when their results feed back into improvement.

Taken together, these theories suggest that improving monitoring is not simply a matter of introducing new tests. It requires aligning what is measured with intended outcomes, ensuring that interpretations are valid, and embedding the use of data in the professional routines of the school.

The study followed a design-oriented qualitative approach. First, the publicly available assessment frameworks and technical documentation of PISA and TIMSS were analysed to identify the design principles that underpin the credibility of their results. Second, these principles were reformulated as requirements that could, in principle, be met by a school or a group of schools with modest resources. Third, the requirements were compared with typical elements of school internal monitoring as reflected in regulatory documents, school monitoring plans and published professional literature, in order to identify recurring gaps. Fourth, a monitoring model was designed to address these gaps, drawing on international research and on the experience of systems that have improved their results. The model is presented as a theoretically grounded proposal; its empirical validation through pilot implementation is identified as the next stage of research.

1. Design principles of PISA and TIMSS relevant to school monitoring

Analysis of the assessment frameworks allows seven design principles to be distinguished. Each of them can be translated into a concrete requirement for school-level monitoring.

Principle one: explicit construct definition. Before a single item is written, both studies publish a framework that defines precisely what is being measured. PISA defines mathematical literacy as the capacity to reason mathematically and to formulate, employ and interpret mathematics to solve problems in a variety of real-

world contexts. TIMSS defines its content domains with reference to curricula across participating countries. The school-level requirement that follows is simple but frequently ignored: every monitoring instrument should be preceded by a short written specification stating which competence it measures, in which content areas, at what cognitive level, and for what decision the results will be used.

Principle two: explicit cognitive demand. TIMSS classifies every item according to cognitive domain, distinguishing knowing, applying and reasoning, and sets target proportions for each. PISA distinguishes processes such as formulating situations mathematically, employing procedures and interpreting results. The requirement for schools is that monitoring instruments should have a blueprint, a simple grid crossing content areas with cognitive levels, and that a deliberate share of items, for example at least one third, should target application and reasoning rather than recall.

Principle three: authentic stimulus and varied item formats. Items in both studies are typically built around a stimulus such as a text, a graph, a table, a diagram or a description of a situation, and combine selected-response items with constructed-response items requiring students to explain or justify their answers. For schools this means moving away from instruments made up exclusively of short, decontextualised questions, and introducing item clusters in which several questions of rising difficulty refer to a single realistic stimulus.

Principle four: described proficiency scales. Results in PISA and TIMSS are not merely numbers; they are linked to descriptions of what students at each level can typically do. This is what makes the results interpretable for teachers and policymakers. The school equivalent is a set of level descriptors, written in plain language, against which students' performance can be reported. Instead of telling a teacher that a class scored 62 percent, the monitoring report should say what proportion of students can only perform routine procedures, what proportion can

apply knowledge in familiar contexts, and what proportion can reason in unfamiliar situations.

Principle five: anchoring for comparability over time. Both studies maintain trend measurement by keeping a set of secure link items across cycles. Without such anchoring, it is impossible to tell whether a change in results reflects a change in learning or simply a change in test difficulty. Most school monitoring lacks this feature, so year-on-year comparisons are largely meaningless. A practical requirement is to keep a confidential core of anchor items that are reused in successive rounds of school monitoring and never used for teaching or published.

Principle six: matrix sampling and contextual questionnaires. International studies do not require every student to answer every item. Using rotated booklets, they cover a broad domain while limiting the testing time for each student. They also collect background information from students, teachers and principals, which allows results to be interpreted in light of factors such as attitudes, learning resources and classroom climate. For schools, rotated test forms make it possible to assess a broader range of competences without increasing the burden on students, while short questionnaires make it possible to ask not only how well students perform but also under what conditions.

Principle seven: careful reporting of uncertainty. International reports routinely present standard errors and avoid overinterpreting small differences. School monitoring, by contrast, often ranks classes and teachers on the basis of differences that are well within the range of random variation, especially in small groups. The requirement here is for reporting rules: no ranking of classes or teachers on single test results, attention to trends across several rounds, and explicit caution when groups are small.

2. Gaps between current school monitoring and these principles

Comparison of these requirements with typical school monitoring practice reveals a consistent pattern of gaps.

The first gap is the absence of specifications and blueprints. Tests are often compiled from available collections of exercises without a prior decision about what exactly they should measure. As a result, the cognitive profile of school instruments is usually skewed towards recall and routine procedures, and the instruments cannot support claims about higher-order competences.

The second gap concerns the format of reporting. Results are usually summarised as percentages of successful completion and grade distributions. Such indicators tell a principal that something is wrong but not what is wrong or what to do about it. Descriptive, level-based reporting is rarely used.

The third gap is the lack of comparability over time. Because instruments change from round to round and no anchor items are kept, schools cannot reliably judge whether their improvement measures are working.

The fourth gap is the separation of achievement data from contextual information. Schools collect data on attendance, behaviour and family circumstances, but these data are rarely analysed together with achievement data in order to understand which groups of students are falling behind and why. This limits the capacity of monitoring to support equity, which is one of the dimensions of quality that international studies emphasise most strongly.

The fifth gap is organisational. In many schools, monitoring is perceived primarily as an administrative obligation carried out for the next level of the hierarchy. There is no regular, protected time in which teachers jointly examine evidence of student learning, formulate hypotheses about the causes of difficulties, agree on changes in teaching and check their effects. Under these conditions, the use of data remains episodic.

The sixth gap concerns stakes and incentives. When monitoring results are used mainly to evaluate or sanction teachers, the predictable response is to narrow teaching to the test, to manage which students sit the tests, or to inflate marking. The literature on accountability, including Koretz's analysis of score inflation, provides ample evidence that high stakes attached to weak instruments damage both the instruments and the teaching.

3. A three-tier model of school-level quality monitoring

To address these gaps, a three-tier monitoring architecture is proposed. Its logic is that different decisions require different kinds of evidence, collected at different frequencies and with different degrees of standardisation.

The first tier is classroom-level formative monitoring. Its purpose is to inform day-to-day teaching. It relies on short, frequent tasks embedded in lessons, such as exit tickets, short reasoning problems, peer explanation and diagnostic questions designed to reveal typical misconceptions. Following Wiliam's work on embedded formative assessment, the emphasis is on eliciting evidence of student thinking and adjusting teaching in response. The connection with international assessments lies in task design: teachers deliberately include tasks at the applying and reasoning levels and tasks built around realistic stimuli, so that students regularly encounter the kind of thinking that PISA and TIMSS value. Results at this tier are not aggregated or reported upwards; they belong to the teacher.

The second tier is school-level benchmarking. Its purpose is to give the school leadership and subject departments a reliable picture of how students are progressing across key competences. Benchmarking assessments are conducted two or three times a year, mainly in the grades that correspond to international studies, namely grades 4 and 8 for TIMSS-type competences and the age group of fifteen-year-olds for PISA-type literacy. The instruments are built from a written specification and blueprint, use clusters of items around realistic stimuli, combine selected-response

and constructed-response formats, and contain a confidential core of anchor items. Constructed responses are scored with common rubrics distinguishing full, partial and no credit, and a proportion of scripts is double-marked to check consistency between scorers. Results are reported against level descriptors rather than as raw percentages. A short student questionnaire covering reading habits, attitudes towards mathematics and science, perceived support from teachers and access to learning resources is administered at the same time.

The third tier is district-level aggregation and support. Its purpose is not to rank schools but to identify patterns that individual schools cannot see on their own, to allocate methodological support and to share effective practices. At this tier, a regional methodological centre can maintain a shared item bank, train scorers, analyse data from groups of schools and, where data allow, take into account students' starting points so that schools serving disadvantaged communities are not unfairly judged. Item response theory scaling, which underpins international assessments, can be introduced gradually at this level, where the necessary expertise and sample sizes are more likely to be available.

The value of this procedure lies in its discipline. It prevents the common jump from a disappointing result straight to a generic remedy such as additional lessons, and directs attention to the specific mechanisms behind the difficulty. It also distributes responsibility for quality across the teaching staff rather than concentrating it in the hands of the administration.

The proposed model rests on the idea that international assessments are most useful to schools as a source of methodology, not as a target. This position has several implications.

First, it shifts the emphasis from instruments to capacity. A school that receives a ready-made bank of PISA-style items but lacks staff who can interpret results and translate them into changes in teaching will gain little. Investment in assessment and

data literacy of principals and teachers is therefore a precondition rather than an addition.

Second, it requires a change in the culture of accountability. As research on school inspections in Europe by Ehren and colleagues suggests, external evaluation influences improvement mainly when schools accept its standards and develop their own capacity for self-evaluation. If monitoring is perceived as a threat, teachers will protect themselves; if it is perceived as a tool for professional problem solving, they are more likely to engage with it honestly.

Third, it calls for realism about resources. Not every school can maintain an item bank, train scorers or apply item response theory. The three-tier design acknowledges this by assigning more technical tasks to the district or regional level while keeping interpretation and action at the school level, close to the classroom.

Fourth, the model must respect national curricula and cultural contexts. International frameworks offer a valuable language for describing competences and cognitive demand, but tasks used in school monitoring should draw on situations familiar to local students and remain consistent with national educational standards. The aim is to match the level of cognitive demand set by international benchmarks, not to copy foreign content.

Finally, it should be acknowledged that the model is a theoretical proposal. Its feasibility, the workload it creates for teachers and its effect on student outcomes need to be tested through pilot implementation in a range of schools, including small rural schools, with careful attention to unintended consequences.

In the preparatory phase, lasting about one academic term, the school or group of schools establishes data teams, trains their members in assessment specification, item writing, rubric scoring and basic data analysis, and agrees on level descriptors

for the target competences. The district methodological centre prepares the first version of the shared item bank and the anchor set.

In the pilot phase, lasting one academic year, two rounds of benchmarking are conducted in the target grades together with the student questionnaire. Data teams carry out at least two full improvement cycles, documenting their hypotheses, actions and results. Particular attention is paid to the workload of teachers and to the clarity of reports.

In the consolidation phase, the instruments and procedures are revised on the basis of pilot experience, the anchor set is used to establish a baseline for trend measurement, and monitoring is integrated into the school development plan and the professional development plans of teachers.

Success at each phase should be judged not only by changes in benchmarking results but also by process indicators: the proportion of teachers participating in data team cycles, the share of reasoning-level tasks in classroom and school assessment, the consistency of scoring between markers and the extent to which improvement measures are based on evidence rather than routine.

International large-scale assessments such as PISA and TIMSS offer schools far more than comparative rankings. Behind their results lies a body of assessment design principles that make those results credible: explicit construct definition, a deliberate balance of cognitive demand, authentic stimulus material, described proficiency levels, anchoring for comparability, matrix sampling with contextual questionnaires and careful treatment of uncertainty. Translating these principles into requirements for school-level monitoring makes it possible to close the gap between national participation in international studies and the everyday practice of schools.

The analysis identified typical weaknesses of current school monitoring, including the absence of specifications, reporting that is not interpretable in terms of

competences, lack of comparability over time, separation of achievement from context, weak routines for collective data use and the distorting effect of high stakes. The proposed three-tier model, combining classroom formative monitoring, school benchmarking and district-level support, and driven by school data teams working in a structured improvement cycle, addresses these weaknesses while guarding against teaching to the test.

The main recommendation is to treat the improvement of monitoring as a matter of professional capacity and organisational culture as much as of instruments. Further research should focus on piloting the model, measuring its effects on both student outcomes and teaching practice, and adapting it to schools of different sizes and contexts.

REFERENCES

1. Biggs J. Enhancing teaching through constructive alignment // Higher Education. 1996. Vol. 32, No. 3. P. 347–364.

2. Campbell D.T. Assessing the impact of planned social change // Evaluation and Program Planning. 1979. Vol. 2, No. 1. P. 67–90.

3. Clarke M. What Matters Most for Student Assessment Systems: A Framework Paper. Washington, DC: World Bank, 2012.

4. Ehren M.C.M., Altrichter H., McNamara G., O'Hara J. Impact of school inspections on improvement of schools: describing assumptions on causal mechanisms in six European countries // Educational Assessment, Evaluation and Accountability. 2013. Vol. 25, No. 1. P. 3–43.

5. Greaney V., Kellaghan T. Assessing National Achievement Levels in Education. Washington, DC: World Bank, 2008.

6. Koretz D. The Testing Charade: Pretending to Make Schools Better. Chicago: University of Chicago Press, 2017.