ASSESSMENT BLUEPRINT

10TH MEETING OF THE GLOBAL ALLIANCE TO MONITOR LEARNING (GAML)

ANDRÉS SANDOVAL-HERNÁNDEZ
PEDRO PINEDA-RODRÍGUEZ
ARTEMIO CORTEZ-OCHOA

UNIVERSITY OF BATH
University of BRISTOL
UNESCO INSTITUTE FOR STATISTICS
GLOBAL ALLIANCE TO MONITOR LEARNING
SUSTAINABLE DEVELOPMENT GOALS

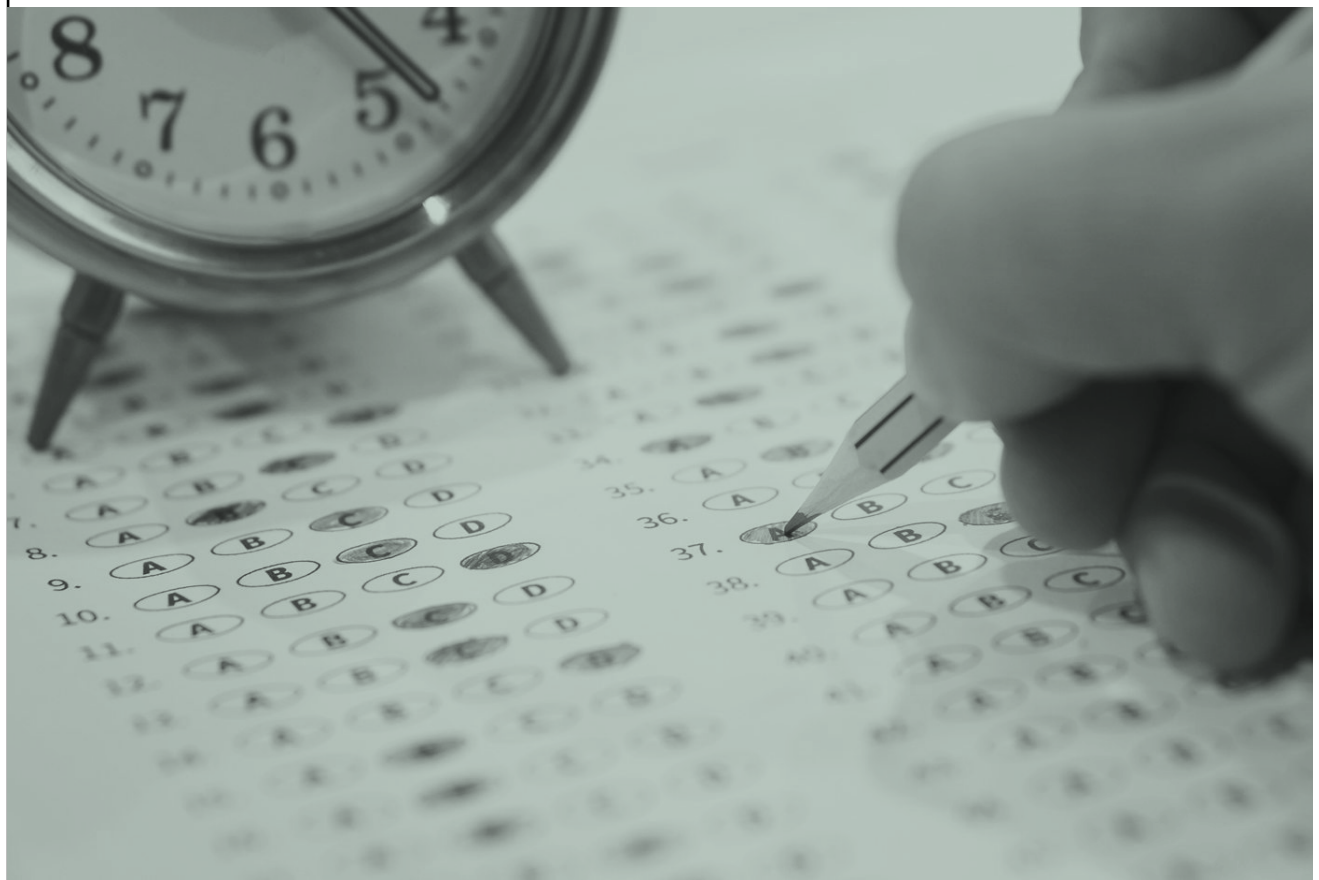
CONTENT

- Background
- Harmonization initiatives
- An Assessment Blueprint
- Discussion

BACKGROUND

STARTING POINTS

STARTING POINTS



- In the context of SDG4, gaining an insight into the areas where progress has been achieved and those where further improvements are needed is critical...

- BUT challenging because the discussion on how to monitor and measure learning outcomes and skills is ongoing in nature.

- Current issues include:

  - a narrow vs a broad scope of learning measurement

  - global vs national goals and targets

  - measurement of learning for all children vs those in schools

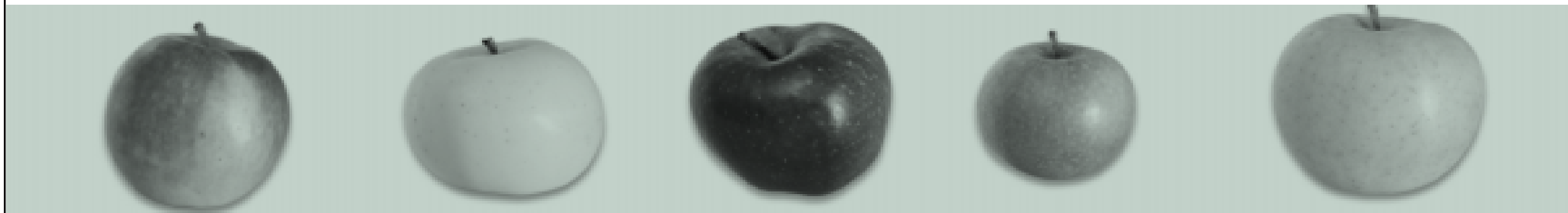
  - top-down vs bottom-up implementation

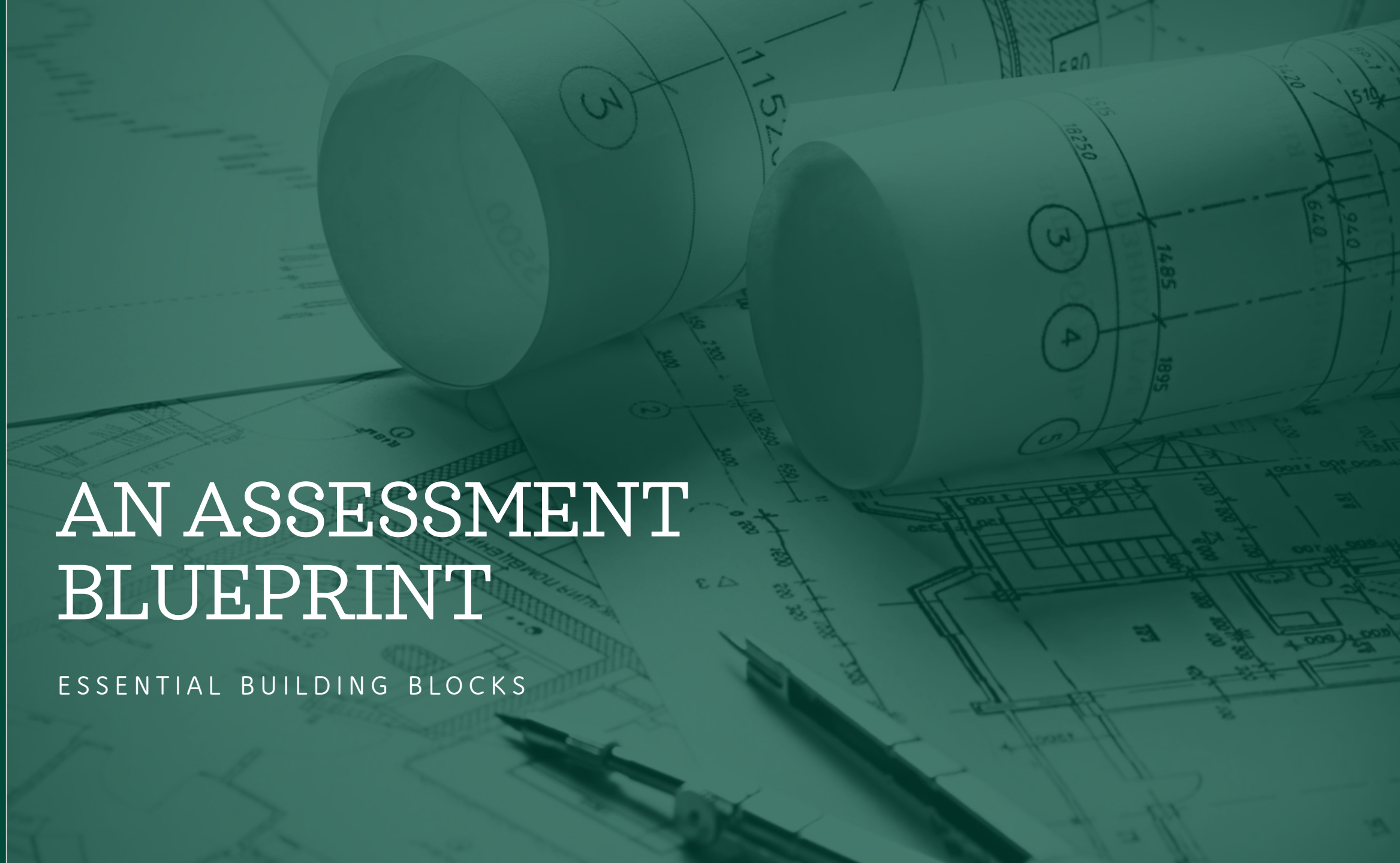
HARMONIZATION INITIATIVES

MAKING DATA COMPARABLE

MAKING THE DATA COMPARABLE

- Considering the limitations, different solutions have been suggested to obtain data that can be used to measure and monitor, for example, SDG 4.1.1.
- **Rosetta Stone** uses psychometric methods to link regional assessments (e.g., ERCE, PASEC) to international assessments (e.g., TIMSS, PIRLS). Potentially applicable to national assessments too.
- **Policy Linking** is a non-statistical method that uses judgment to align and match items from the national assessment with the Global Proficiency Framework (GPF).
- **Assessments for Minimum Proficiency Levels (AMPLs)** are tools targeted at measuring the attainment of a single proficiency level for reading and mathematics at a given level of the education cycle.
- **Pairwise Comparison Method:** is a systematic approach to ranking or selecting from a group of alternatives by comparing them against each other in pairs.





AN ASSESSMENT BLUEPRINT

ESSENTIAL BUILDING BLOCKS

WHAT FOR?

- The ideal strategy would be the creation and systematic maintenance of a harmonized international dataset, that provides longitudinal indicators of educational achievement at the country level.

- This harmonized dataset would include indicators from national, global and regional large-scale school assessments that **meet the minimum quality requirements**.

- While initiatives like the Rosetta Stone, Pairwise Comparison, or Policy Linking have worked to harmonize different educational assessments, a standardized blueprint is still needed to systematically evaluate which assessments are suitable to include in these harmonization efforts

ASSESSMENT BLUEPRINT

OVERVIEW

Category	Specific Property	Description	Example of Property Compliance*	Hypothetical Example of Property failing to Comply
Psychometric Properties	Alignment	Assessments should be closely aligned with SDG 4 targets and should articulate with their social expectations and be in harmony with international assessments (Molina et al., 2021).	ERCE 2019 assessed students' academic abilities in third grade in language, and mathematics, providing a comprehensive measure that is highly aligned with the goals of SDG 4.1.1a.	A national "Mathematics Literacy" scale focuses only on advanced calculus, ignoring basic numeracy skills. This does not comply with SDG 4 targets because it neglects basic literacy.
	Validity	The assessment must accurately measure what SDG 4 intends to measure, demonstrating construct validity, content validity, and criterion validity. Construct validity ensures that the test accurately represents the features it intends to describe, explain, or theorize, as confirmed by its scope and psychometric attributes. Content validity ensures that the test covers all relevant aspects of the subject under investigation, aligned with SDG 4 targets. Criterion validity confirms that the test results are effective predictors of a future outcome or are in agreement with a present outcome, thereby aligning with SDG 4 metrics (Cohen et al., 2018)	ERCE 2019 attributes are thoroughly examined to ensure accurate representation of the educational constructs it aims to describe and evaluate (see ERCE 2019 Assessment Framework). ERCE 2019 is designed by UNESCO for UNESCO providing a good example of content validity by ensuring that the content of the test is relevant and addresses the key areas outlined in the SDG 4.1.1a. ERCE 2019 also presents criterion validity as research has shown that its results predict future educational outcomes or align with present outcomes related to SDG 4 metrics (e.g., Carrasco, Rutkowski & Rutkowski, 2023)	A "Reading Literacy" assessment for grade 2 only measures word decoding skills through having students read words aloud from a list. It does not have them read grade-level texts aloud or answer comprehension questions, which better represents overall reading proficiency.
	Reliability	To effectively contribute to reliable Sustainable Development Goal metrics, an assessment must consistently yield stable and unvarying results over multiple time points, as emphasized by psychometric research (Naglieri, 2013). This property enables reliable repetition over time to track progress in meeting SDG 4. Ensuring such reliability, it is recommended that assessments achieve a test-retest reliability coefficient, typically using Pearson's r, of at least 0.9 (Price, 2017). This threshold indicates that the assessment maintains a high degree of stability in its measurements over time.	TIMSS utilizes a well-defined methodology, including rigorous sampling and instrument piloting, to ensure that its assessment of math and science skills is reliable from one cycle to the next.	A Mathematics Literacy test changes its format and question types annually, making it impossible to compare results from year to year. It fails to comply with repetition viability because it cannot reliably track progress over time.
	Difficulty Level	The assessment should be precisely calibrated to measure the specific educational level and context targeted. It is crucial to make accommodations that do not compromise the test's validity or alter the difficulty level of the items, thereby ensuring that the constructs being measured remain consistent (Willis et al., 2013).	PIRLS targets fourth-grade students and is careful to use language and question formats that are age-appropriate, ensuring that the assessment is tailored to its intended audience (see PIRLS 2021 Assessment Frameworks).	A "School Infrastructure" survey uses overly technical language, difficult for local school administrators to complete. This does not comply with difficulty level because it is not accessible to its intended audience.
	Item discrimination	The assessment should effectively differentiate between different levels of achievement (Cizek, 2001). In this context, it is important to consider the trade-off between item discrimination across a range of ability levels and the accuracy of assessment around the critical proficiency levels of interest.	PASEC 2019 includes a wide range of questions that cover varying levels of difficulty, allowing the test to distinguish clearly between high, medium, and low performers (see PASEC International Reports).	An ICT Skills assessment has too many easy questions, making it hard to distinguish between levels of competence. This does not comply with discrimination because it fails to differentiate between skill levels.
	Item Design Clarity	The design of each assessment item must be clear, unambiguous, and directly aligned with the intended measurement goals. Before being used in large-scale applications, items should be rigorously vetted through cognitive testing, pilot testing, re-testing, and refining. This comprehensive process is crucial for ensuring that each item is understandable and effectively measures the intended construct. The methodologies of Item Response Theory (IRT) or Classical Test Theory (CTT) can be employed to gauge the reliability and validity of these assessments (UNESCO, 2019).	The ICCS uses unambiguous language and provides clear instructions to ensure that students from different cultural backgrounds can understand what is being asked (see ICCS Technical Report).	A test on historical and civic knowledge includes questions on peace education, but uses the term 'peace' ambiguously. This leads to different interpretations by students of different religions, such as peace as a spiritual state evoked by shalom and salaam in Hebrew and Arabic, or as the absence of violence according to the Western tradition, derived from the Latin word pax (peace, paz, paix, pau, pace), which refers to the absence of violence (Pineda & Celis, 2021; Pineda et al., 2019). This does not comply with item design clarity because the questions are not straightforward, causing confusion among test-takers.
Data quality	Representativeness	The sample for the assessment must be reflective of the diversity of educational status, ensuring not only that participants represent an available population but also the target population to which findings are intended to be generalized (Cohen et al., 2018). The chosen approach must be well-defended, taking into account factors such as alignment with the language of the Sustainable Development Goals, the economy of field costs, and agreeableness to the national government. A typical difficulty that should be considered at the school level is scheduling; assessments must be planned at times convenient for both participants and administrators and should avoid vacation periods.	SEA-PLM includes both public and private schools, from both urban and rural settings in multiple countries, and ensures a representative sample of the target population by implementing a rigorous sampling methodology (see SEA-PLM Technical Standards).	An Enrolment Rates study only samples urban schools, ignoring rural areas. It fails to comply with representativeness as it does not cover the full spectrum of educational diversity.
	Comparability	Procedures for administering the assessment should be standardized to enable comparison across regions (Rutkowski & Rutkowski, 2017). Furthermore, it is crucial that these procedures, along with their standardization processes, are thoroughly documented, maintained on file, and made publicly accessible to ensure transparency and reproducibility in the assessment's application and analysis.	TIMSS provides strict guidelines to all participating countries on how to administer their Mathematics and Science tests, ensuring comparability. TIMSS also implements strict technical procedures to produce scale scores that are comparable between countries and across time (see TIMSS 2019 Technical Report).	State A and State B administer their own versions of the Abitur exams with differing academic rigor and testing criteria (see Kühn, 2012). Due to these variations, a high score in State A may not signify the same level of achievement as a similar score in State B. This lack of standardization poses challenges for comparability, making it difficult to use the exam results for measuring SDG 4.
	Transparency	The process of creating and conducting assessments must incorporate well-documented design, sampling techniques, and analysis procedures, and these details should be clearly and publicly documented at the time of the assessment's deployment.. This transparency is essential for meeting the increasing demands for reliable measures and high-quality documentation (Stancel-Piątak & Schwippert, 2022).	ICCS provides comprehensive methodological reports available publicly, detailing the data collection, sampling methods, and analysis techniques (see ICCS Technical Report).	A School Infrastructure assessment lacks any available documentation on how the survey was conducted or analysed. It fails to comply with documentation, hindering transparency.
	Test security	To prevent potential issues such as teaching to the test or excessive test preparation, it is essential not to make specific test items public. This approach safeguards the integrity of the assessment process, ensuring that students are evaluated based on their understanding of the broader curriculum or assessment framework rather than focusing solely on memorising or practicing specific test items. Rigorous vetting through cognitive testing, pilot testing, re-testing, and refining should be conducted to maintain the clarity and effectiveness of each assessment item (Gölogl Demir & Kaplan Keles, 2021).	The PIRLS Item Release Policy states that responses to all items used in the assessment are included in the database. After each cycle, however, some of the items are made available for restricted use by the public. The remaining items are kept secure, thus ensuring the possibility of measuring trends over time. The item release policy is described in the Item Release Plan. Access to the restricted use items is subject to approval by the IEA Amsterdam. However, the response data for all items used in the assessment are publicly and freely available in the data files (see information on PIRLS 2016 Database for details).	A" Reading Literacy" assessment makes all items used in its cognitive test public in order to ensure transparency.
	Suitable Technical Infrastructure	The technological framework required for administering assessments must not only be reasonable and achievable but also centrally coordinated to ensure uniformity (Hastedt & Sibberns, 2022)	LLECE assessments use basic multiple-choice questions that can be answered using ordinary computers or even paper-based tests, ensuring broad accessibility. Another example is the Quality Assurance Program implemented by PIRLS (see a description here)	A test in Geographical Knowledge required the use of advanced of complex geographic information system software, which most schools do not have access to. This created an unnecessary technological hurdle, excluding schools that could not afford or implement the required software.
	Stakeholder Involvement	Stakeholder Involvement refers to the active participation of subject matter experts, educators, and community members in the design and implementation of educational assessments or programs. This involvement also extends to the dissemination of results and recommendations to facilitate informed decision-making at various levels of educational policy and practice, all while maintaining the integrity of the evaluations (Ababneh et al., 2016).	TALIS involves teachers, principals, and education researchers in the design and interpretation phases of their survey.	A Completion Rates survey, developed solely by a bureaucratic government department, lacks the insights that teachers, parents, and educational researchers could have provided. This resulted in questions that are not reflective of the educational environment.
Ethics	Feasibility	Feasibility in the context of educational assessment refers to the consideration of both financial and time-related costs for all parties involved in the testing process (Rutkowski et al., 2023).[1]	UWEZO tests are deliberately designed to be administered within a single school day and are low-cost enough to be managed by local volunteers.	To administer a Science Achievement test, schools are required to purchase specialized, expensive equipment and allocate additional staff hours. These excessive requirements led to many schools opting out of the test.
	Accessibility	Measures must be implemented to guarantee equitable access to assessments for all individuals, especially those with disabilities. When disabilities are considered, they should be consistently acknowledged and documented across different places of application (Meinck & Vandenplas, 2022).[2]	In Italy, for national assessments at the primary and secondary levels, the National Evaluation Center provides tests in special formats (e.g., tests recorded in MP3 audio files, tests in large print or Braille format for visually impaired children, tests specifically adapted for deaf students) (see Italy's chapter in PIRLS 2016 Encyclopaedia).	The School Infrastructure survey was designed without taking into account the needs of individuals with disabilities, failing to offer alternative formats like braille or audio descriptions.
	Digital Accessibility	Tests are administered in various formats, including paper-based, computer-based, or a combination of both, depending on the specific assessment cycle and technological developments (Kyriakides et al., 2022)[3]	PIRLS offers its Reading Literacy test online, thereby ensuring it is accessible to a wider audience who can take the test remotely.	A Completion Rates survey is only distributed in print, without an online alternative, limiting its reach and ease of participation. It does not comply with digital accessibility.
	Data Privacy	Measures must be in place to protect the confidentiality and privacy of participants' data (Walford, 2005).	TIMSS anonymizes all participant data and stores it in secure databases, accessible only to authorized researchers (see TIMSS 2019 Data Protection Declaration).	In an ICT Skills assessment, participants found that their personal data, including their names and scores, were published on a government website without their consent, breaching data privacy norms.

ASSESSMENT BLUEPRINT

PSYCHOMETRIC PROPERTIES

- Alignment
- Validity
- Reliability
- Difficulty Level
- Item Discrimination
- Item Design Clarity

DATA QUALITY

- Representativeness
- Reproducibility
- Transparency
- Test Security
- Suitable Technical Infrastructure
- Stakeholder Involvement

ETHICS

- Feasibility
- Physical Accessibility
- Digital Accessibility
- Data Privacy

DISCUSSION

AND NOW WHAT?

HOW CAN WE MOVE FORWARD?

What do we have?

- Different sources of data to measure an monitor SDG 4 indicators
- Several options for harmonization of educational assessments (Rosetta Stone, Pairwise Comparison, Policy Linking, MPL).
- A proposal for an assessment blueprint

What do we need?

- Finalize the assessment blueprint
- Use it to identify the best-quality data sources for each SDG 4 indicator
- Support (where needed) the development of quality assessment data for measuring and monitoring SDG 4 indicators.
- Strengthen and decide on a harmonization strategy to be applied to each SDG 4 indicator (+ background questionnaires).
- Produce high-quality comparable data to measure and monitor SDG 4 indicators.
- Strengthening stakeholder collaboration.

THANK YOU!

QUESTIONS?

@asandovalh