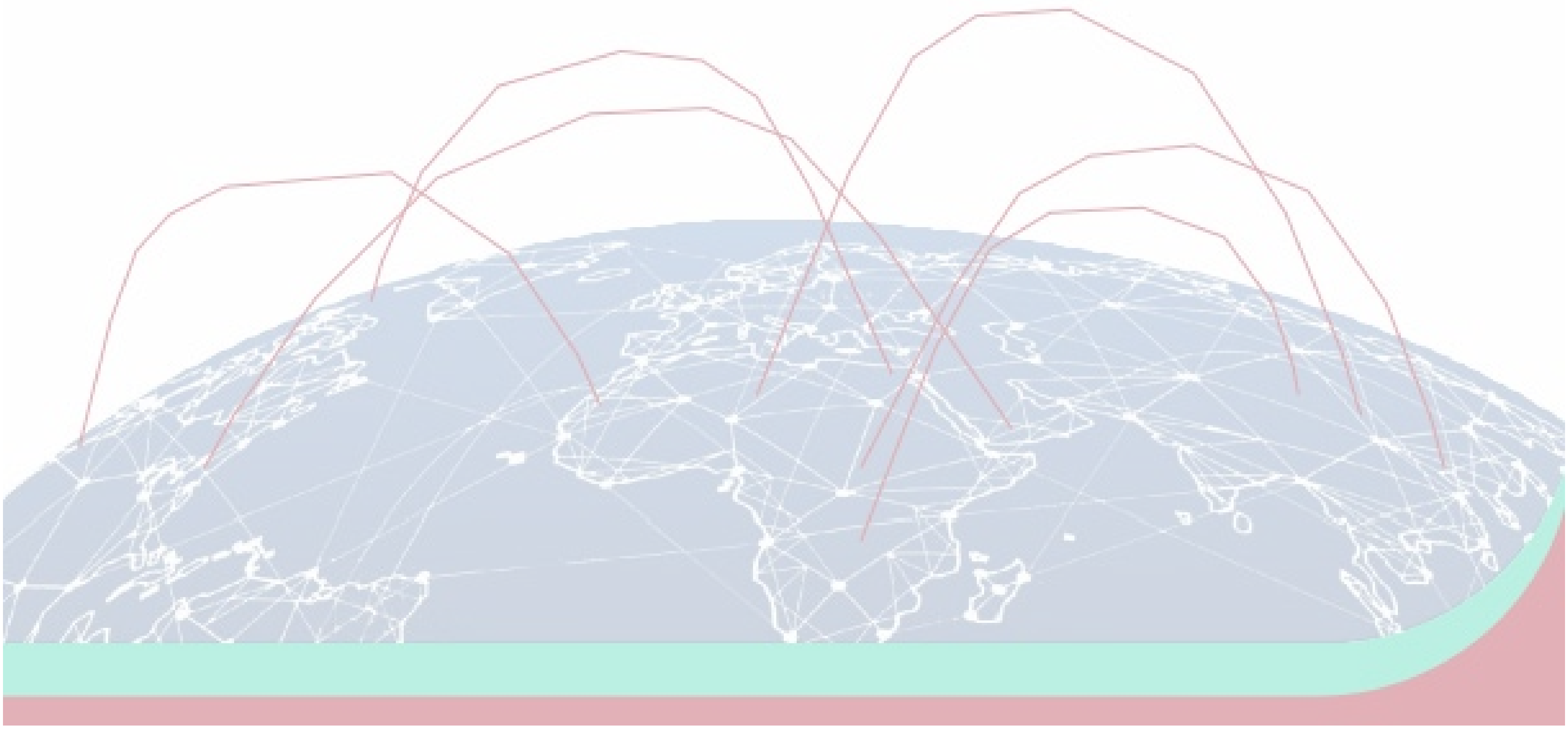
WG/GAML10/ST.REP/1

Guidelines for reviewing and integrating education assessments into the evaluation of SDG4 indicators

Tenth meeting of the Global Alliance to Monitor Learning (GAML)

Paris, 6 - 7 December 2023



GUIDELINES FOR REVIEWING AND INTEGRATING EDUCATION ASSESSMENTS INTO THE EVALUATION OF SDG4 INDICATORS

Draft 3

Pedro Pineda Rodriguez & Andrés Sandoval-Hernández

University of Bath

TABLE OF CONTENTS

INTRODUCTION

The increasing emergence of national and international large-scale assessments (Ramirez et al., 2018) highlights the need for a standardised blueprint to systematically evaluate these assessments' potential for inclusion in evaluation initiatives. This is especially vital in light of their linkages to the measurement of global objectives such as the Sustainable Development Goals (SDGs). Setting out standard reporting requirements for assessments to be used for measuring and monitoring SDG 4.1 to be admissible is a crucial initiative for the global education community. Efforts like the Rosetta Stone, Policy Linking, and AMPLs have already made strides toward harmonising diverse educational assessments. However, it is still important to have clear criteria that assess the quality, comparability, and suitability of these assessments for integration. Establishing such standardised criteria will allow the global community to better leverage large-scale assessments to track progress on SDG 4.1 and other shared education objectives.

To address this challenge, the present document provides a comprehensive blueprint to evaluate educational assessments in relation to their alignment with SDG targets 4.1, 4.4, and 4.7. This blueprint outlines the critical factors that each assessment must meet to be deemed suitable for integration. These criteria include alignment to learning standards and frameworks, psychometric properties, representativeness, and transparency of processes, among others.

The blueprint is compiled by categorising and detailing the essential properties that educational assessments should possess to align with SDG objectives. Each category is described with a specific property, an explanation of its importance, and an illustrative example showing how this property links to educational indicators mapped to the pertinent measurements or sub-scales. Table 1 sets the criteria for evaluating assessments. By utilising this blueprint, UNESCO can rigorously evaluate educational assessments, determining their appropriateness and technical capacity for integration. This will pave the way for stronger global standards and participation, strengthening the efforts of networks like GAML and ensuring that the international educational assessment community is aligned in its mission.

THE BLUEPRINT

Table 1 Expected Properties for Assessments d Indicators for Evaluating Educational Assessments in Alignment with SDG 4 Objectives

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	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.

Category	Specific Property	Description	Example of Property Compliance*	Hypothetical Example of Property failing to Comply
			metrics (e.g., Carrasco, Rutkowski & Rutkowski, 2023)	
	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	PASEC 2019 includes a wide range of questions that cover varying levels of difficulty, allowing the test to distinguish	An ICT Skills assessment has too many easy questions, making it hard to distinguish between levels of competence. This does

Category	Specific Property	Description	Example of Property Compliance*	Hypothetical Example of Property failing to Comply
		trade-off between item discrimination across a range of ability levels and the accuracy of assessment around the critical proficiency levels of interest.	clearly between high, medium, and low performers (see PASEC International Reports).	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). When utilizing CTT, a common measure of internal consistency is Cronbach's Alpha. The accepted norms for this metric vary: an Alpha above 0.90 is indicative of very high reliability, a score from 0.80 to 0.90 suggests high reliability, and a range from 0.70 to 0.79 is typically deemed acceptable for most research purposes. Alpha values	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.

Category	Specific Property	Description	Example of Property Compliance*	Hypothetical Example of Property failing to Comply
		between 0.60 and 0.69 are considered marginally reliable, whereas scores below 0.60 reflect unacceptably low reliability (Cohen et al., 2018). In contrast, when using IRT or other test development strategies, alternative metrics are applied to ensure consistent measurement of the theoretical construct, moving beyond the focus on internal consistency inherent in Cronbach's Alpha. These standards help ensure that assessment items are not only well-designed but also consistently measure what they are intended to measure.		
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;	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.

Category	Specific Property	Description	Example of Property Compliance*	Hypothetical Example of Property failing to Comply
		assessments must be planned at times convenient for both participants and administrators and should avoid vacation periods.		
	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	The PIRLS Item Release Policy states that responses to all items used in the assessment	A" Reading Literacy" assessment makes all items used in its cognitive test public in order to ensure transparency.

Category	Specific Property	Description	Example of Property Compliance*	Hypothetical Example of Property failing to Comply
		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 (Göloglu Demir & Kaplan Keles, 2021).	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).	
	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.

Category	Specific Property	Description	Example of Property Compliance*	Hypothetical Example of Property failing to Comply
	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	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	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.

[1] While time limits on standardized tests are often set for logistical reasons like cost efficiency and ease of administration, the search for feasibility should not lead to think ghat the speed at which a task is completed is generally not the primary construct being measured, especially in K-12 settings. As a result, some U.S. states have eliminated time restrictions on their assessments to focus more accurately on the constructs of interest.

Category	Specific Property	Description	Example of Property Compliance*	Hypothetical Example of Property failing to Comply
		application (Meinck & Vandenplas, 2022).[2]	print or Braille format for visually impaired children, tests specifically adapted for deaf students) (see Italy's chapter in PIRLS 2016 Encyclopaedia).	
	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.

[2] This consistency is crucial in maintaining the comparability of results, particularly in International Large-Scale Assessments (ILSA), where there has been a concerning trend of increasing exclusion rates for students with learning disabilities.

[3] Transition to digital formats, like "eTIMSS" and "digitalPIRLS," is increasing, allowing for more precise instruments and options for entities. Future cycles, such as PISA 2025, are expected to further refine and expand computer-based assessments, including optional tasks in open-ended, digital learning environments.

** PISA - Programme for International Student Assessment
PIRLS - Progress in International Reading Literacy Study
TIMSS - Trends in International Mathematics and Science Study
ERCE - Estudio Regional Comparativo y Explicativo (Regional Comparative and Explanatory Study)
SACMEQ - Southern and Eastern Africa Consortium for Monitoring Educational Quality
PASEC - Programme d'Analyse des Systèmes Éducatifs de la CONFEMEN (Program for the Analysis of Education Systems)
UIS - UNESCO Institute for Statistics
World Bank - World Bank Group
EGMA - Early Grade Mathematics Assessment
ASER - Annual Status of Education Report
UWEZO - Uwezo ('capability' in Swahili) is part of the PAL NETWORK
ICCS - International Civic and Citizenship Education Study
ICILS - International Computer and Information Literacy Study
IELS - International Early Learning and Child Well-being Study
PIAAC - Programme for the International Assessment of Adult Competencies
SEA-PLM - Southeast Asia Primary Learning Metrics
TALIS - Teaching and Learning International Survey
TALIS Starting Strong 2018 - Starting Strong Teaching and Learning International Survey 2018
PIRLS 2021 - Progress in International Reading Literacy Study 2021
TIMSS 2023 - Trends in International Mathematics and Science Study 2023
PISA 2022 - Programme for International Student Assessment 2022
PASEC 2019 - Programme d'analyse des systèmes éducatifs de la CONFEMEN 2019
SACMEQ IV Study - Southern and Eastern Africa Consortium for Monitoring Educational Quality IV Study
SEA-PLM 2019 - Southeast Asia Primary Learning Metrics 2019
TALIS 2018 - Teaching and Learning International Survey 2018

CONCLUSIONS

The present document serves as an initial proposal for the integration of educational assessments, particularly as they relate to the Sustainable Development Goals (SDG 4s), specifically focusing on SDG 4 targets. It lays out a comprehensive blueprint for evaluating these assessments based on various criteria such as alignment to learning standards, psychometric properties, representativeness, and transparency among others. This proposal is aimed at guiding UNESCO and other stakeholders in systematically evaluating educational assessments for their suitability in international harmonization efforts.

The preliminary conclusion suggests that measuring the targets of SDG, specifically 4.1.1a, 4.1.1b, and 4.7, is feasible due to the well-established knowledge in educational assessment and the increasing availability of assessments that meet these criteria. The properties that educational assessments should possess have been discussed for many decades, and there is consensus in the educational community about what these should be. In addition, the comprehensive list of international assessments shows that appropriate tools are currently in place that are closely aligned with the targets and indicators of SDGs 4.1, 4.3 and 4.7. However, it should be noted that for the target '4.7.2 Percentage of schools providing life skills-based HIV and sexuality education,' national surveys still require implementation for accurate measurement. Both developments open up opportunities for the educational community to apply these universally accepted properties in the measurement of the SDGs. The tables included in the document are designed to serve as tools for UNESCO to rigorously scrutinize assessments, ensuring their alignment with global objectives and enhancing international cooperation in the educational domain. UNESCO and relevant stakeholders can now refine this tool and suggest further improvements, so that it can serve the purpose of supporting the selection of national assessments for measuring the SDGs.

References

Ababneh, E., Al-Tweissi, A., & Abulibdeh, K. (2016). TIMSS and PISA impact – the case of Jordan. *Research Papers in Education*, *31*(5), 542-555. https://doi.org/10.1080/02671522.2016.1225350

Cizek, G. J. (2001). *Setting performance standards: concepts, methods, and perspectives*. Mahwah.

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in Education*. Routledge.

Göloglu Demir, C., & Kaplan Keles, Ö. (2021). The Impact of High-Stakes Testing on the Teaching and Learning Processes of Mathematics. Journal of Pedagogical Research, 5(2), 119–137.

Hastedt, D., & Sibberns, H. (2022). Future Directions, Recommendations, and Potential Developments of ILSA. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International Handbook of Comparative Large-Scale Studies in Education: Perspectives, Methods and Findings*. Springer.

Kühn, S. M. (2012). Zentrale Abiturprüfungen im nationalen und internationalen Vergleich mit besonderer Perspektive auf Bremen und Hessen. In K. M. Merki (Ed.), *Zentralabitur: Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (pp. 27-44). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-94023-6_2

Kyriakides, L., Charalambous, C. Y., & Charalambous, E. (2022). Using ILSAs to Promote Quality and Equity in Education: The Contribution of the Dynamic Model of Educational Effectiveness. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International Handbook of Comparative Large-Scale Studies in Education: Perspectives, Methods and Findings*. Springer.

Meinck, S., & Vandenplas, C. (2022). Sampling Design in ILSA 23: Methods and Implications. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International Handbook of Comparative Large-Scale Studies in Education: Perspectives, Methods and Findings*. Springer.

Molina, J., Hai, N. V., Cheng, P.-H., & Chang, C.-Y. (2021). SDG's Quality Education Approach: Comparative Analysis of Natural Sciences Curriculum Guidelines between Taiwan and Colombia. *Sustainability (United States)*, *13*(6). https://doi.org/10.3390/su13063352

Naglieri, J. A. (2013). Psychological assessment by school psychologists: Opportunities and challenges of a changing landscape. In *APA handbook of testing and assessment in psychology, Vol. 3: Testing and assessment in school psychology and education.* (pp. 3-19). American Psychological Association. https://doi.org/10.1037/14049-001

Pineda, P., & Celis, J. (2021). Rejection and mutation of discourses in curriculum reforms: peace education(s) in Colombia and Germany. *Journal of Curriculum Studies*, 1-23. https://doi.org/10.1080/00220272.2021.1904006

Pineda, P., Celis, J., & Rangel, L. (2019). The worldwide spread of peace education: Discursive patterns in publications and international organizations. *Globalisation, Societies and Education*, *17*(5), 638-657. https://doi.org/10.1080/14767724.2019.1665988

Price, L. R. (2017). *Psychometric Methods: Theory into Practice*.

Ramirez, F. O., Schofer, E., & Meyer, J. W. (2018). International Tests, National Assessments, and Educational Development (1970–2012). *Comparative Education Review*, *62*(3), 344-364. https://doi.org/10.1086/698326

Rutkowski, D., Rutkowski, L., Valdivia, D. S., Canbolat, Y., & Underhill, S. (2023). A Census-Level, Multi-Grade Analysis of the Association Between Testing Time, Breaks, and Achievement. *Applied Measurement in Education*, *36*(1), 14-30. https://doi.org/10.1080/08957347.2023.2172019

Rutkowski, L., & Rutkowski, D. (2017). Improving the Comparability and Local Usefulness of International Assessments: A Look Back and A Way Forward. *Scandinavian Journal of Educational Research*, *62*(3), 354-367. https://doi.org/10.1080/00313831.2016.1261044

Stancel-Piątak, A., & Schwippert, K. (2022). Comprehensive Frameworks of School Learning in ILSAs. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International Handbook of Comparative Large-Scale Studies in Education: Perspectives, Methods and Findings*. Springer.

UNESCO. (2019). *Recommendations on Assessment Tools for Monitoring Digital Literacy within UNESCO's Digital Literacy Global Framework*. UNESCO Institute for Statistics.

Walford, G. (2005). Research ethical guidelines and anonymity1. *International Journal of Research & Method in Education*, *28*(1), 83-93. https://doi.org/10.1080/01406720500036786

Willis, J. O., Dumont, R., & Kaufman, A. S. (2013). Assessment of intellectual functioning in children. In *APA handbook of testing and assessment in psychology, Vol. 3: Testing and assessment in school psychology and education.* (pp. 39-70). American Psychological Association. https://doi.org/10.1037/14049-003