unesco
Institute for Statistics

25 YEARS

Education Data and Statistics Commission

WG/GAML/11/4.1

BUYER'S GUIDE TO INTERNATIONAL STUDENT ASSESSMENT

Global Alliance to Monitor Learning

Contents

Why a "Buyer's Guide" for Student Learning Assessments?[1]

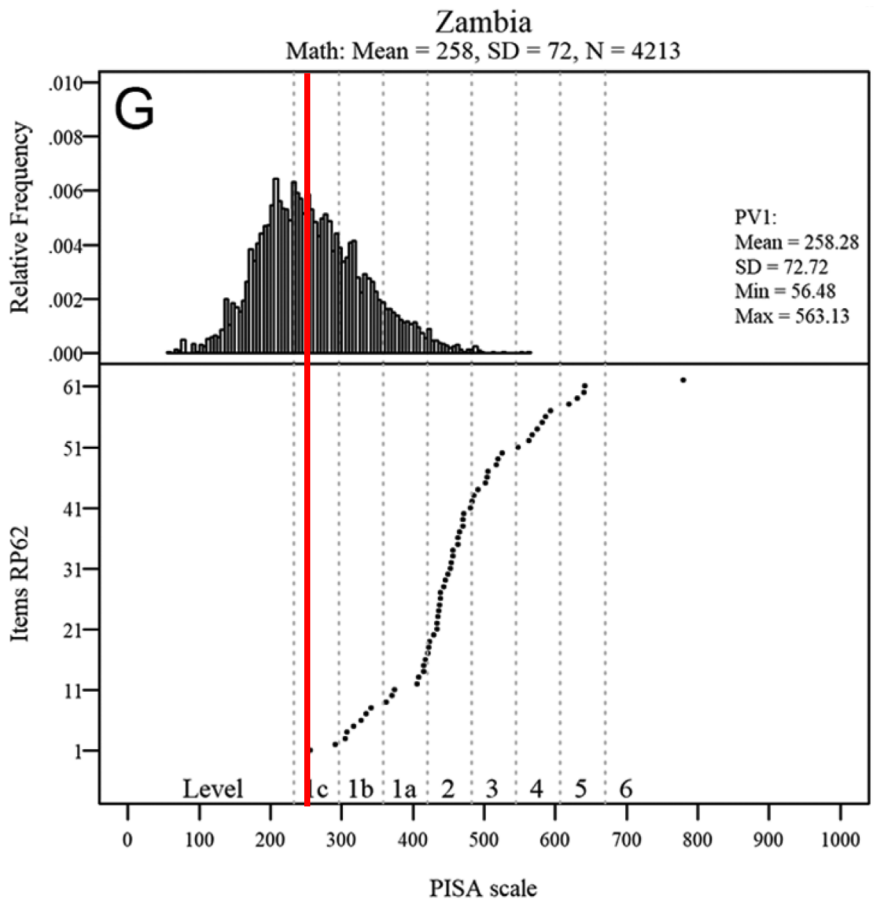
The market for student learning assessments is inefficient and unequal

That there is a market for student assessments may not be immediately apparent, but countries are, in effect, buyers choosing among different assessments (i.e.: products) offered by a number of sellers. Country participation in international student assessments has been aptly characterized as a market: there are buyers (countries), sellers (assessment organizations) and products (the assessments themselves) (Montoya & Crouch 2022a). There are a range of assessments available that differ in skills measured (e.g.: TIMSS versus PISA), age of students being assessed (8th grade TIMSS versus 4th grade TIMSS versus EGRA), the breadth of countries that results can be compared to (e.g.: international versus regional versus national assessments), and so on. Participating in any of these assessments is of course not free: assessment organizations (e.g.: the IEA for TIMSS, the OECD for PISA) charge fees to participate effectively making them sellers and the countries (typically governments but also civil society) buyers. Once purchased, there are additional costs and resource requirements to countries for participating (e.g.: government staff time, school staff time, national workshops, printing, etc.) that vary by the assessment purchased. Even if participation in an international assessment is financed by an external donor, the country is still in effect buying the assessment with donor funds.

However, the market for learning assessments expresses many of the hallmarks of an inefficient and inequitable market. In economic theory, markets are efficient when there is competition, consumers have all the information about the products needed to make an informed choice including fully understanding the costs and benefits, that production technologies are known to all and can be copied, there are no barriers for new sellers to enter, and prices are transparent and uniform. Markets can be inequitable for many of the same reasons, for instance when wealthier or larger consumers can influence prices or have better information about products, but even if efficient, market inequities can arise from differences in purchasing power. From this perspective, the market for learning assessment appears to fall well short of the conditions needed for an efficient and equitable market (Montoya & Crouch 2022b). There is limited competition for a specific type of assessment as assessments differ in various ways. It is difficult for new providers of assessments to enter the market because of cost in developing a new assessment but also because countries want to participate in assessments that already have a large number of participants in order to compare their performance globally (similar to what economists refer to as network externalities). Also, countries may not be willing to switch from one assessment to a new one because switching requires effort (and cost) to learn about new assessment programs, plan and promote it, develop expertise, etc. Switching might also be impractical if the new assessments' results are not comparable with a country's existing assessment's historical results, resulting in a loss of across-time comparability. There are also a lack of transparency and discrimination on prices with negotiation on fees as well as negotiation between third parties (e.g.: development partners) and providers.

[1] Authored by Kevin McDonald under the guidance of Silvia Montoya and Luis Crouch.

Figure 1. Distribution of student achievement and test items: Zambia in PISA-D



Source: Rutkowski, Rutkowski and Liaw (2019)

A number of countries participate in international assessments that are too difficult, resulting in data that cannot be used to inform teaching practices and in potential embarrassment for government. For example, an analysis by Rutkowski, Rutkowski and Liaw (2019) found that PISA D's test items did not cover the lower half of the student ability distribution in Zambia (see **Figure 1**). This means that identifying the skills and competencies that the bottom half of the distribution of students need support on is not possible with this data. Van Davier et al. (2024) noted that Cote d'Ivoire's performance in TIMSS in mathematics and science was "not reliably measured because the percentage of students with achievement too low for estimation exceeds 25%." In other words, even the overall score for Cote d'Ivoire cannot be well measured. While both cases clearly highlight the need for improving learning outcomes, this same conclusion could have been made using an assessment that is more closely aligned to the country's ability distribution, resulting in data that would be more useful for improving learning. The UIS has recently developed the Rosetta Stone (UIS 2022) which provides insight into how countries would score in TIMSS based on their performance in ERCE or PASEC (see discussion below).

Assessments are being conducted in countries sometimes repeatedly with little capacity building being institutionalized. In Sierra Leone, for example, between 2014 and 2020, donors have financed a number of nationally representative student assessments for a variety of grade levels but a national assessment agency has yet to be fully established. This is despite significant financing from donors: cost data is partial, but about 4.5 million USD were disbursed by donors for four assessments between 2017 and 2020 (Varly 2022) while Antonis (2024) estimates that as much as 15 million USD has been disbursed by donors between 2014 and 2022 on assessments. A national assessment unit was created officially in 2021 but budget constraints have resulted in the unit being only partially staffed (Varly 2022). The same issues occur with larger assessment programs as well in Africa. For example, many of PASEC countries teams consist of officials from different parts of the education ministry but are not dedicated to implementing assessments. The result is that the expertise and experience does not get institutionalized; when staff members change, the expertise and knowledge gained from experience in implementing assessments are lost.

Box 1. How do countries choose assessments in practice?

International and regional assessments were generally established by countries sharing common educational goals and curricula, though Western donors have played a significant role in the establishment of regional assessments as well. For example, the OECD established PISA to measure relevant skills for countries' economies and the global economy, the IEA established PIRLS and TIMSS reflecting a shared view of countries' curricula. Regional assessments were also established reflecting participating countries' common aspects of curricula; however, regional assessments were often initiated or financed by donors from or dominated by the global north. Regional assessments remain to varying extents reliant on financing and expertise from the global north, although LLECE has become substantially more independent.

Donor-financed investment project M&E is also a major determinant of choice of assessment, reflecting the goals of the project. EGRA and national assessments have been financed by development partners in part or wholly (especially for EGRA) for the evaluation of project impacts, and the use of assessment for project M&E is an important way that countries end up choosing an assessment or investing in an assessment. There is nothing wrong with this in itself as measuring the impact of education programs on learning outcomes is critical to improving learning outcomes. The main drawback has been the lack of capacity building, especially within government (see discussion below).

Less well understood is how countries decide to join an assessment program that they were not originally part of. For middle- and high-income countries that are able to finance their own participation in international assessments, the choice to join is likely driven by policy needs to promote competitiveness and economic growth (e.g.: joining PISA or TIMSS). Countries also join to help set standards for their own curriculum and assessment. For low- and middle-income countries where resources are far scarcer, donors are often the main financer of participation in international assessments. The motives of development partners (in addition to project M&E) are development-focused: to generate dialogue about education quality and investment needs, support government efforts to improve transparency, support international measurement (e.g.: 4.1.1) and policy research more broadly. There appears to be interest in donors and countries to participate in the large international assessments, but as discussed, these may not be the best choice for countries depending on the difficulty and how the data will be used.

For these reasons, governments need to be well prepared to choose and participate in an international assessment. The need for countries to be well prepared to participate in an international assessment is emphasized repeatedly throughout this Buyer's Guide. This includes having a clear strategy and policy about which assessments a country should prioritize (including national assessment among the choices) as well as establishing a strong enabling environment. A clear strategy and policy would help better coordinate donor efforts, and help ensure countries fully benefit from their participation in international assessments.

Senior policy makers are asking for help in navigating the market

The UNESCO Institute for Statistics routinely receives requests for information about comparisons between international assessments--that is, how to choose which assessments are best suited. The types of questions that UIS field staff receive include asking the difference between assessments aimed at low- and middle-income countries, for example between AMPL and LANA, about the purpose and suitability of the large international assessments, what assessments are available for countries to participate in, and what they measure. They ask about the difference between international and national assessments and whether national assessments can be used to report on SDG 4.1.1., for example. In other words, government staff tend to understand the basic characteristics of the main international student assessments, but are less aware of how to compare or choose among the different assessments.

The buyer's guide aims to help empower senior policy makers through better preparedness

The need for the Buyer's Guide stems primarily from the following observed issues:

1. **Countries participating in assessments that are too difficult resulting in data that cannot be fully taken advantage of---**as discussed above, it offers shock value but otherwise the data may not be useful--and there may be other ways to create comparative learning measures.
2. **Countries not benefiting from capacity building or building capacity sustainably**—as in the examples given above and requests from government for technical assistance on using assessment data.
3. **Country requests to UNESCO for guidance on choosing assessments**--including what assessments would be best suited for countries, what are the choices, and how do they compare.
4. **Power imbalance between country governments, donors and assessment programs**—particularly among low and lower-middle income countries who rely on external donors to finance assessments

A fundamental driver for these adverse outcomes is that countries lack preparedness—especially a formal strategy—for participating in an international assessment program and/or pursuing a national assessment program. A lack of an assessment strategy that clearly specifies or prioritizes goals for participating in assessments and how the country will capture benefits is a major constraint. Having a strategy implies that government has worked through which assessment program best fits the country's needs, has planned for how capacity can be built sustainably and how the results would be used. International donors respect existing country policy and strategies and as a result, a country would be

better prepared to work with donors to ensure capacity can be built and the assessment being financed meets government goals. Indeed, the goal of the Buyer's Guide is to promote better preparedness among countries for learning assessments.

Senior policy makers understand the assessment landscape in principle but there are several differences between assessments that tend to be overlooked and are critical for making a good choice of assessment. Senior policy makers know the differences between the major international and regional assessments, including who participates, grade (or age) level assessed, the subjects assessed, and the types of skills at a general level, being assessed. What is less well understood is how to choose between assessments. For example, what are the pros and cons of participating in a major international assessment versus a regional assessment? Or versus establishing a national assessment? The Buyer's Guide offers guidance on making these decisions and in particular highlights five differences between assessment programs that appear to be less understood by senior policy makers and are quite important for choosing. These are (1) how usable the results are for improving learning, (2) how they differ in difficulty and when this matters, (3) how they differ in capacity building opportunities and knowledge transfer, (4) how they differ in country ownership, and (5) how they differ in costs—but also in the benefits from that cost, in other words, thinking of costs in terms of investment.

The objective of the Buyer's Guide is to promote better preparedness for selecting and participating in student learning assessments. It takes the perspective of a buyer's guide to help senior policy makers recognize that they are buyers in a market for student assessments and they have the power of choice. It also helps inform how that choice can be made and what are the key characteristics (often overlooked or misunderstood) of different assessments that need to be considered. The choice includes not only international assessment but national assessments as well.

The purpose of the Buyer's Guide is <u>NOT</u> to (1) disparage the main international assessments (PISA, PIRLS, TIMSS) or (2) to measure SDG 4.1.1. While the Buyer's Guide offers critique about the participation of some countries in the main international assessments, particularly when the assessments prove to be too difficult for a country's students and the data is not well used, the immense value that the main international assessments contribute is well understood globally. When a country participates in one of the main international assessments and the assessment is of limited use because it was too difficult or the data does not get used, this benefits neither the country nor the assessment provider. Finally, while this Buyer's Guide is produced by the UNESCO Institute of Statistics, it recognizes that reporting on SDG 4.1.1 is just one of many goals for participating in an assessment that are considered (see **Box 2** below).

There are many resources that already exist that help provide guidance on selecting and planning for large scale student assessments; this Buyer's Guide highlights key issues and aims to promote these resources. There are a number of resources related to the issues raised in this Buyer's Guide, and the purpose of the Buyer's Guide is to highlight to senior policy makers the key considerations for choosing among different student assessment options--and planning and preparing a strategy or policy on student assessment--that often get overlooked.

Table 1. Resources related to preparing for student assessments including policy and enabling environment issues (not technical aspects)

Title	Organization	Reference	Description
Primer on Large Scale Assessments of Educational Achievement	World Bank	Clarke & Luna-Bazaldua 2021	Provides guidance to stakeholders on strengthening countries' participation in student assessments including developing political and financial support, careful planning, precise implementation, technical capacity, and timely and clear reporting.
CORE-E / Capacity Needs Assessment	ACER	e.g.: Teo 2024	Assists countries with the successful preparation and implementation of PISA during all stages of the assessment, from planning and contextualization to implementation, analysis and reporting.
ANLAS - analysis of national learning assessment systems	ACER & GPE	ACER & GPE 2019	Provides a resource for developing country partners to build effective and sustainable learning assessment systems leading to evidence-based decision making in education policy and practice.
SABER Assessment	World Bank	Clarke 2012	Provides a framework for assessing the strength of the enabling environment for country's large scale student assessment

Ready to buy? Five (high-level) questions to ask yourself first

Country preparation for an assessment is critical for benefiting from participating—there are many resources on how to do this, but here are five high-level questions countries appear to be overlooked frequently. The need for a strategy or policy that reflects country needs and goals and how to benefit from participating in a student assessment (whether international or national) is well documented (e.g.: World Bank Primer, ACER policy work, etc.). The following are five high-level questions that a strategy would address (in substantially more detail) but are often overlooked by senior policy makers.

1. Why am I participating in an international (or national) assessment?

Clearly specifying goals is important because (1) assessment vary by how well suited they are to different objectives and (2) goals can conflict when choosing a specific assessment. For example, if a country's economic policy is to strengthen competitiveness and attract foreign investment, then benchmarking students' skills to other countries around the world may be a goal and the major international assessments (TIMSS and PISA) may be best suited. If a country is struggling with foundational skills and this stems from primary level pedagogy, then an assessment targeting earlier grades may be best suited. Countries may want detailed analysis of the types of skills and competencies specific to their own national curricula and perhaps a national assessment would be best suited. There are many different reasons or goals for why countries participate in a student assessment (see **Box 2**). Goals can also conflict. For example, an assessment that aims to benchmark skills that are relevant from a global economic perspective may not offer rich enough assessment of specific skills and competencies that would be needed to develop teacher professional development. An assessment that aims to raise citizen awareness and advocacy, offering understandable measures of learning to parents, may not be designed psychometrically to measure progress in learning across time, or if run a by an NGO, may not offer capacity building for government.

2. How am I going to use the assessment results?

Assessment results are and can be used in different ways from identifying professional development to improve learning to raising public awareness about the quality of education. This is a question that, though clearly part of why, is often overlooked in planning for assessments—often countries participate in international assessments and their results are published, but there is no use of the data afterwards, either because policy makers are unsure how to use the data or do not have the resources for analysis. Assessment data has a number of uses beyond comparing or ranking against other countries: setting standards, basis of curriculum reform, informing allocation or targeting of resources, identifying PD practices to improve learning, raising awareness, etc. (Kellaghan, Greaney, and Murray 2009).

Planning on how the data will be used needs to be done from the beginning to ensure you are choosing the right assessment and to ensure sufficient budget and personnel. For example, if you plan to use the data to identify where students are struggling to inform teaching practices, a number of questions arise: (1) is the assessment suited for this given the level of your students (see below), (2) what expertise is needed and do you have it? (3) what resources will be needed to finance the use of the data afterwards?

Box 2: Potential goals for participating in an international student assessment

- **Raise public awareness:** through the media and civil society to better understand education quality, accountability
- **Enable academic research:** promote better educational outcomes, policies, accountability, etc.
- **Compare to other countries:** to understand how competitive countries are
- **Measure progress across time:** to understand whether learning is improving or not
- **Benchmark:** to international or national norms (e.g.: proficiency levels, global MPL, national reference points)
- **Report on international commitments:** SDG 4.1.1 and other related indicators
- **Identify interventions to improve learning:** pedagogical decisions, informing teacher professional development and coaching, teaching resources, etc.
- **Evaluate:** specific interventions or investments or policy reforms,
- **Set national targets and plan:** to allocate national resources, make policy decisions
- **Revise national benchmarks and curricula:** based on international assessment frameworks
- **Build national capacity:** for assessment design, implementation and policy analysis (i.e.: facilitate above)
- **Ownership and pride:** ensure that assessments are culturally relevant and reflective of a region's values

Also, assessment participation can have unexpected consequences, for example, what happens if your country performs poorly, how will you manage the public's reaction? Many governments, particularly those of low- and middle-income countries seem to be taken by surprise by how poorly their student perform in international student assessments. This is, in part, due to countries not understanding how assessment programs differ in terms of difficulty but more fundamentally about not carefully planning which assessment to choose and planning ahead for managing expectations. Media (as discussed below) tend to focus on results and often are critical of government, but do not discuss how results can be improved or what the problems are.

3. Do I have a strategy or policy that addresses participation in assessments?

To prepare for participating in an assessment program, a strategy is needed specifying how a country's goals are to be met—this strategy may be implicit in policy already but many countries lack one. This is not a new point but highlighting this point is a goal of the Buyer's Guide. Assessment system diagnostic tools by the World Bank (2016), ACER (e.g.: Teo 2024), and others discuss the importance of the policy or legal enabling environment and how to develop it. Countries with strong assessment capacity establish an enabling environment either through explicit assessment strategies or policies or implicitly through education strategies and policies more broadly. In many counties, particularly in low- and middle-income countries, such a strategy does not exist. The result is missed opportunities to build capacity, duplication of efforts, and donor financing not well orientated towards larger term goals or needs of government.

An effective strategy would specify goals, needs from assessments programs, and how to improve capacity for assessment which includes everything from personnel's technical skills to legal frameworks.

ACER's Capacity Needs Assessment framework (to draw on one example) used by PISA's Capacity Building and Implementation Support option examines three dimensions for strengthening the assessment system: enabling environment (laws, institutional organization, budget sustainability, system alignment, etc), organization-level capacity (staffing, IT resources, infrastructure, etc), individual-level (availability of skills to do various tasks). Under legislation and policy, the framework assessed to what extent large scale assessments are subject to laws or policies. For example, Tajikistan was rated as "established" under this category as laws were established to create a national assessment center to administer large scale assessments (Teo 2024) while Egypt was rated as "emerging" as only national, not international assessment, was addressed by law or policy (OECD 2023).

4. Am I prepared to work with donors on financing an international assessment?

Donors are willing to finance student assessments for a variety of reasons, and governments need to be ready to leverage donor interest towards their own goals for participating in an assessment program. External donor finance student assessments in order to help identify policy or investment programs, measure global indicators including SDG 4.1.1, facilitate research and monitor and evaluate projects. At the same time, donors' financing of student assessment can be project-focused, and they can overlook overarching assessment needs of the country or government goals. For example, financing participation in a major international student assessment would be beneficial for a donor project that aims to create dialogue around improving competitiveness and skills, but it may not be aligned with students' ability resulting in limited use for improving pedagogy. Implementing EGRA multiple times may be beneficial to evaluating the impact of a teacher professional development project, but if implemented by experts outside of government or separately from government, then it may not create national capacity.

Having a strategy or policy for student assessment would help ensure that donor efforts are better aligned with national goals and lead to sustainable capacity building. Generally, international donors especially international organizations respect government policies and procedures and capacity building is a goal of investment projects. Having a clear policy or strategy towards assessments would help direct donor financing of assessment towards national goals. For example, if the assessment policy is prioritizing assessments that can inform pedagogy needs for students, then this would send a clear signal to donors that if they choose to finance an assessment, it should address this need. If a strategy states that the national assessment agency must implement assessments including donor financed ones for projects, then the assessment agency would gain the expertise and experience in running this type of assessment.

5. Do I have an institutional home for assessment that can accumulate expertise?

International assessments are often conducted in-country by an ad hoc team or by a permanent team that lack an institutional home—the result is loss of institutional memory and expertise. Having an institutional home is important because the training and experience that go along with participating in an international assessment can be retained; e.g.: through the development of implementation manuals and guidelines, detailed descriptions of how previous assessments were implemented, etc. In cases without an official institutional home, changes to the team implementing the assessment (if the team is permanent) result in a loss of experience and history. In many cases, the team may consist of a group of individuals from the Ministry that do not normally work together. Establishing an institutional home often

goes hand-in-hand with national laws or policies, for example the national assessment centre for Tajikistan was established through legislation on assessment participation. One challenge is that international assessments are not conducted frequently, potentially resulting in “down-time” for staff. One solution to this is to use the same agency for national assessments and even national examinations (e.g.: Ethiopia’s National Assessment and Examination Agency). Assessments can also be conducted by a research agency (for example, Thailand’s Institute for the Promotion of Teaching Science and Technology is responsible for PISA).

Key characteristics of assessments that senior policy makers need to know and weigh

This Buyer's Guide focuses on several characteristics of international assessments that are critical for choosing between assessment programs and are not given sufficient attention. Large-scale assessments differ in numerous ways, and senior policy makers are generally aware of the "obvious" differences between assessment programs (grade or age of student assessed, subjects, other countries that participate, etc.). Given that the goal of the Buyer's Guide is to improve country decision making across assessments, the Buyer's Guide offers a comparison across features that are important for choosing but are often overlooked until after the assessment has been conducted. Note that these comparisons aim to highlight key decisions and are not exhaustive, and they ideally would inform the development of an assessment policy or strategy.

The suitability of a specific assessment to a certain goal is not always static—assessment programs can and do evolve to meet demands of participating countries. Throughout this analysis assessment programs are compared on different features; however, it is recognized that these features can change. For example, assessment programs can offer new capacity building options (e.g.: PISA see below) or change governance practices, or even add additional items to enable measurement of ability better matching a specific country. In other words, a number of these characteristics of assessment programs are not fixed but rather they are what has been done historically.

Table 2. Comparison of Student Assessments

Some information is still to be collected and noted as "to be added" in the table.

Assessment	Actionability	Difficulty	Capacity Building		Ownership: Governing board	Fees as a percent of total cost
			Typically included	Extra / ad hoc		
International Assessments						
PIRLS	Second most actionable: (a) can potentially understand specific skills and knowledge students are struggling and (b) 4th grade TIMSS , and PIRLS conducted at earlier grades where outcomes more malleable; however, evidence is lacking on how this information can translate into improved results, plus other caveats as discussed	International assessments tend to be most difficult as they originally emerged for higher-income countries	Implementation	Implementation, reporting and use of assessment (e.g.: through IEA-ETS)	Countries able to be represented by one or more entity (e.g.: Ministry of Education or research institute) but must be able to finance participation	29%
TIMSS 4th grade						29%
TIMSS 8th grade	Third most actionable: can be used to determine knowledge and skills students are struggling with, but being later, the results are also an accumulation of previous years'					
LANA		(*to be added*)				(*to be added*)
PISA		International assessments tend to be most difficult as they originally emerged for higher-income countries	Implementation	Enabling environment, implementation, reporting and use of assessment (e.g.: with ACER)	OECD members only with a few exceptions	28%

	learning, potentially making the results less malleable, also for comparative policy work but this has longer term horizon.					
Regional assessments						
PASEC	Second most actionable, see PIRLS and TIMSS 4th grade above	Lower difficulty, based on Rosetta Stone mapping	Item design, implementation	use of assessment	PASEC steering committee members?	41% (PASEC 2024)
ERCE		Lower difficulty than TIMSS based on Rosetta Stone mapping	Item design, implementation, enabling environment, use of assessment	(*to be added*)	(*to be added*)	44% (TERCE)
SACMEQ		Lower difficulty, based on Global Proficiency Levels and/or Minimum Proficiency Levels mappings	(*to be added*)	(*to be added*)	(*to be added*)	30%
SEAPLM		Similar to TIMSS and ERCE based on Global Proficiency Levels and/or Minimum Proficiency Levels mappings	Implementation, use of assessment (for education reform)	(*to be added*)	(*to be added*)	0%?
PILNA		Lower difficulty, based on Global Proficiency Levels and/or Minimum Proficiency Levels mappings	(*to be added*)	(*to be added*)	(*to be added*)	27%

Others						
EGRA	Most actionable: interventions to improve results are well-known and have been rigorously evaluated	Foundational reading skills, but still many zero scores	N/A	N/A	N/A	N/A
AMPL	Less actionable-- Used for making international comparisons	Designed to be incorporated into an assessment of any level of difficulty	Implementation	(*to be added*)	UNESCO membership	(*to be added*)
Nat. Assessments	Generally, second most actionable if early (see PIRLS above) or third most if later grades (see TIMSS 8th grade above)	N/A -- depends on assessment design, but theoretically capable of matching the student ability level well	N/A	N/A	Country government	N/A

How actionable are different assessments' results?

Student assessments vary but are generally limited in their ability to identify (especially specific) policies or interventions that will improve their results. For example, assessment data identifies the types of skills and knowledge students are struggling with, but there is currently little evidence that designing interventions based on this information can improve results—with the exception of EGRA. Generally, assessment data provides insights and motivation for interventions or broader education reforms, the fruits of which are only occasionally captured in assessment results.

Low- and middle-income countries often experience poor results in an international assessment with highly negative coverage in the media with little constructive (i.e.: actionable) criticism. A familiar experience is that a country has poor results in an international assessment (perhaps participating for the first time), and the government being chastised in the media for these poor results. Media coverage offers little in terms of how to improve the results, but rather focuses on how poor they are, especially how the country ranks against others. For example, the OECD conducted a study of media coverage about PISA 2006 and found that news articles discussed results but only 2 percent of news articles mentioned policies and evidence for how to improve results (Lockheed, Prokic-Breuer & Shadrova 2015). And one reason for why media does not cover what needs to be done to improve assessment scores is that, very often, the data does not offer a clear answer—in other words, the results of many assessments are not clearly actionable.

However, international assessments vary in how prescriptive they are for improving results, and EGRA is arguably the most actionable. Development partners often receive requests from governments for technical assistance on how to use their assessment data to improve learning outcomes after participating in an international assessment; however, the ability of the data to be prescriptive, that is, to offer an indication of how to improve learning is often more limited than policy makers expect. This is for a number of reasons. One is that international assessments measuring learning in secondary school may not be malleable if the bottlenecks to learning are earlier in the education cycle, for example, poor pedagogy in primary leading to a lack of foundational skills—whatever improves could be identified from a secondary-level assessment would not have impact if foundational skills are lacking. Another reason is that assessments may not offer information about student skills and competencies that is specific enough to adjust teaching practices. Item analysis can help, but often test items are not publicly released. A third reason is the lack of a model of taking the assessment results and transforming them into, for example, a professional development program. This is why EGRA emerges as the most actionable of the assessment programs: it measures foundational skills at the very beginning of primary school. It offers measurement of very specific skills and competencies that students need in order to learn to read, and there is a well-defined professional development program that has been rigorously evaluated using randomized-controlled trials that can improve results (Piper & Korda 2011; Piper, Zuilkowski & Mugenda 2014; Lucas et al. 2014; Kerwin & Thornton 2021 Macdonald & Vu 2018; Macdonald et al. 2018). In other words, with EGRA results, there is a clear path for how to improve these results. However, it should be noted, that EGRA can be resource intensive as it requires one- to-one testing, and it may not be applicable for contexts where students are not struggling with foundational reading skills. Also, how to implement the "EGRA" intervention at scale is not well researched.

International assessments especially at earlier grades have the potential to be prescriptive by identifying skills and knowledge that students struggle with most, but evidence on the effectiveness of these

interventions is currently limited. Assessment data can identify the skills and competencies that children struggle with which in turn can form the basis for professional development. This can often be done through item analysis, by identifying how difficult the item is (or whether many children struggle with it), knowing the underling skill being tested, knowing what the common misconceptions are (e.g.: distractors), etc. The amount of detail varies from assessment to assessment as assessments differ in where items are on the difficulty scale and the scope of a topic being assessed. National assessments offer an advantage because the assessment agency would have access to the test items and can design them to fit their own curriculum but also to target perceptions of where students are struggling. It may be that the reason why students are struggling with the identified skills and competencies is due to poor foundational skills as student assessment effectively capture the accumulation of learning over the course of a student's educational experience. As a result, assessments done earlier, in primary and even at earlier grades within primary, may be the most relevant in terms of professional development being able to improve learning. For this reason, the earlier assessments are ranked higher in our Buyer's Guide than later assessments, although for more advanced education systems with very strong foundational skills, the later assessments may be more relevant. However, a significant limitation is that (to our knowledge), there is very little rigorous evidence showing that professional development derived from item analysis can improve learning outcomes (unlike EGRA where there are number of randomized-trials showing impact). Effective PD for learning does not solely rely on its content but also modality, including type of training and coaching, and these other aspects complicate the generation of evidence for these approaches.

International assessments offer crucial information for comparing policies across countries as well as (though rarely) evaluation of reforms. So far, the discussion has been around using assessment data to identify specific professional development interventions to improve learning, as the most directly or actionable channel for using assessment results. International assessments are more likely to be used for doing cross-country policy comparisons (e.g.: see Clarke & Luna-Bazaldua 2021; Lockheed, Prokic-Breuer & Shadrova 2015. In some cases, this analysis can offer quite actionable, that is, specific recommendations. For example, Burundi's 2nd grade PASEC score far exceeded other African PASEC countries, demonstrating the important of language of instruction matching the language used by families. International student assessment data (as can national assessment data) can also be used to evaluate specific countries' reforms. For example, Jakubowski et al. (2010) evaluated the impact of Poland's 1999 education reform, finding that it improved learning outcomes. This provides actionable advice in that the reform should be sustained. Generally, however, the international assessment data informs research but does not offer a very specific prescription for countries.

How do assessments differ in difficulty and when does this matter?

The difficulty of regional assessments as well as small scale assessments tend to be closer to the ability of their students while PISA, TIMSSS and PIRLS tend to be closely aligned to high income countries. If an assessment is too difficult, the data may be less informative for improving teaching and potentially embarrassing for governments; however, performing poorly on a major assessment can also be a powerful wake-up call for education reform and investment.

Assessments differ in their difficulty meaning that the range of student abilities being tested may not overlap well with the range of student abilities in a country. For example, an analysis Rutkowski, Rutkowski and Liaw (2019) compared the student ability level tested by PISA-D's test item and the

distribution of student ability in Zambia which participated in PISA-D (see **Figure 1** above). The study found that PISA-D test item were concentrated around an ability level much higher than Zambia's. About half of students scored at a level below the achievement level tested by PISA-D. In other words, the PISA-D assessment was far too difficult for Zambian students. The result was two-fold. First, the country had a very low achievement score (and low ranking) which clearly exposes the challenges faced in education quality and would also have exposed government to criticism. Second, the information that could obtained from analyzing the test items (as discussed previously) would only cover the top half of students in Zambia; there would be very little information to inform professional development for the lower half of the distribution.

For example, based on various studies aiming to compare international assessments including UNESCO's Rosetta Stone, regional assessments tend to be less difficult than the PISA and TIMSS. Comparing difficulty in assessments is well researched in the literature but in broad terms could be done either by reviewing definitions of proficiency levels between assessments or by psychometrically linking assessments by having assessments incorporate test items from other assessments. For the Buyer's Guide, the difficult of assessments was determined based on (1) UNESCO's Rosetta Stone project (UIS 2022) that linked ERCE and PASEC to TIMSS and (2) the policy linking work of GAML (see: UIS 2023b) which aims to link proficiency levels between assessments. The Rosetta Stone found that (by comparing means) that PASEC was the easiest, followed by ERCE with TIMSS being substantially more difficult[2]. Based on GAML's policy mapping, mathematics proficiency levels for PASEC, PILNA and SACMEQ were lower (suggesting easier assessments) compared to the proficiency levels for TIMSS, SEA-PLM and ERCE. Broadly speaking, the regional assessments tend to be substantially easier than the international assessments, especially considering that comparisons are being made between 6th grade regional assessments to 4th grade TIMSS. The reason why is quite clear: regional assessments are designed to assess the skill level of their country members whereas TIMSS and PISA began predominantly for high income countries. The result is that countries that participate in the main international assessments are likely to do much more poorly than in their own regional assessments. While this may result in a true international benchmark for student performance, the information may noy be as helpful for identifying professional development or other interventions needed to improve learning.

National assessments and some small-scale assessments including EGRA are highly adapted to local context and to foundational skills and tend to test more within the range of student ability. National assessments have the most potential for assessing a country's difficulty level correctly because they are owned nationally and can be tailored to the country's own ability distribution. Other small scale learning assessments also exist that are highly adapted to local contexts as well. UNESCO's AMPL offers modules that can be applied to assessments to permit measurement of global minimum proficiency in order to improve comparability across assessments that are highly adapted to specific country contexts; when this occurs, AMPL is effectively an international assessment with a difficulty level tailored to each individual

[2] From the Rosetta Stone, average 6th grade ERCE and PASEC scores for mathematics were mapped to 4th grade TIMSS scores of 465 and 336, respectively—note the grade difference, suggesting that TIMSS with an average of 500 is substantially more difficult than ERCE or PASEC. For reading, 6th grade ERCE and PASEC were mapped to TIMSS 4th grade scores of 482 and 317 on average. These mappings are estimates and the mapping values presented here are the point estimates.

country. Finally, foundational learning assessments including EGRA may also be better matched to countries with low learning outcomes due to poor foundational skills, although even EGRA tends to have a high number of "zero scores" in low-income countries.

Having an assessment with a difficulty level well matched to the abilities of your students enables you to better to identify the skills and knowledge that students lack and to improve teaching. Assessments that are too difficult, as in the Zambia case discussed previously, will not provide data that is informative for identifying specifically the types of skills and competencies that students need strengthening and subsequently the types of PD needed. Likewise if the assessment is too easy. If the goal of participating in an international assessment is not only to measure learning but also to identify how to improve learning, then ensuring correct difficulty would be a top priority for assessment choice.

However, when an international assessment is very difficult and a country scores or ranks poorly, this can serve as a wake-up call to government and citizens and motivate educational reform and investment. Participating in an assessment that is too difficult but that also that represents a global benchmark (e.g.: PISA or TIMSS) can offer a shock to the system to demonstrate to policy makers, the public and donors about the realities of a country's education system. The potential for poor performance on an international assessment to motivate government and public to reassess the quality of the education system is demonstrated by Germany's "PISA shock" (Breakspear 2012). Participating in an assessment that is too difficult would clearly expose low learning in a country relative to others. However, being too difficult would also result in the data not being highly relevant to the country's ability distribution, limiting the use of the data to identify PD. Also, the development of tools like AMPL enable international comparison between assessments that are well suited to student abilities of each country.

How do assessments differ in capacity building and knowledge transfer?

Capacity building activities offered by assessment providers have evolved beyond training to include the broader enabling environment for assessments including policies and financing; however, there are important differences in the type of capacity building that is included with assessment provider fees versus at additional cost or provided on an ad hoc basis. There are also differences in the extent to which local experts are involved in item design.

Capacity building in the form of training and knowledge transfer but also institutional and regulatory reform offers a significant benefit to countries from participating in an international assessment. Building the skills of government staff and local experts has long been a benefit of participating in international learning assessments, and traditionally this has been accomplished through training to country representatives including a national coordinator or a country team. Typically, countries are responsible for adapting and implementing an international assessment, and the national coordinator receives training on how to adapt the instrument to the local language, determine the sampling frame and generate the sample, and on the standards for implementation and rolling out the assessment. By participating in an international assessment, country coordinators or teams develop expertise in assessment implementation through their training and experience. This expertise obtained by government staff becomes transferable to other assessments including development of national assessments if government is well prepared (see discussion above). Traditionally this capacity building has been focused

on the implementation of the assessment, but more recently there is recognition of the need for a broader capacity building including use of assessment and development of countries' enabling environment including national staff expertise, infrastructure and laws and policies. However, the type of capacity building that is included in international assessment fees and those that require additional fees, and those that are typically provided by donors, still vary by assessment, and generally there is little involvement of countries in the development of test items.

The IEA and OECD now offer a broad range of capacity building including using assessment results and the enabling environment, but accessing this requires additional fees. Prior to 2015, capacity building offered by IEA and the OECD was typically focused on assessment implementation including technical quality standards, adaptation, sampling requirements, hands-on software training among others (Lockheed, Prokic-Breuer & Shadrova 2015). Capacity building offers have since broadened. The IEA and ETS jointly run the International Education Research Institute (IERI) which offers capacity building on a wide range of statistical topics related to assessment data. OECD now offers a broader set of capacity building through its Capacity Building and Implementation Support Option (CBIS) option; this is implemented by ACER and assesses a country's capacity for international assessment at the enabling environment level (legal, institutional and financial context), the level of the implementing organization (staffing levels, infrastructure, IT availability) and at the individual level (skills of specific team members).

Regional assessment programs tend to include a wide range of capacity building but depending on the assessment may be ad hoc and reliant on donor financing. A review of websites and knowledge of activities by different assessment providers provides a picture of the type of capacity building provided by countries and to what extent it exceeds a focus on implementation. For example, Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE, providers of ERCE) provides capacity building on assessment design and on the use of assessment results, that is, beyond implementation of assessments. PASEC has organized a number of workshops on capacity building on use of assessments that are donor financed, but significant constraint is that many of its member countries do not have established assessment units. SEA-PLM offers training on use of assessment and linking assessment results to education reform.

Few assessment programs have country experts involved in core design of the assessment including item development which limits knowledge transfer on an important aspect of assessment design. For both PISA and TIMSS, item design is conducted by select group of experts rather than representatives from various member countries, this is in part practical given the large number of countries involved. But the result is that country participants do not gain experience in designing items for assessments. Regional assessments tend to involve their member country teams in item design though to differing extents. For example, PASEC recently organized a workshop on item design, and Laboratorio offers assessment design as part of its capacity building. What is not clear is how hands-on these workshops are and whether participants are actually developing experience in item design versus learning about item design. Item design is an important skill to develop for establishing and improving the quality of national examinations; it also helps ensure countries' needs are represented in the design of the assessment.

How do assessments differ in country ownership over the program?

Country influence over assessment programs generally depends on membership in the overarching international body that owns the assessment and on the number of members, but country influence is also affected by dependency whereby assessment programs may be difficult to change after they have been established and running and by the de facto influence from international experts and donors.

Ownership is an important consideration for student assessment choice, reflecting the ability of countries to ensure assessments stay relevant to their contexts, but also for accountability and pride. Countries have different contexts and therefore different needs from assessments, and the ability of countries to have ownership over assessment programs helps determine the relevance of an assessment program for a country. This includes difficulty: regional assessments may be more closely aligned to the student ability of a country if other countries in the region have similar student ability levels. But it also includes the assessment framework and the types of skills being assessed, and how these might related to local job markets or local curricula, as well as capacity building activities offered. Ownership also ensures that assessment programs stay relevant as contexts change, as skill levels increase or as policy makers become interested different aspects of an education system. Ownership and accountability are closely related; strong ownership results in countries having more control and oversight over the quality of the assessment products for their countries. Finally, assessment programs, especially regional assessments offer a point of pride for countries, in having their assessment that meets their own needs and is not imposed upon from elsewhere.

Assessment programs vary in the extent to which countries have ownership including through representation on boards and the number of members. For example, PISA is owned by the OECD, and its governance board includes OECD members only (with a few exceptions) while partner countries (non-OECD) members have observer status. This is not to say that the needs of low and middle income countries are not being addressed by PISA, but the fact is that low- and middle-income countries have a *de jure* disadvantage in their ownership over PISA and an unequal status. IEA is governed by IEA members, and countries are able to join the IEA through membership of one or more of their educational institutions. Regional assessments offer a higher level of ownership generally as they are smaller and owned by regional international organizations in which their countries are members. For example, PASEC is part of CONFEMEN but government by PASEC members specifically. PILNA is governed through the South Pacific Community. SEA-PLM is implemented by UNICEF but its main governing body, the Regional Steering Committee, consists of the South East Asian Ministers of Education (SEAMEO) countries. National assessments, by definition, have the highest ownership as they are owned and implemented by governments themselves.

Even with official representation on governing boards, country influence can be mitigated by path dependency—historical precedent of an assessment—and by de facto influence of international experts and donors. The *de jure* authority that countries have over assessment programs can be mitigated by a few factors. First, some features of assessment programs cannot change quickly and historical precedent plays a significant role in international assessments; for example, PISA would be unlikely to change to assessing early grade primary students in a reasonable planning horizon. Grade levels and skills assessed tend to persist across time. In regional assessments, especially, international experts from outside the region may participate in scientific committees or other technical groups, and their opinions, by virtue of

their reputation for instance, might override or conflict with the opinions of regional experts. Donors also have substantial influence by virtue of their financing of assessment programs and this gives donors, often Global North countries, *de facto* influence over assessment program decision making away from official governing members of an assessment. This is also not static but a process. For example, LLECE was established with significant financing and technical support from the global North but has since become much more autonomous in terms of its financing and expertise.

How do assessments differ in costs--and in returns?

The cost of participating in an international student assessment goes well beyond the contributions to the assessment provider and include staff time for implementing the assessment as well as using the assessment data afterwards. Different have assessments have different cost drivers. However, assessments also differ in the extent to which they provide future return. For example, on paper, a national assessment may be more costly to establish than joining an international assessment program, but the cost of a national assessment is in fact an investment in a country's own assessment capacity.

The cost of implementing an international assessment go far beyond the "assessment fees" that are charged by providers and include staff-time, printing, analysis and dissemination. When countries participate in international assessments, they are typically responsible for financing a wide range activities related to the design and implementation of the assessment. These activities include test translation and back translation, designing country-specific background questions, pilot testing the assessment, printing, disseminating and supervising the assessment, data entry and marking, and the overhead costs of government supervision and implementation (see UIS 2016). A study by Wagner, Babson and Murphy (2011) reviewed international and national assessments in a selection of countries to identify costs (see also Wolff 2007). Based on their work, the cost of test application (printing, distributing field testing and supervising) was the largest cost, ranging from 34 percent up to 80 percent of the total cost of the assessment while post-processing including data entry, marking and analysis was the second highest cost item, ranging from 3 percent up to 25 percent, depending on the assessment (see **Table 3**). In their study, the assessment fees paid to the assessment provider were only ranged from 5 to 7 percent of the total cost, but this data was available only for PISA 2009.

Table 3. Examples of activities needed to design and implement and assessment and their cost as percent of total cost (range)

Activity	National assessments 2003-04	PASEC 2010	SACMEQ III 2007	PISA 2009	EGRA 2008
Test preparation: test items, pilot testing, training	9% - 20%	18%	6% - 8%	3% - 17%	6% - 12%
Test application: design, printing, distribution, field testing, supervision	41% - 54%	50%	55% - 80%	36% - 74%	34% - 38%

Processing and analysis: data entry, marking open-ended questions, analysis	13% - 25%	7%	21%	9% - 25%	3% - 6%
Dissemination: reports to schools, national reports, public relations	1% - 15%	17%	1% - 2%	1% to 7%	1%
Institutional costs: personnel (staff + consultants), infrastructures, equipment	33%	7%	12% - 16%	7% - 28%	43% - 49%
Test Fees	n.a.	n.a.	n.a.	5% - 7%	n.a.

Adapted from Wagner, Babson & Murphy 2011

Data on the total costs of assessment programs is currently limited; however, a compilation of studies offers insights into the cost of assessments per student assessed. One significant driver of cost is the number of students assessed, and the variation in sample size is usually reflective of the extent to which countries want accurate estimates for sub-populations versus having an accurate estimate for the population of students as a whole. For example, Wagner, Babson and Murphy (2011) noted that Chile's national assessment, SIMCE, in 2004 sampled 300,000 students while Tanzania's SACMEQ III in 2007 sampled only 3,000 students. Looking at the cost of per student assessed offers an initial comparison of costs across different assessments (see **Table 4**). Based on this measure and the studies available, national assessments tend to have the lowest cost per student sampled, ranging from 11 to 27 USD (in 2024 terms, adjusted for inflation). PASEC tends to have the highest costs per student, at 114 USD for PASEC 2024. The reasons for why assessments cost more or less per student are not well documented so far. One driver is the number of students assessed, as there are economies of scale, for example, spreading of the design and setup costs across more students. Note that the data available is quite limited, and the actual costs are expected to vary substantially across countries, given the range of costs already observed. Finally, the World Bank is undertaking a rigorous costing of international assessments in Sub-Saharan African which will provide a better picture of cost comparisons across assessment activities[3].

Table 4. Cost per student of selected assessments (converted to 2024 USD)

Assessment	Sample	Cost per student	Study
National assessments	3 LAC countries	11 - 27 USD	Wagner, Babson & Murphy 2011
PASEC 2010	Average	48 USD	Wagner, Babson & Murphy 2011
PASEC 2019	Average	150 USD	Hounkpodoté 2024
PASEC 2024	Average	114 USD	Hounkpodoté 2024
SACMEQ III 2007	2 countries	71 to 76 USD	Wagner, Babson & Murphy 2011
PISA 2009	4 LAC countries	median 52 USD	Wagner, Babson & Murphy 2011
PISA 2015	Average	51 USD	UIS 2016
TIMSS 2015	Average	75 USD	UIS 2016
PIRLS 2016	Average	64 USD	UIS 2016
ICILS 2013	Average	81 USD	UIS 2016
ICCS 2009	Average	85 USD	UIS 2016
EGRA 2010	2 countries	43 - 89 USD	Wagner, Babson & Murphy 2011

[3] The findings of World Bank's costing work will be included in updated drafts of the Buyer's Guide.

Adapted from Wagner, Babson & Murphy 2011; USD adjusted for price inflation

The assessment fee as a percent of the total cost to countries for an international assessment provides a useful comparator for understanding relative costs and additional costs countries may incur. UNESCO has been compiling cost to countries for participating in different international assessments for a variety of institutional documents (e.g.: UNESCO 2023; Montoya 2023; UIS 2018). The source of their data includes previous studies (including Wagner, Babson & Murphy 2011) and data collected directly from assessment providers. The comparison table (**Table 2** above) presents the assessment fees as a percent of total costs. This cost share ranges from 27 percent for PILNA up to 44 percent for ERCE. These costs may not be presented as fees, but rather as "international costs", but the important thing to note is that the international costs or fees do not comprise the majority of the cost to the country, the additional costs do. Second, the actual costs vary by country and depend on a wide range of factors (cost of staff time, geographic size, sample size, and so on). Finally, these costs are based on what has been happening historically with the assessment. Countries may opt to pay more to do additional capacity building (see above) or additional analyses.

Even if donor-financed, it is critical that governments understand the cost given that donor financing has clear opportunity cost. In many cases for low- and middle-income countries, donor financing is provided to cover both the international fees or contributions for participating in an assessment as well as the national costs of implementing the assessment. However, these costs should not be ignored--this represents financing that could be used for other purposes within the country, in other words, there is always an opportunity cost.

Cost, however, only tells half of the story: how that spending translates into an investment with a future benefit to countries can vary and should not be ignored. Assessments differ in their ability to meet the needs and goals of a particular country, for example in how much one can learn about an education system, how they inform professional develop needs, and the type of capacity building provided, among others. The benefits in terms of capacity building is demonstrated in comparisons of national versus international assessment costs, and the opinions that emerge in the debate about whether to establish and run a national assessment or participate in an international assessment. However, this, in the view of the Buyers' Guide, is really a false dichotomy. First, national and international assessments are suited towards different goals, for example, ensuring assessment is well aligned with the country's own curricular goals, students' needs, and ability of students while international assessments offer a higher level of comparability across countries. Second, in this debate, some note that national assessments are much more costly to establish and run than participating in an international assessment. Partly this higher cost emerges from larger sample sizes in national assessments. For example, Chile's national assessment SIMCE in 2004 cost nearly two million USD (well more than double the cost of its participation in PISA in 2009), but SIMCE sampled 300,000 students versus less than 6000 in PISA 2009; this resulted in the national assessment having a lower per student cost (Wagner, Babson & Murphy 2011). Similarly, Uruguay's 2003 national assessment cost 8 USD per student while participating in PISA in the same year cost 21 USD--again the national assessment cost more because of a larger sample. However, the important thing to note is that establishing and running a national assessment involves creating an assessment unit including the cost of personnel, training and infrastructure which often ignored when participating in an international assessment. In this sense, the country is benefiting more from a national assessment because of the

creation of capacity which not only may not happen from participating in a one-off international assessment, but certainly should be done in order for the country to fully benefit from the international assessment. Even if national assessments cost more, there can be a larger benefit through the capacity created or in other words, the cost of national assessment represents an investment rather than a recurrent cost.

Buyer beware: common pitfalls and how to avoid them

The following highlights and reiterates a few key pitfalls that the buyer's guide is aiming to help countries avoid. There are numerous pitfalls in choosing and implementing a student assessment; however, the Buyer's Guide focuses on three: participating in overly difficult assessments, not using assessment data, and not retaining capacity.

1. Participating in an overly difficult assessment

As discussed previously, a number of countries have participated in international assessment that are too difficult for their students but do help with international benchmarking. The rationale to participating in an overly difficult assessment is generally to benchmark student achievement to a global standard, in comparison with a large number of countries. For example, participating in PISA or TIMSS allows a county to assess their student based on these assessment programs' assessment frameworks including their proficiency levels which are viewed as being relevant in a global sense. It also allows comparing results to a large number of countries. These assessments are also very rigorous about their comparability over time; other assessments (some national ones, some versions of assessments such as EGRA), sometimes are not quite as rigorous in their comparability over time or may not even have been designed to allow it. Doing poorly on assessments like these offer highlights for government, the public and development partners, the urgent need to invest and improve education quality.

However, data may not be informative for improving results (especially if foundational skills are weak) and governments often experience embarrassment from the poor results. Looking at PISA and TIMSS participants, there are handful of countries that score very low. There are also cases of countries being embarrassed by poor results and these poor results are often cited (in commentary) for reasons why countries may not continue participating in an assessment program. The example of Zambia participating in PISA-D (above) illustrates how the misalignment between test items and student ability limits the useability of data. More broadly, the issue is that the major international assessments are assessing higher level cognitive skills late in the education cycle when the true constraint is in fact foundational skills that are developed early in primary.

One way to avoid this pitfall is (a) truly understand how difficult the assessment is and (b) be clear on why you are participating in an overly difficult assessment and know the consequences. Current attempts to link major assessments using either the Rosetta Stone project or by technical linking (e.g., Benchmark method) to the Minimum Proficiency Level (see discussion above), provide guidance on which assessments or more difficult than others. The Rosetta Stone mapping tables offer a very easy estimate of what a country's score would be in TIMSS based on the country's score in ERCE or PASEC. If your country would score quite poorly on TIMSS based on the mapping, then you may wish to do further analysis to verify whether participating in TIMSS would provide you with reliable and usable estimates of student ability. Based on this, and knowing that you would do poorly in an assessment like PISA or TIMSS, then you would need to weigh your goals. Does the need for international benchmarking outweigh the reputational risk to government (not to mention that the data may not be usable to improve outcomes)?

Note that with the advent of tools like AMPL, countries can link their own assessments (including national assessments) to a global proficiency level, allowing global benchmarking without participating in an overly difficult assessment. AMPL provides modules that can be inserted in an assessment, including a national assessment, to enable measuring of what proportion of students achieve the global minimum proficiency level. To some extent, this nullifies the tradeoff between participating in an overly difficult assessment like PISA and TIMSS and international benchmarking. This allows countries to assess at a level of difficulty that is relevant to the ability distribution of their students and still provide international comparison. It does not, however, eliminate the reputational risk because comparison is still possible, and it does not replace PISA or TIMSS in terms of the global relevance of the skills these assessment programs assess.

2. Not using the assessment results (or not knowing how results can be used)

Many countries participate in international assessments, but their assessment data is not used. This represents a significant lost opportunity, and two main constraints are (1) that countries are unsure of how to use the data, (2) countries may not have the expertise to do this analysis, and (3) in cases where assessments are too difficult, the data's use may be limited. As described previously, there are a wide range of uses of assessment data by countries. Further analysis to understand how sub-populations have performed, item analysis to identify the skills and competencies students lack, and setting standards for countries own curricula and assessment are several ways in which countries can benefit from their assessment data (e.g.: see Clarke & Diego Luna-Bazaldua 2021).

Understanding how assessment data can be used and planning ahead for analysis after the assessment would help countries get more value out of participating in international assessment. Increasing the value countries get from their assessment data relies on good preparation prior to the assessment. First, how assessment data will be used should be part of answering why you are participating in an international assessment. If the goal is to identify how learning can be improved, then a clear plan for using the data is needed. Part of this is understanding what the data can do and how actionable the data is, which links back to whether the difficulty of the assessment is appropriate for the challenges your students face. An assessment early on would provide data that is more actionable than an assessment later if foundational skills are a constraint for learning. Being able to access the test items to conduct item analysis would also be a consideration. A subsequent question would be about resources and expertise to do this analysis. Answering these questions prior to participating in an assessment is critical because the answers will help the country determine how beneficial a particular assessment would be.

3. Not retaining capacity from assessments

There also many cases in which countries participate in learning assessments, including as part of donor financed project M&E, but do not accumulate the expertise and experience. This can happen especially if participating in an assessment (e.g.: EGRA) is part of the evaluation of a donor financed project. There is a tendency for donor projects to contract firms to implement assessment to monitor the project as this is generally, from a procurement and project-management perspective, far simpler than having a government agency implement the assessment. Some donor partners are unable or prefer to support non-government entities which further complicates capacity building. The example of Sierre Leone described

previously demonstrates this, but the outcome is that government does not gain the capacity to assess student learning.

An assessment policy and institutional home for assessment activities helps address this. If governments have clear policies that assessment activities need to be implemented by a national assessment unit then donor partners generally would comply. It is possible that the national assessment unit alone does not have the expertise to conduct a specific type of assessment, in which case, donor projects would need to help build this expertise as much as possible by ensuring that agency staff are gaining real, hands-on experience. Non-government providers may be needed as a source of human resources for enumerators; however, the more that the national assessment agency does in contracting firms or universities is helpful because provides experience to the agencies and helps the agency build its network of experts.

Towards a more efficient market for student assessments

While the Buyer's Guide presents key considerations for countries in choosing student assessments, the following is highlighting to help strengthen the market for student assessments. Senior policy makers understand the different assessments that exist, and the Buyer's Guide aims to strengthen their knowledge on *how* to choose, that is, to understand the main differences in assessments for choice and to incorporate this into a larger strategy or policy. However, there are other aspects of the market for student assessments that need to change as well in order to strengthen its efficiency. The following are highlighted as they are linked to the Buyer's Guide work or are not well publicized.

1. Donor should encourage countries to have a strategic approach to assessment participation prior to financing assessment: Donor have significant influence over how countries benefit from student assessment, and a greater emphasis should be placed on the enabling environment and more specifically assessment strategy, even if the donor project uses assessment as M&E. This would help countries fully benefit from participating in learning assessments and ensure that assessment is well orientated to the needs of countries. Assessment programs are already recognizing the important of the enabling environment as part of capacity, and donors can take advantage of this capacity building. Note that having a strategy for assessment is not the same as enabling environment: the strategy could simply describe process for developing these, but the main thing is the goals and how to get there, which assessments would be best

2. AMPL offers countries the ability to make international comparisons of minimum proficiency using assessments that are well aligned with student ability. In effect, if countries wish to demonstrate that their student achievement is lagging well behind international norms, incorporating AMPL modules into either a country's national assessment or regional assessment would enable this comparison. It would also allow reporting on SDG 4.1.1.

3. The need for the Virtual Fund for International Assessment: Currently, donor financing for participating in student assessment is often country and project specific (with exceptions of large financing for regional assessments). Creating a common (virtual) fund for financing assessments would improve the efficiency of donor financing for assessment by ensuring that assessments closely meet the needs of each country and that best-practice is present in planning for assessments (e.g.: having assessment strategy, capacity, etc.). The idea in many ways mirrors the Gavi, which pools financing together for provision of vaccines.

References

ACER and GPE (2019). Analysis of National Learning Assessment Systems [ANLAS]. Manual. Global Partnership for Education

Antoninis, M. 2024 On the way forward for SDG indicator 4.1.1a: supporting countries' development needs. Accessed: https://world-education-blog.org/2024/03/26/on-the-way-forward-for-sdg-indicator-4-1-1a-supporting-countries-development-needs/

Breakspear, S. 2012. The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance. OECD Education Working Papers No. 71. Paris: OECD

Clarke, Marguerite. 2012. "What Matters Most For Student Assessment Systems: A Framework Paper." SABER—Systems Approach for Better Education Results series, Working Paper No. 1. World Bank, Washington, DC

Clarke, Marguerite, and Diego Luna-Bazaldua. 2021. Primer on Large-Scale Assessments of Educational Achievement. National Assessments of Educational Achievement series. Washington, DC: World Bank. doi:10.1596/978-1-4648-1659-8.

Hounkpodoté, H. (2024). PASEC: Evaluation in the service of better management of educational systems. Global Partnership for Education. https://www.globalpartnership.org/blog/pasec-evaluation-service-better-management-educational-systems

Jakubowski, Maciej and Patrinos, Harry Anthony and Porta, Emilio Ernesto and Wisniewski, Jerzy, The Impact of the 1999 Education Reform in Poland (April 1, 2010). World Bank Policy Research Working Paper No. 5263, Available at SSRN: https://ssrn.com/abstract=1585062

Kellaghan, Thomas, Vincent Greaney, and Scott Murray. 2009. National Assessments of Educational Achievement, Volume 5: Using the Results of a National Assessment of Educational Achievement. Washington, DC: World Bank.

Kerwin, Jason T. & Rebecca L. Thornton, 2021. "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures," The Review of Economics and Statistics, MIT Press, vol. 103(2), pages 251-264, May.

Lucas, A. M., McEwan, P. J., Ngware, M., & Oketch, M. (2014). Improving early-grade literacy in East Africa: Experimental evidence from Kenya and Uganda. Journal of Policy Analysis and Management, 33(4), 950–976.

Macdonald, Kevin; Brinkman, Sally; Jarvie, Wendy; Machuca-Sierra, Myrna; McDonall, Kris; Messaoud-Galusi, Souhila; Tapueluelu, Siosiana; Vu, Binh Thanh. (2018). Intervening at Home and Then at School : A Randomized Evaluation of Two Approaches to Improve Early Educational Outcomes in Tonga. *Policy Research Working Paper Series* ;No. 8682. World Bank, Washington, DC

Macdonald, K. A. D. and Vu, B. T. 2018. A randomized evaluation of a low-cost and highly scripted teaching method to improve basic early grade reading skills in Papua New Guinea (English). World Bank Policy Research Working Paper Series No. 8427. Washington, D.C. : World Bank Group

Montoya, S. (2023). Reporting learning outcomes in basic education: Country's options for indicator 4.1.1. Montreal: UNESCO Institute for Statistics

Montoya, S. and L. Crouch (2022a) The learning assessment market: pointers for countries – part 1 https://world-education-blog.org/2019/04/26/the-learning-assessment-market-pointers-for-countries-part-1/

Montoya, S. and L. Crouch (2022b) The learning assessment market: pointers for countries – part 2 https://world-education-blog.org/2019/05/20/the-learning-assessment-market-pointers-for-countries-part-2/

Lockheed, Marlaine, Tijana Prokic-Breuer, and Anna Shadrova. 2015. The Experience of Middle-Income Countries Participating in PISA 2000–2015. Washington, DC, and Paris, France: World Bank and OECD Publishing. doi:10.1787/9789264246195-en

OECD 2023. PISA Capacity Needs Assessment. Egypt. Paris: OECD

Piper, B., & Korda, M. (2011). *EGRA Plus: Liberia (Program evaluation report)*. Durham, NC: RTI International. Doi:10.1016/0742-051X(89)90027-9

Piper, B., Zuilkowski, S. S. & Mugenda, A. (2014), 'Improving reading outcomes in Kenya: First-year effects of the primr initiative', International Journal of Educational Development 37, 11–21.

Rutkowski, L., Rutkowski, D., & Liaw, Y. L. (2019). The existence and impact of floor effects for low-performing PISA participants. Assessment in Education: Principles, Policy and Practice, 26(6), 643–664. https://doi.org/10.1080/0969594X.2019.1577219

Teo, I. 2024. PISA Capacity Needs Assessment, Tajikistan, Dushanbe. Paris: OECD

UIS (2016). The cost of not assessing learning outcomes. Information Paper No. 26. Montreal: UNESCO Institute for Statistics

UIS (2018). The Investment Case for SDG 4 Data Concept Note Technical Cooperation Group on SDG 4–Education 2030 Indicators January 2018. Montreal: UNESCO Institute for Statistics

UIS (2022). Rosetta Stone Policy Brief: Establishing a concordance between regional (ERCE/PASEC) and international (TIMSS/PIRLS) assessments. Montreal: UNESCO Institute for Statistics

UIS (2023) Aligning and reporting on indicator 4.1.1: UIS annotated workflow. Montreal: UNESCO Institute for Statistics

UIS (2023b) POLICY LINKING FOR MEASURING GLOBAL LEARNING OUTCOMES TOOLKIT Linking Assessments to the Global Proficiency Framework. Montreal: UIS. https://gaml.uis.unesco.org/wp-content/uploads/sites/2/2021/03/Policy_Linking_for_Measuring_Global_Learning_Outcomes_Dec-2020.pdf

Varly, P. 2022. Challenges In Establishing A Learning Assessment System: The Example Of Sierra Leone. Background paper for the 2022 Spotlight On Basic Education Completion And Foundational Learning In Africa. Paris: UNESCO

von Davier, M., Kennedy, A., Reynolds, K., Fishbein, B., Khorramdel, L., Aldrich, C., Bookbinder, A., Bezirhan, U., & Yin, L. (2024). *TIMSS 2023 International Results in Mathematics and Science*. Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.timss.rs6460

Wagner, D. A., A. Babson, K. M. Murhpy (2011). How Much is Learning Measurement Worth? Assessment Costs in Low-Income Countries. *Current Issues in Comparative Education*, Vol. 14: 3-23

Wolff, Laurence. 2007. "The Costs of Student Assessments in Latin America." Working Paper No. 38, Partnership for Educational Revitalization in the Americas, Washington, DC.

World Bank (2016). SABER Student Assessment. Washington, D.C.: The World Bank. https://documents1.worldbank.org/curated/en/410721496308370728/pdf/SABER-Student-assessment.pdf