The promise of large-scale learning assessments

Acknowledging limits to unlock opportunities

UNESCO Education Sector

Education is UNESCO's top priority because it is a basic human right and the foundation on which to build peace and drive sustainable development. UNESCO is the United Nations' specialized agency for education and the Education Sector provides global and regional leadership in education, strengthens national education systems and responds to contemporary global challenges through education with a special focus on gender equality and Africa.

The Global Education 2030 Agenda

UNESCO, as the United Nations' specialized agency for education, is entrusted to lead and coordinate the Education 2030 Agenda, which is part of a global movement to eradicate poverty through 17 Sustainable Development Goals by 2030. Education, essential to achieve all of these goals, has its own dedicated Goal 4, which aims to *"ensure inclusive and equitable quality education and promote lifelong learning opportunities for all."* The Education 2030 Framework for Action provides guidance for the implementation of this ambitious goal and commitments.

UNESCO **Education Sector**

United Nations
Educational, Scientific and
Cultural Organization

Education 2030

The promise of large-scale learning assessments

learning assessments

Acknowledging limits to unlock opportunities

Foreword

Standardized tests are typically associated with end-of-year school examinations for the certification and selection of students to progress to higher education levels. More recently, standardized test results have increasingly been used to gauge the performance of education systems as reflected in the growing prominence of large-scale learning assessments (LSLAs). Conducted at national or cross-national levels, such initiatives have not only drawn attention to issues related to levels of learning, but also to determinants of learning, including teacher qualification; the quality of school environments; parental support and guidance; as well as social and emotional health in and outside schools.

LSLAs have never been as much in demand. They are now regarded as being critical to improve the quality of education and they have contributed greatly to strengthening policies and strategies aimed at enhancing effective and relevant learning. Increasingly used to ensure excellence in education, they are at the centre of discussions related to the quality and effectiveness of education systems at national, regional and global levels. But our pursuit of *excellence* and *quality* in education must not come at the expense of *equity*. The 2030 Agenda is unambiguous in defining equity and inclusion as the central policy lens for educational development. It is a commitment to ensure that all children, youth and adults have the right to effective and relevant learning. While LSLAs are a central and an indispensable tool for progressing towards these new targets, under certain circumstances, they may unintentionally undermine the Education 2030 commitments made to ensure relevant, equitable and quality learning for all.

The Promise of Large-scale Learning Assessments: Acknowledging limits to unlock opportunities addresses the more contentious aspects of these standardized assessments. Drawing on UNESCO's extensive experience in the area – from involvement in the direct implementation of assessments and as a knowledge broker and convener of networks – this publication presents the Organization's critical take on such initiatives. It aims to balance the debate on LSLAs by reviewing their benefits while raising awareness on their potential risks and pitfalls. The focus of discussions in this publication is on LSLAs conducted in formal and school-based education.

It is hoped that a range of actors will benefit from the reflections in this piece. Particularly, governments who undertake or plan to conduct an LSLA; international and regional organizations who provide technical assistance in the planning and implementation of such initiatives; academics who critically analyze such processes; donor governments who are increasingly mandating them; and private businesses who may be looking to enter this niche market. The discussions in this publication are intended to provide greater insight and understanding about LSLAs by exploring a number of key questions: Why have LSLAs gained such prominence over the past decade? In what way can they support the 2030 commitments

to ensure inclusive and equitable quality learning for all? Or undermine them? Are they worth the time and resources? Do they really target all children and serve to promote equity? Do they assess relevant knowledge and skills, or the competencies needed to function in today's society?

The current growth in LSLAs will continue unabated – not only in numbers, but also in sophistication as they become more digitalized and adaptive. While this holds promise for the future, we must be wary of the potential risks involved. What will the future of assessment and the use of resulting data look like? How can we ensure that large-scale assessments do not continue to unintentionally constrain efforts to ensure quality learning for all, or worse still, exacerbate social disparities in learning? Looking ahead, we need to ensure that we fully understand the limits of LSLAs – both in their design and potential use – and apply these insights in the development of future assessments.

Stefania Giannini,
Assistant Director-General for Education, UNESCO

Acknowledgements

Table of contents

List of boxes, figures and tables

Acronyms

ACER	Australian Council for Educational Research
ADEA	Association for the Development of Education in Africa
ANLAS	Analysis of National Learning Assessment Systems
ASER	Annual Status of Education Report
CONFEMEN	*Conférence des ministres de l'Éducation des Etats et gouvernements de la Francophonie*
DFAT	Department of Foreign Affairs and Trade
DFID	Department for International Development
EGMA	Early Grade Mathematics Assessment
EGRA	Early Grade Reading Assessment
EQAP	Educational Quality and Assessment Programme
ERCE	*Estudio Regional Comparativo y Explicativo* [Regional Comparative and Explanatory Study]
ESD	Education for Sustainable Development
FFA	Education 2030 Framework for Action
GAML	Global Alliance to Monitor Learning
GDP	Gross Domestic Product
GEMR	Global Education Monitoring Report
GPE	Global Partnership for Education
HIV/AIDS	Human immunodeficiency virus and acquired immune deficiency syndrome
IBE	UNESCO International Bureau of Education
ICCS	International Civic and Citizenship Education Study
ICILS	International Computer and Information Literacy Study
IEA	International Association for the Evaluation of Educational Achievement
ICT	Information and Communications Technology
IIEP	UNESCO International Institute for Educational Planning
KIX	Knowledge and Innovation Exchange
LAMP	Literacy Assessment and Monitoring Programme
LaNA	Literacy and Numeracy Assessment
LLECE	*Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación* [Latin American Laboratory for Education Quality Assessment]
NALA	Network for African Learning Assessment
NAEP	National Assessment of Educational Progress
NEQMAP	Network on Education Quality Monitoring in the Asia-Pacific
NGO	Non-governmental organization
OECD	Organisation for Economic Co-operation and Development
OAA	Optimizing Assessment for All
PAL	People's Action for Learning
PASEC	Programme d'analyses des systèmes éducatifs de la CONFEMEN

PERCE .... *Primer Estudio Regional Comparativo y Explicativo* [First Regional Comparative and Explanatory Study]

PILNA .... Pacific Islands Literacy and Numeracy Assessment

PIRLS .... Progress in International Reading Literacy Study

PISA .... Programme for International Student Assessment

RTI .... Research Triangle Institute

SACMEQ .... Southern and Eastern Consortium for Monitoring Educational Quality

SEAMEO .... Southeast Asian Ministers of Education Organization

SERCE .... *Segundo Estudio Regional Comparativo y Explicativo* [Second Regional Comparative and Explanatory Study]

SDGs .... Sustainable Development Goals

SEA-PLM .... Southeast Asia Primary Learning Metrics

SIMCE .... Sistema de Medición de la Calidad de la Educación [Education Quality Measurement System]

TALENT .... Teaching and Learning: Educator's Network for Transformation

TERCE .... *Tercer Estudio Regional Comparativo y Explicativo* [Third Regional Comparative and Explanatory Study]

TIMSS .... Trends in International Mathematics and Science Study

UIS .... UNESCO Institute for Statistics

UN .... United Nations

UNESCO .... United Nations Educational, Scientific and Cultural Organization

UNICEF .... The United Nations Children's Fund

USAID .... United States Agency for International Development

Executive summary

Large-scale learning assessments (LSLAs) have been gaining prominence in the last decades. They are a form of national or cross-national standardized testing that provide a snapshot of learning achievement for a group of learners in a given year and in a limited number of learning domains.

The use of these assessments has been steadily increasing – they now reach more than half of the countries around the globe. They have also broadened in scope. Going beyond measuring reading and mathematics, they now increasingly target a greater number of domains, including digital skills, computer and information literacy, socio-emotional skills, or the understanding of concepts and issues related to civics and citizenship.

LSLAs have gained centrality in education debates at both local and global levels. This has been further encouraged by the new emphasis on equitable, effective and relevant learning for all inherent to the 2030 Agenda for Sustainable Development and the resulting focus on using data to improve policies and strategies. LSLAs are increasingly incorporated into the work programmes of international and regional organizations, and are supported by donor agencies through financial and technical assistance.

The benefits and advantages of LSLAs have been extensively reviewed in specialized literature, and there is growing recognition of the potential of assessment data to inform policy in a variety of ways. As a result, increasingly higher expectations have been placed on LSLAs as drivers of policy change over the past few decades. They are indeed expected to serve a range of uses, including monitoring, accountability, agenda-setting and analysis. The potential of such tools reaches far beyond reporting purposes. They can provide insight into areas in need of improvement and help pinpoint the most appropriate, promising and effective policy interventions.

However, both national and cross-national learning assessments have raised some concerns. A growing volume of evidence calls attention to a range of unexpected and even negative effects resulting from such exercises. Drawing on a diverse body of evidence, including scholarly literature and the experience of a range of international organizations, development partners and assessment specialists, this publication reflects on the possible unintended consequences of LSLAs. Some concerns stem directly from the characteristics inherent to their design while others centre around the (mis)uses of data to inform agenda-setting and policy formulation.

Valuing more than what can be measured: Limits in large-scale learning assessment design

A first set of unintended consequences of LSLAs stem directly from the way they are designed. Four main dimensions that are likely to be negatively affected are our conceptualization of education and learning; our understanding of education quality; the range of skills, knowledge and attitudes we value; and our attention to inclusiveness and respect for diversity.

Constraining the conceptualization of education and learning

Education is expected to serve multiple purposes in personal and social development for individuals and communities alike. This conceptualization of the integrated multiple purposes of education is well-captured in the four pillars of education put forth by the Delors Commission in 1996. It is arguably even more relevant today as a means to advance towards sustainable human and social development in a rapidly changing world. Yet, LSLAs generally focus on a limited range of learning dimensions and address a finite number of the multiple purposes of education. This is largely due to the design of LSLAs that limits measurable skills and competencies to a small number to allow for system-level comparability. While this is not inherently problematic, it is likely to end up constraining our understanding of what is valuable in education – particularly if LSLAs acquire a disproportionate prominence over alternative forms of assessments.

Narrowing our understanding of education quality

Learning outcomes as measured by LSLAs constitute one of the key indicators and dimensions of education quality. However, learning outcomes cannot be understood in isolation or as the only valid measure of education quality. Rather, they must be viewed in relation to other relevant dimensions, including societal outcomes and benefits required for social cohesion; procedural variables relative to the teaching-learning process; and classroom and school settings. However, progress in these dimensions is not necessarily measured by learning assessments. In fact, it remains unclear if this would be a possible or even desirable shift. If we are to capture the different dimensions of education quality, standardized tests must be understood in combination with a range of other qualitative and quantitative indicators.

Restricting what we value

The limited number of targeted domains measured can be considered an intrinsic feature of LSLAs. This has raised a number of concerns that such exercises cannot adequately assess the broad range of areas of competence cultivated through education. Even within the domains assessed, LSLAs are insufficiently attentive to the breadth of knowledge, skills, attitudes and values encompassed in education. While some organizations are exploring new strategies to capture the missing areas of competence, the viability and desirability of such measures again remains questionable. The neglect of certain knowledge areas could arguably be better addressed through the combination of LSLAs with other forms of evaluation so that learning outcomes do not become the single focus of attention.

Limiting who we value

LSLAs hold great promise as instruments to promote equity. However, the inclusiveness imperative is not always observed in the design and administration of LSLAs. The exclusion of vulnerable and marginalized children, youth and adults can occur during the sampling process, but also at the classroom or school level – as a consequence of the very design or formulation of the test items. While growing efforts are made to accommodate all students and appropriately capture their progress, this remains a challenge as special arrangements are likely to create both perverse incentives and comparability issues.

Beyond data as an end in itself:
Limits to the use of large-scale learning assessments

The limitations associated with LSLAs also relate to the use of their results or lack thereof. LSLA data serve various purposes, including monitoring trends; evaluating and orienting policies; setting agendas by raising awareness on certain issues; and holding various agents accountable.

Despite good intentions for the use of the data to improve educational processes, three main concerns arise: its under-use, over-use and the combination with (or subordinate to) accountability measures.

Under-using assessment data

LSLAs can only become a driver of change and improvement if concerned stakeholders act upon them. However, a range of factors constrain the effective use of assessment data to adequately inform policy debates and guide policy action. These include limited levels of institutionalization of assessment practices as well as limited ownership of funding, design, data management and dissemination. As administrating assessments require a lot of resources, making use of the resulting data is of paramount importance so the exercise is not futile or perceived as a waste of public resources.

Over-using assessment data

Excessive focus or credence given to assessment results can have a misleading and diverting effect on the priorities and behaviors of a wide range of stakeholders. This occurs in certain circumstances and through a number of processes. Firstly, the impact of assessment data on shaping policy will be squandered if the policy action is solely or mainly oriented at improving figures rather than engaging in substantive change. Secondly, the use of rankings within and between countries can encourage competition dynamics that, in turn, foster ill-informed policy borrowing practices driven by a desire to emulate rather than learn from certain countries. Thirdly, assessment data can encourage a careless use of causality language that may lead to ill-founded conclusions on the effects of learning. Finally, the incorporation of assessment

data in results-based funding schemes increasingly used by the donor community entails an important risk as it might inform strategies exclusively oriented to affect assessment results.

Combining assessment with accountability

Many of the risks associated with LSLAs are in fact the product of their use in combination with high-stakes accountability schemes. The unintended effects of such arrangements have been extensively reviewed in specialized literature. Some of the most frequently discussed pitfalls include teaching to the test dynamics and narrowing of the curriculum. Admittedly, the present understanding and recognition of the consequences of low-stake assessments is limited and warrants further investigation.

Overall, LSLAs hold great potential but harbor certain risks. While such limitations do not render assessments useless or inherently problematic, there is cause to exercise vigilance in the design of tools and use of the resulting evidence. Since LSLAs are increasingly a taken-for-granted feature in most education systems, it is important to identify the areas that could benefit from further enquiry, reflection and research. For instance, questions related to transparency and data ownership; the role of other modalities of assessment for diagnostic purposes; and the potential of emerging varieties of LSLAs.

Introduction

Education in the 2030 Agenda for Sustainable Development places new emphasis on equitable, effective and relevant learning for all. This central policy focus implies greater efforts to monitor the results of educational processes, which is consistent with the growing recognition of the potential of assessment data to serve a variety of policy-related purposes.

The advantages of large-scale learning assessments (LSLAs) have been well established by an expanding body of specialized literature. LSLA systems are a critical component of monitoring at the global, national or even local levels. The spread of LSLAs worldwide in the past two decades is a testimony to this crucial role. They have indeed contributed to the improvement of national policies and strategies aimed at ensuring effective and relevant learning for all. As they continue to serve a growing range of purposes, however, they also raise concerns on the potential for unintended consequences and the possible misuse of assessment data to shape education policy and practice. This publication contributes to the global debate by examining the evidence on the benefits, limits and potential side effects of LSLAs. It calls for greater caution in the design and implementation of LSLAs as well as the use of resulting data. These safeguards must be taken if we are to meet our collective commitment to ensure effective and relevant learning for all.

Focus questions

This publication addresses the following questions: What are LSLAs? What are the drivers of the expansion of such initiatives? How do LSLAs relate to the global 2030 commitments? And, most importantly, what are the potential unintended consequences in the design of LSLAs and use of the resulting data for policy and practice? Overall, the objective of the publication is to better understand how the design and use of LSLAs can be improved to ensure quality and equitable education. In short, we seek to comprehend how the potential of LSLAs can be unlocked.

Organization

Chapter 1

This chapter provides some conceptual clarification on the notion of LSLAs to adequately contextualize them within the broader scope of education monitoring. It contains an overview of the evolution of LSLAs over the past few decades and charts their broadening scope and rise in number, prominence and visibility. Possible rationales driving the expansion are also examined along with the evolving culture of evaluation and the shift in focus from access to learning in global education policy discourse.

Chapter 2

This chapter situates LSLAs against the backdrop of the principles and commitments that underpin the Education 2030 Framework for Action as well as against the broader global education policy discourse on learning. It discusses how the adoption of the Sustainable Development Goal on Education (SDG 4) has galvanized attention on learning outcomes and explores how LSLAs relate to and support the new mandate. Finally, the role of development partners in securing attention on assessment as a privileged means to foster progress towards the achievement of SDG 4 is examined.

Chapter 3

This chapter presents the first set of unintended consequences associated with the design and development of LSLAs. It considers four main dimensions that risk being negatively impacted as a result of the very design of LSLAs. These include our conceptualization of education and learning; our understanding of education quality; the range of skills, knowledges and attitudes we value; and our commitment to inclusiveness and respect for diversity.

Chapter 4

This chapter examines a second set of unintended consequences associated with the uses of LSLA data by a wide range of stakeholders. Three main groups of concerns are addressed. First, the under-use of assessment data with particular attention to the constraints that result from limited ownership. Second, the over-use of data – i.e. excessive attention or credence given to assessment results – that can provoke a misleading or diverting effect. Lastly, the combination of under- or over-use of data with (or subordinate to) high-stakes accountability frameworks is examined.

Overall, the publication aims at highlighting the tensions between the potential of LSLAs as tools at the service of quality, equity and accountability in education, and the potential negative impact they have in practice as a consequence of their design or unintended misuses. Ultimately, this tension calls for a careful and considered use of LSLAs to mitigate the potentially detrimental consequences.

1. The growth of large-scale learning assessments

This chapter provides an overview of the evolution of large-scale learning assessments (LSLAs) over the last decades with special attention to their growth in scope, number, prominence and visibility. First, some conceptual clarifications will be presented to appropriately contextualize LSLAs. Later, a review is given of the spread and evolution of both national and cross-national LSLAs. Finally, the chapter reflects on the possible factors driving the growing demand for and interest in such tools – both at the global and national levels.

Contextualizing the discussion on large-scale learning assessments

LSLAs have only gained prominence over the last decades in global, national and even local education debates. However, they need to be understood as a particular variety of assessments with well-defined characteristics that are integrated into larger, highly complex assessment systems. This first section provides conceptual clarification, defining LSLAs and situating them in a broader context.

For the sake of simplicity and clarity, this paper focuses on LSLAs administered to school-aged children – although, not necessarily in the school context. Most of the debates addressed in this publication are particularly relevant to primary and secondary education. It should be noted, however, that LSLAs that focus on other age groups are affected by a distinct set of challenges and limitations that, albeit interesting, lie beyond the scope of this paper.[1]

Assessment *of* learning or assessment *for* learning?

Assessment is central to any learning process. Pedagogues have often distinguished between assessment *for* learning and assessment *of* learning. Assessment *for* learning includes classroom or formative assessments used by teachers to adapt their teaching strategies or as means to provide individual grading to students at the end of a certain period of instruction. These individual assessments are intended to assess and monitor learner knowledge, skills, attitudes and values for a full set of academic domains. These include diagnostic assessments used to identify pupils who have learning difficulties and to tailor instruction and learning activities to their needs (Clarke, 2012). Assessment *of* learning, on the other hand, pertains to examinations or summative assessments used to certify or select learners in a given grade or age for further schooling, training or work. They are a form of standardized assessment that are consistent and in line with the curriculum – albeit generally focused on core academic subjects. **Table 1** provides an overview of these different assessment modalities. It should be noted that the distinction of assessment *for* learning and assessment *of* learning (and its equation to the

---

1   For instance, this includes the case of debates on the measurement of adult literacy (Addey, 2014) or on the implications of a lifelong learning perspective in terms of assessment.

formative vs. summative assessment distinction) is an analytical one and as such, does not hold up perfectly in practice – with a growing number of hybrid assessments combining both functions (Clarke, 2012).

Table 1. Overview of assessment modalities

Internal or school-based assessment	Formative classroom assessment	
	Summative classroom assessment	
External or standardized assessment	Central examinations (also known as exit, public or end-of-cycle examinations) – High stakes for students	
	Large-scale or system-level assessments (generally low-stakes for students)	Cross-national assessments
		National and sub-national assessments

Grey cells: Assessment for learning
White cells: Assessment of learning

Source: Based on Clarke, 2012; Fiske, 2000; OECD, 2013.

Learning assessment systems

Both assessments *of* learning and *for* learning are essential dimensions of learning assessment systems. Learning assessment systems consist of a set of policies, structures, practices and tools to generate data on learning outcomes (Clarke, 2012; UNESCO, 2017a). They comprise different types of assessments that yield information at individual, institutional or system levels for teachers, students, planners, policy-makers and other stakeholders. Such systems can help learners to learn, teachers to improve instruction, planners to decide on resource allocations and policy-makers and governments to address system inefficiencies or evaluate education programmes.

What are large-scale learning assessments?

LSLAs are a specific subset of learning assessment systems. They are system-level assessments that provide a snapshot of learning achievement for a given group of learners (based on age or grade) in a given year and in a limited number of domains. They are often categorized as national or cross-national (regional/international) assessments.[2] LSLAs are uniform and standardized in content, administration process, timing and scoring – they are, in fact, frequently referred to as *standardized tests* – particularly within Anglo-Saxon countries and literature. They are generally sample-based, although, over the last decades, an increasing number of countries have adopted a census-based approach (Verger et al., 2018). They can be school- or household-based; curriculum-based or not; and generally, teachers and schools may have a stake in the outcome while they hold low/no stakes for the test-takers.

> LSLAs are system-level assessments that provide a snapshot of learning achievement for a given group of learners in a given year, and in a limited number of domains. They are often categorized as national or cross-national assessments.

---

2   See Annex 1 for a list of cross-national assessments and more detailed information.

LSLAs are generally used by education authorities to determine learners' overall achievement levels. They help governments monitor changes in learning outcomes over time and highlight inequalities in learning achievement among population groups. By identifying correlates of learning outcomes and by providing a deeper insight on how a range of variables interact, LSLA data also provide a better understanding of the dynamics behind the performance of education systems. In doing so, they inform the design of policies and strategies aimed at improving student knowledge and competences as well as equity in learning.

Principles of rigor in design and implementation

One of the defining (albeit not exclusive) characteristics of LSLAs is the fact that they are designed and implemented to meet a certain level of standards.[3] Despite there being no vetted international standard for the characteristics that define robust LSLAs (i.e. that yield reliable data), there is substantial agreement among test developers, statisticians and psychometricians on the technical requirements of such assessments. LSLAs must be developed and implemented based on at least three principles.

First, LSLAs must be technically sound. In other words, the assessment methodologies, analysis and interpretation of data must follow scientific principles. For example, the assessment needs to include well-constructed contextual instruments; valid and reliable assessment tools; linguistic quality control for translation of instruments; rigorous scientific sampling procedures; and a selection of appropriate analytical methods (ACER, 2017).[4]

Second, LSLAs must follow standardized field operations. For example, the data needs to be collected under similar technical conditions in an efficient and secure manner – independent from the administration context. All personnel involved should receive adequate training and instruments must be administered according to established procedures.

Third, LSLAs must be designed to be ethical, fair and inclusive of the target population. Fairness and ethical principles must be applied to each one of the assessment stages, including:
- Development - e.g. an effort should be made to minimize measurement bias;[5]
- Implementation - e.g. the well-being of participants during the assessments should be taken into account when devising the length of the assessment, for instance;
- Analysis and dissemination stages - e.g. the confidentiality of participants should always be ensured by anonymizing data prior to analysis and public release.

In addition, the assessments should be inclusive. They should be designed to be relevant for as many members of the target population as possible and designed to respond effectively to all

---

3   While standardized design and administration constitute one of the most defining features of LSLAs, it should be noted that such characteristics are not exclusive to this modality of assessment. For instance, standardization is indeed also a key feature of most national examinations.

4   Currently, there is no international standard for characteristics of a 'technically sound' learning assessment. However, The Principles of Good Practice in Learning Assessment prepared by ACER in collaboration with the UNESCO Institute for Statistics and the Global Alliance to Monitor Learning (ACER-UIS, 2017) provides an independent articulation of good practices that accommodates the diversity of LSLA activities undertaken around the world. See http://uis.unesco.org/sites/default/files/documents/principles-good-practice-learning-assessments-2017-en.pdf for more information.

5   'Measurement bias' is where a test or contextual instrument consistently discriminates against a particular group for reasons unrelated to the learning domain assessed or the background data collected.

learners (UNESCO, 2017b). Thus, the sample should capture a representative pool of learners regardless of their particular attributes, including their background (socio-economic or other), geographical location or educational performance. Likewise, these principles of equity and inclusion have implications for example, on the translation and adaptation of test materials to suit different population groups or to accommodate children with special needs.

The promises of LSLAs

The multiple benefits and advantages of LSLAs have been extensively discussed in specialized literature. There is indeed growing recognition of the potential of assessment data to serve a variety of policy-related purposes. It has thus been noted that although assessment is part of the monitoring and evaluation phase of the policy cycle, their uses extend far beyond this stage. As argued by Best et al. (2013), LSLAs also provide key evidence necessary to inform agenda-setting (i.e. the identification of priority issues), orient policy-formulation and guide policy implementation. Similarly, Tobin et al. (2015) document how LSLA data

> There is growing recognition of the potential of assessment data for a variety of policy-related purposes. It has been noted that LSLAs provide key evidence necessary to also inform agenda-setting and the identification of priority issues, orient policy-formulation and guide policy implementation.

impact a range of education policies – most notably, system-level policies, including curricular reforms and the establishment of performance standards; resource-allocation policies; and teaching and learning policies affecting school- and classroom-level practices.[6] As noted by these authors, LSLAs can serve a variety of agendas, including promoting quality, equity and accountability. The potential of LSLAs has also been increasingly highlighted by specialized education agencies that have made the case for sustained investment in such arrangements (e.g. UIS, 2018a or The World Bank, 2018).

Overall, this body of literature makes it clear that the benefits of LSLAs are manifold and that they should be a central feature of any education system. The absence of learning data poses a variety of risks that challenge progress towards greater equity and quality. However, efforts to identify and synthesize limitations, challenges and risks associated with LSLAs are scarcer and remain at the developmental stage. The objective of the discourse in this publication is precisely to delve into such limitations in a systematic manner to identify the ideal conditions in which LSLAs are more likely to live up to their promise and have an impact on education quality and equity.

A phenomenon on the rise

In the last two decades, LSLAs, both national and cross-national, have increasingly gained currency worldwide, expanding well beyond more affluent countries. Indeed, a growing number of middle- and lower-income countries are investing in the implementation of national assessment systems and are participating in cross-national assessments (UNESCO, 2015b).

---

6   In fact, a number of international large-scale assessments provide individual or cross-country analysis on certain subjects. See Annex 1 in this report for a list of links to websites containing consolidated research and analysis curated by the cross-national assessments in place.

Between 2013 and 2017, nationally-representative assessments (including both cross-national and national assessments) were administered in at least in 177 countries at the primary-education level and at least in 150 countries at the lower-secondary education level (UIS, 2019a). This section provides an overview of the spread and evolution of both varieties of assessment.

National assessments

Unlike examinations that are used to certify student achievement and select students for further study, the main purpose of national assessments is to provide feedback to policy-makers at national and sub-national levels on the learning achievement of a given target population (Greeney, 2008; Clarke, 2012). National assessments are generally confined to a few subjects (Greeney, 2008) – mainly reading, mathematics and sometimes, sciences, social sciences and foreign language (Benavot and Köseleci, 2015).[7] Such assessments have spread from 70 countries in 1990-1999 to 135 countries in 2013 (Benavot and Köseleci, 2015) – most of them emerging in developing countries. These include the applications of the citizen-led initiatives and Early Grade Reading Assessments (EGRA) in Grades 2 and 3,[8] which have been administered in more than 70 developing countries in over 100 languages. In approximately a dozen countries, these have expanded from small sample surveys to nationally representative assessments (UIS, 2019a).[9] Although at the time of writing it is not possible to provide an exact figure on the number of national assessments in place, extensive efforts are currently underway to map out country-level initiatives – with the UIS Database of Learning Assessments[10] featuring as the most comprehensive effort to date.

Cross-national assessments

Similarly, the number of countries[11] participating in cross-national assessments has almost doubled since the first rounds (1999-2000) with 137 countries engaging in one or more cross-national assessment, according to recent estimates (UIS, 2018b).[12] For instance, the number of countries participating in the Trends in International Mathematics and Science Study (TIMSS) shifted from 42 in 1995 to 64 in 2019; and in the case of the Programme for International Student Assessment (PISA), the number of participants rose from 42 in 2000/2001 to 79 in 2018 – along with seven additional countries engaging in PISA for Development (Raudonyte, 2019). It is also worth noting that while most participating countries (60%) are still from Europe and North America (UIS, 2019a), participation has expanded over the years to all regions of the world. New waves of countries from the global South are joining upcoming cycles of cross-national assessments and new regional assessments have been developed. In sub-Saharan

---

7   Some of the most developed and well-established national learning assessments include the USA's National Assessment of Educational Progress (NAEP) and Chile's SIMCE (Education Quality Measurement System or *Sistema de Medición de la Calidad de la Educación*).

8   Although Early Grade Reading Assessments are national and often sub-national in scope (as they are country-specific instruments and not intended for cross-country comparability purposes), the fact that they rely on a common methodological framework has led some analysts to classify them as international or hybrid assessments (see Wagner, 2011; Raudonyte, 2019).

9   For more information on citizen-led and early grade assessments, consult Annex 2 in this report and the Global Reading Network website: https://globalreadingnetwork.net/

10   http://uis.unesco.org/en/uis-learning-outcomes

11   It should be noted that, while cross-national assessments generally compare different countries, sub-national regions can also constitute a differentiated unit of analysis, particularly in the case of federal systems or local governments with high levels of administrative autonomy.

12   See Annex 1 for a list of cross-national assessments and more detailed information.

Africa, for instance, the Democratic Republic of the Congo, Gabon, Guinea, Madagascar and Mali will join Programme d'analyses des systèmes éducatifs de la CONFEMEN (PASEC) in 2019 while in Latin America, Bolivia, Cuba and El Salvador will rejoin the fourth edition of the Regional Comparative and Explanatory Study (ERCE).[13] In Asia and the Pacific, the Southeast Asia Primary Learning Metrics (SEA-PLM) assessment will be conducted for the first time in 2019 in 11 countries in the region.

Initially designed to assess foundational skills among children of primary and lower-secondary school age, cross-national learning assessments have expanded to reach young children as well as youth and adult populations in a wider range of learning domains. These include computer and information literacy; the understanding of concepts and issues related to citizenship; and civic attitudes and engagement.

Opting out, but not for long

While participation in cross-national learning assessments is on the rise, it is interesting to note that some countries are opting out for political, financial or technical reasons (Addey and Sellar, 2019).[14] El Salvador, for instance, ceased participation in all international assessments in the mid-2000s, including in the TIMSS, Latin American Laboratory for Education Quality Assessment (LLECE) and the Literacy Assessment and Monitoring Programme (LAMP) after the newly elected government put a halt to all system-level assessment activities. Argentina opted out of PIRLS after the first round in 2001. These changes might have come about because certain countries are more interested in comparing themselves to their regional peers – as documented among Latin American countries by Ferrer and Fiszbein (2015). It could also be argued that PIRLS did not hold any added-value as the existing regional assessment (LLECE) also measures the reading skills of children in the lower grades of primary education.[15] These trends suggest that participation in cross-national assessments is motivated or driven by both instrumental and symbolic reasons.

Some countries have also dropped their national learning assessments. This is the case in South Africa, which had conducted the Annual Learning Assessment for all students Grades 1-9. The assessment was linked to teacher and school accountability, so the controversial decision was made to discontinue the assessment. Uganda, on the other hand, has had to change its approach to yearly assessments due to funding issues and has opted for implementation of such testing every three years instead. However, these cases remain the exception rather than the rule. The global trend for LSLAs is an upswing in number and a broadening in scope.

---

13 At the time of publication, it was still too early to identify new countries joining the Progress in International Reading Literacy Study (PIRLS) 2021 and PISA 2021.

14 On a different scale, opt-out movements in the form of grassroots, citizen-led initiatives have also been observed, mainly in the United States. Opt-out movements can be characterized as collective efforts encouraging the refusal to take standardized tests. These have generally been led by parents of school-aged children (see Mitra et al., 2016; or Pizmony-Levy and Saraisky, 2016). This discussion, however, lies beyond the scope of this publication.

15 It is, however, important to note that Latin American countries are interested in comparing their secondary-school population to highly industrialized countries through assessments, such as PISA that targets 15-year-olds (Ferrer and Fiszbein, 2015).

In fact, it appears that some countries that have dropped out are likely to rejoin upcoming cycles of cross-national initiatives. Austria, Bosnia and Herzegovina, and the Philippines will be rejoining TIMSS 2019 after having ceased participation in the assessment for several years. In Latin America, Bolivia, Cuba and El Salvador will join ERCE 2019 after having opted out of the assessment for one or more rounds. Algeria had opted out of cross-national assessments since its participation in TIMSS 2011 but joined PISA in 2015.

What drives the rise in large-scale learning assessments?

The rationales for developing a national learning assessment or participating in cross-national initiatives are multiple and driven by a wide range of often interconnected factors. Therefore, engagement in LSLAs cannot be assumed to be a simple function of wealth, development status or to serve a particular role within the international relations realm. This section reflects on four main factors behind the rise and spread of LSLAs – a growing number of perceived benefits, an evolving global culture of evaluation, a shift in the focus of global education policy, and the priorities and demands of development donors.

What are the perceived benefits?

While some LSLAs are mandated by bilateral and multilateral donors (Benavot and Köseleci, 2015) and/or driven by concerns for results and accountability (Kamens and McNeely, 2010), participation is also driven by rationales related to perceived benefits of the assessment process (see **Box 1**). This is particularly true for cross-national initiatives. These benefits include developing positive international relations, improving economic outcomes as well as informing curriculum and pedagogy. Less explicitly acknowledged perceived benefits are of a more symbolic nature, such as participating in a "global ritual of belonging," emulating so-called "reference societies," and pushing the personal political or career agendas of national policy-makers (Addey, 2014; Addey et al., 2017).[16] While the rationales for engagement vary, the common thread seems to be that fewer and fewer countries imagine achieving the status of a "good society" without both higher levels of educational attainment and accompanying efforts to monitor learning outcomes (Kamens and McNeely, 2010).

> The rationales for developing a national learning assessment or participating in cross-national initiatives are multiple, and driven by a wide range of often interconnected factors, and cannot therefore be assumed to be a simple function of wealth, development status, or a particular role within the international relations realm.

---

16 In this sense, the rise of LSLAs cannot be dissociated from broader changes and, most notably, the emergence of new hegemonic policy paradigms as New Public Management (Addey et al., 2017). The policy models put forward by this particular approach to public service emphasize the evaluative function of the State (in education or other areas). As countries face increasing levels of pressure to reform their education systems according to these principles, the adoption of LSLAs is perceived as a relatively inexpensive and (politically) rewarding way of doing so (Verger et al., 2018).

Box 1. Perceived benefits associated with participation in cross-national initiatives

- **Providing access to networks** at regional and international levels, both for researchers and policymakers;

- **Provoking dialogue** among different groups and allowing for catalyzing debate on education;

- **Producing evidence** for educational phenomena and advancing the use of technical and complex education indicators;

- **Motivating regulatory and behavioral policy reforms**, including on teaching, learning and assessment;

- **Attracting media attention and increasing transparency** regarding education system outcomes and human capital development in national and cross-national contexts;

- **Developing capacities of professionals** who participate in the assessment – from a technical and operational perspective.

Source: Based on Addey, 2014; Addey et al., 2017; Diaz and Flores, 2008; Fenwick and Edwards, 2014; Lockheed et al, 2015; Wagner et al., 2012; OECD, 2016b.

An evolving culture of evaluation

The growth in LSLAs is also associated with a change in assessment culture. Over time, the OECD and International Association for the Evaluation of Educational Achievement-led (IEA-led) assessments have shifted from what were initially research studies focused on in-depth singular education systems to cross-national initiatives with policy-oriented rationales linked to accountability and competition (Pizmony-Levy, 2013; Addey, 2018). This shift in assessment culture is also observed elsewhere. The PASEC assessment is one example where the approach shifted from individual national assessments to a comparative study. Between 1991 and 2012, PASEC conducted individual national assessments in 24 French-speaking countries and territories in Africa, the Middle East and East Asia. Until then, the data were not comparable across countries. It was not until 2014 that PASEC changed its approach to conduct its first cross-national assessment in ten countries. This evolution in assessment cultures is a result of intensifying globalization and reflects the growing interest in global mandates (Greeney, 2008) and the increasing emphasis on LSLAs in global education discourse.

A shift in focus from *access* to *learning* in global education policy discourse

Global education discourse – particularly in the years leading up to the adoption of the 2030 Agenda – shifted from a focus on access to schooling to a focus on learning. There has indeed been an overall shift in global education discourse from a concern with education inputs, such as pupil enrolment rates, school facilities, classroom materials and teacher education and training, to a concern with education outcomes – i.e. the knowledge, skills, attitudes and values that pupils acquire as a result of education. This is visible in strategy

> Equity and quality have now taken center stage in global education debates and LSLAs have come to be seen as tools to support countries to achieve quality and equity in education.

documents recently produced by key international development partners (see **Table 2**), which together with others, advocate for a paradigm shift from ensuring equitable access to relevant

and effective learning for all (UNESCO-IBE, 2013; Barrett, 2016). Equity and quality have now taken centre stage in global education debates and LSLAs have come to be seen as tools to support countries as they strive to achieve quality and equity in education.[17]

Negotiations around the Education 2030 agenda played a crucial role in securing and consolidating the interlinking of education quality, learning outcomes and LSLAs. This was particularly due to the selection of target indicators. To be sure, various reports and declarations issued during the lead-up to the adoption of the new agenda emphasized a different understanding of quality. While some of them chose a narrow approach to quality by equating it to numeracy and literacy outcomes, the understanding that finally prevailed was notably broad and comprehensive in scope. However, the translation of the final list of targets into indicators entailed a relative narrowing of the original ambition by emphasizing the measurement of proficiency in reading and mathematics as an adequate proxy (King, 2017; Unterhalter, 2019).

Table 2. **Shift to a focus on learning – selected publications**

Publication/Initiative	Organization	Focus
Learning for All: Investing in People's Knowledge and Skills to Promote Development. World Bank 2020 Education Strategy (2011)	World Bank Group	Ensuring that all children and youth should not only attend school but also acquire the knowledge and skills they need to lead healthy, productive lives and secure meaningful employment.
LEARNING to Realize Education's Promise. The World Development Report 2018 (2018)	World Bank Group	Learning outcomes
Global Education First Initiative (2012-2016)	UN Secretary-General's Global Initiative on Education	Every child in school; Quality of learning; Global citizenship.
Learning Metrics Task Force. Access plus learning (2012-2016)	UNESCO Institute for Statistics and the Center for Universal Education at Brookings	Strengthening assessment systems and the use of assessment data.
The Learning Generation. Investing in education for a changing world (2016)	The Education Commission (The International Commission on Financing Global Education Opportunity)	Mobilizing and deploying resources to ensure that all children and youth have the opportunity to gain the education and skills they will need to become productive and successful adults.

Note: This selection of initiatives is not exhaustive. It is intended to exemplify the focus on learning across various initiatives, financing mechanisms and organizational strategies.

17 Although LSLAs have been at the centre of international debates, it should be noted that the attention and conceptualization of these tools as key means to monitor and encourage education progress dates back at least to the early days of the Education for All agenda. The report *Assessing Learning Achievement*, published in 2000 as part of the *Status and Trends* series (Fiske, 2000), already inquired into the key role of the assessment of student outcomes as a tool to promote quality education for all. In this sense, it is worth noting that the report fleshes out and reflects on most of the questions along which current debates are still structured.

An important focus on producing and using data

Today, multilateral development partners and donor agencies are increasingly mandating some form of learning assessment to accompany their loans and aid (Kamens and McNeely, 2010). It is common for donor agencies to use results from LSLAs to evaluate the effectiveness of education policies and practices. Assessment data can help monitor the evolution of learning gaps across and between groups based on factors such as gender, income status, residence and home language. The Global Partnership of Education's strategic plan for 2020, for example, introduced the use of results-based financing to help governments create robust assessment systems and encourages them to undertake all forms of learning assessment to monitor learning outcomes (GPE, 2016). To access GPE funding, all grantees must demonstrate that they have good quality data systems or a strategy for building them (GPE, 2018). Similarly, the World Bank's Learning for All Education Strategy supports government efforts to increase the availability of data on learning and skills, and to develop a culture of results monitoring and assessment (World Bank, 2011). The OECD's Learning Framework 2030 (OECD, 2018a) and UNESCO's Education Strategy (2014-2021) also include an important focus on the assessment of learning as a means to monitor foundational and transferable skills as well as using the data to make informed decisions to improve teaching and learning.

A similar prioritization is observed in development strategies among some of the main bilateral donors in education. For instance, in its new Education Strategy, DFID UK highlights its commitment to work with decision-makers to improve the quality and use of education data, prioritizing good national and cross-national measures of learning (DFID UK, 2018). The Australian Department of Foreign Affairs and Trade Education strategy (2015-2020) also includes an important focus on improving learning assessment systems and the use of evidence to inform good policy and practice (DFAT, 2015). likewise, the governments of Canada, Finland, Norway, Sweden and the USA make specific references to the importance of learning assessments as essential tools for monitoring progress in learning (NORAD, 2018; SIDA, 2018; DFAT Canada, 2014; USAID, 2011; Ministry of Education, Finland, 2006).

2. Learning assessment in the 2030 agenda

While both large-scale learning assessments (LSLAs) and learning outcomes have recently gained traction within the development community, the adoption of SDG 4 has galvanized attention on these education components in an unprecedented way. This chapter reflects on how LSLAs relate to the Education 2030 agenda, and how a variety of development partners are encouraging a greater global focus on learning data as a means to foster progress towards the achievement of SDG 4.

A renewed mandate for learning data

The Education 2030 agenda has been instrumental in consolidating the attention and centrality of assessment and learning outcomes – thus contributing to posit LSLAs as indispensable to monitoring learning and progress towards the new targets. This section reflects on the centrality of LSLAs in advancing this new vision of education – not only as tools to monitor and report but to orient and encourage policy action as well to ultimately enable the realization of SDG 4.

The Education 2030 vision

How do LSLAs relate to Education 2030? As outlined in the 2015 Incheon Declaration, Education 2030 is inspired by a humanistic vision of education and development based on principles of respect for human rights and human dignity; social justice; inclusion; protection; cultural, linguistic and ethnic diversity; as well as shared responsibility and accountability. It reaffirms the principle of education as a public good, a fundamental human right and a basis for guaranteeing the realization of other rights. It positions education as essential to human fulfilment and sustainable development within a framework where economic growth is guided by a central concern for social justice and environmental stewardship. Education in the 2030 Agenda for Sustainable Development has refocused attention on equity and inclusion, as well as on the quality and relevance of education within a lifelong learning approach. This is essentially articulated in Sustainable Development Goal 4 (SDG 4), which aims to "ensure inclusive and equitable quality education and promote lifelong learning opportunities for all." (UNESCO, 2016c, p. 7).

> The Education 2030 agenda has been instrumental in consolidating the attention and centrality of assessment and learning outcomes – thus contributing to posit LSLAs as indispensable to monitoring learning and progress towards the new targets.

SDG 4 focus on the results and outcomes of education

SDG 4 centres on the results and outcomes of educational processes. It focuses on both effective acquisition of relevant knowledge, foundational and transferable skills and competencies as the basis for learning throughout life as well as on the relevance of learning for the world of work and for civic, social and cultural life (UNESCO, 2016a). This is clearly reflected in the outcomes-based global targets and indicators proposed for monitoring progress on SDG 4 (see **Table 3**).

Given the centrality of learning outcomes in the new agenda, learning data is increasingly viewed as an essential component of monitoring mechanisms to measure progress towards SDG 4 – at the global, regional and national level.

Table 3. Learning-related SDG 4 targets and indicators

Target 4.1. By 2030, ensure that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes.	
Indicator 4.1.1	Proportion of children and young people (a) in Grade 2 or 3; (b) at the end of primary education; and (c) at the end of lower secondary education achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex.
Target 4.2. By 2030, ensure that all girls and boys have access to quality early childhood development, care and pre-primary education so that they are ready for primary education.	
Indicator 4.2.1	Proportion of children under 5 years of age who are developmentally on track in health, learning and psychosocial well-being, by sex.
Target 4.4. By 2030, substantially increase the number of youth and adults who have relevant skills, including technical and vocational skills, for employment, decent jobs and entrepreneurship.	
Indicators 4.4.1	Proportion of youth and adults with information and communications technology (ICT) skills, by type of skill.
4.4.2	Percentage of youth/adults who have achieved at least a minimum level of proficiency in digital literacy skills.
Target 4.6. By 2030, ensure that all youth and a substantial proportion of adults, both men and women, achieve literacy and numeracy.	
Indicator 4.6.1	Proportion of population in a given age group achieving at least a fixed level of proficiency in functional (a) literacy and (b) numeracy skills, by sex.
Target 4.7. By 2030, ensure all learners acquire knowledge and skills needed to promote sustainable development, including among others through education for sustainable development and sustainable lifestyles, human rights, gender equality, promotion of a culture of peace and non-violence, global citizenship, and appreciation of cultural diversity and of culture's contribution to sustainable development.	
Indicators 4.7.4	Percentage of students by age group (or education level) showing adequate understanding of issues relating to global citizenship and sustainability.
4.7.5	Percentage of 15-year-old students showing proficiency in knowledge of environmental science and geoscience.

Source: UNESCO 2016a.
Note: A full list of the ten SDG 4 targets and their corresponding 43 indicators is available in Unpacking Sustainable Development Goal 4.[18] Data on the 43 thematic indicators is available in the UIS database http://data.uis.unesco.org/.

---

18   See *Unpacking SDG 4:* http://unesdoc.unesco.org/images/0024/002463/246300E.pdf

An indispensable tool for monitoring SDG 4 commitments

As a number of the global SDG 4 targets and indicators are outcomes-based, LSLAs have come to be regarded as an important means of providing timely and reliable evidence to monitor learning-related commitments (UIS, 2018c). In fact, the different outcome-related targets described in Table 3 are expected to be monitored through some form of system-level assessment of learning achievement. As a result, LSLAs are considered indispensable tools to monitor learning – at least until new methodologies incorporating other data sources (e.g. administration data, national examinations, censuses or household surveys) and estimates and/or projections are developed.

Beyond global education targets

Learning metrics and LSLAs are central to the Education 2030 agenda as they enable reporting on a wide range of outcome-related targets but their potential goes well beyond this purpose.[19] Learning metrics also provide crucial evidence on the determinants of student achievement as well as on the impact and efficiency of policy initiatives aiming to support and improve it.

LSLAs can be used to monitor progress over time. They also contain a set of background questionnaires for students, teachers, principals and/or parents to capture information on socio-economic and demographic conditions as well as practices or experiences that influence learning outcomes (see **Table 4**). Moreover, some assessments generate important information on the education system. This includes information not generally available in international education databases, such as on types of early childhood education; national policies on school entry, promotion and examinations; and instructional time devoted to a given subject matter.

This wealth of information allows countries to gauge their educational health, identify correlates of outcomes (e.g. teacher qualifications, educational resources, access) and understand how a range of variables interact to provide a clearer picture of the dynamics in education systems. The data collectively provide insights into areas related to student achievement and its association to health, equity, quality of school inputs, levels of student support at home, or the type and location of schools. In this sense, LSLAs play a key role in the measurement of equity – one of the key principles orienting the Education 2030 agenda.

While cross-national learning assessments tend to include a variety of contextual questionnaires on home, community, school and student factors, this is not often the case for national learning assessments. Most national learning assessments are accompanied by a student questionnaire (8 in 10), and half by a school or teacher questionnaire yet only 20% of them collect information on home practices, environment or parental support – all of which are important determinants of learning achievement (UIS, 2019a). There are also other factors that neither assessment captures. Information on student health and disability, ethnic or minority status is often strictly regulated and not easily obtained through national or cross-national learning assessments.

---

19 The activities of different organizations charged with the production of assessment data is reflective of such potential. For instance, UNESCO and IEA prepared a booklet for the release of the PIRLS 2016 results to further highlight how the LSLA can be used to monitor progress towards SDG 4 by capturing learning achievement data and more.

Table 4. Contextual information collected alongside learning achievement results

Student questionnaire	Teacher questionnaire	Parent questionnaire	School questionnaire (completed by principal)
• Gender, age, language spoken at home • Household posessions [e.g. study desk, internet, bedroom, number and types of books] • Home practice and parental support [e.g. use of computer, reading for pleasure, activities with parents] • Health and wellness [absence from school, feelings about school, proper nutrition at home] • Their school [bullying, enjoyment of the different lessons, impressions about teachers, use of library] • Self-reported measures on reading and mathematics proficiency.	• Gender, age, educational attainment, main areas of study and field of specialization • Teaching experience and formal professional development • Attitudes and impressions about the school, its resources, environment and factors linked to academic success • Reflections on behaviours/practices/resources hindering learning in the classroom • Teaching practices and assessment • Family involvement in the school.	• Type and frequency of activities with children [reading, singing, counting, playing, etc] • Migration status and language(s) spoken at home • Enolment of their children in early childhood education programmes and duration • Age of entry of their children into primary school • Impressions on their children's reading, math or science levels • Practices of studying outside school • Impressions on the child's school • Parental literacy practices, highest level of educational attainment, type of employment.	• Principal experience and education • Number of students enrolled and distribution by socio-economic background • geographical location of school • Condition of school infrastructure • Provision of meals • Instructional time • Resources and technology [e.g. library, computer labs, supplies, accomodations to students with disabilities] • School discipline and safety • Parental and community contribution to the school.

Source: Authors. For more examples of context questionnaires see Cresswell et al., 2015; OECD, 2017; UIS, 2016b; IEA, 2015; IEA, 2016; UIS, 2016b.
Note: No questionnaire contains all the elements described in this table. It is intended to exemplify the wealth of information that can be collected and analyzed alongside learning achievement levels based on the accompanying context questionnaires.

To monitor policy commitments implemented over time, accompanying questionnaires and other supporting data on the education system also help capture policy-relevant information. Details that can be captured include the impact of specific language of instruction policies, how teacher policies are changing and the relative emphasis of subdomains assessed in the curriculum. In sum, background questionnaires have the potential not only to provide valuable information on the learning outcomes of specific population groups, but also on the inputs related to such factors as teachers, classrooms, parental support and school resources that contribute to this learning. This wealth of information, in turn, plays a key role in guiding and orienting policy action at multiple levels (global, national, local) – ultimately enabling the achievement of SDG 4.

Nurturing the new global education agenda

Initiatives to support the measurement of learning outcomes have flourished over the last decades. This section presents an overview of the range of projects and programmes that have recently been implemented at different levels to support the advancement and consolidation of LSLAs.

Supporting LSLAs: global and regional initiatives

As previously mentioned, LSLAs are fast becoming indispensable tools to monitor learning-related outcomes in the 2030 agenda. This move has been supported by donor governments, international and regional organizations, development partners and other private actors who recognize the enormous value of these instruments for educational improvement. These organizations also recognize the importance of assessment data in building national education development strategies and the need for greater national, regional and global efforts to overcome the large data gaps in learning achievement levels observed across the world (UIS, 2018a).

International and regional organizations are advocating for LSLAs to help drive system improvement. Well-established initiatives include, for instance, the regional assessment program led by the Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (Latin American Laboratory for Assessment of the Quality of Education, LLECE) at the UNESCO Regional Bureau for Education in Latin America and the Caribbean. Since 1997, up to three regional assessments have been conducted in the region, covering the mathematics, science and reading domains. More recently, considerable efforts are being made to help countries generate evidence and use it in areas critical to improving learning processes and outcomes (see **Box 2**). These new initiatives and approaches represent great strides in how we think about using learning data and evidence to improve education systems.

> International and regional organizations have been advocating for the importance of LSLAs for system improvement, support[ing] countries to generate evidence and using it in areas which are critical to improving learning processes and outcomes.

Over the past few years, international organizations have grown increasingly data-focused, supporting the creation of robust data collection systems. Even before the adoption of the 2030 Agenda, initiatives supporting the measurement of learning outcomes mushroomed with diverse options for establishing global metrics for reading and numeracy. These included the Learning Metrics Task Force (2012-2016),[20] the Learning Metrics Partnership (2014)[21] and the Center for Global Development Study Group on Measuring Learning Outcomes (2013).[22] In 2016, the Global Alliance to Monitor Learning (GAML), led by the UNESCO Institute for Statistics (UIS), was launched to consolidate different processes. The GAML provides a space for dialogue on political perspectives and technical expertise in the establishment of strategies,

---

20 For more information on the Learning Metrics Task Force: https://www.brookings.edu/product/learning-metrics-task-force/
21 For more information on the Learning Metrics Partnership: http://uis.unesco.org/sites/default/files/documents/learning-metrics-partnership-a-capacity-support-and-policy-strengthening-initiative-to-develop-and-use-common-learning-metrics-mathematics-reading-2014-en_1.pdf
22 For more information on the Study Group and its output: https://www.cgdev.org/publication/ft/schooling-not-education-using-assessment-change-politics-non-learning

methodologies and tools to monitor learning outcomes at the global level (see **Box 3**). The establishment of the GAML marks a major step towards greater synergy by fostering collaboration across a range of stakeholders, many of which compete for a share of the learning outcomes measurement "market."

Box 2. Recent regional and global initiatives promoting large-scale learning assessments

TECHNICAL SUPPORT

- UNESCO-IIEP is expanding its support to Member States (mainly under the form of research and training) to encourage the use of learning assessment data in the planning and management of education systems (UNESCO-IIEP, 2017).

- The Center for Universal Education at the Brookings Institute launches the Optimizing Assessment for All (OAA) initiative to support countries to improve assessment, teaching and learning of 21st century skills, in collaboration with the NEQMAP initiative at UNESCO Bangkok and the TALENT initiative at UNESCO Dakar (more on these projects below).[23]

- The Association for the Development of Education in Africa (ADEA) established the Network for African Learning Assessment (NALA) in 2016 as a research programme seeking to promote SDG 4-related assessments with an emphasis on the assessment of non-cognitive skills.

- The Analysis of National Learning Assessment Systems (ANLAS) developed by ACER with the support of GPE aims to help countries undertake a comprehensive analysis of their assessment systems.

NETWORKS AND KNOWLEDGE SHARING

- UNESCO's Network Education Quality Monitoring in the Asia Pacific (NEQMAP) aims to improve the capacity of Member States to use student learning assessments to strengthen education systems.

- In sub-Saharan Africa, UNESCO's Teaching and Learning: Educator's Network for Transformation (TALENT) also works directly with Member States to strengthen areas critical to improving learning – curricula, teacher policy and training, and assessment of learning outcomes.

- In 2016, UNESCO-IIEP launched the Learning Portal to provide up-to-date, relevant and neutral information on learning issues from primary through secondary education.

FINANCING

- The Global Partnership for Education's (GPE's) Assessment for Learning initiative works at national levels to ensure that investments in learning assessment systems are based on a needs analysis, linked to policy processes and mainstreamed into education sector plans. Recently, GPE has launched the Knowledge and Innovation Exchange (KIX) mechanism, which also includes a specific focus on supporting learning assessment systems. The KIX mechanism builds on the Assessment for Learning (A4L) initiative launched in 2017, oriented at capacity-building and supporting countries to strengthen their learning assessment systems.

Note: For more information on UNESCO's activities in the area of learning assessments, consult the brochure *Learning assessment at UNESCO: Ensuring effective and relevant learning for all* (UNESCO, 2017c).

---

23   For more information on the OAA initiative, consult https://www.brookings.edu/wp-content/uploads/2017/12/overview_optimizing-assessment-for-all.pdf

---

Box 3. The Global Alliance to Monitor Learning

The Global Alliance to Monitor Learning (GAML) is a multi-stakeholder initiative led by the UIS to support national strategies to measure learning and to enable international reporting on the SDG 4 indicator framework. The GAML brings together UN Member States, international technical experts, implementation partners (Member States, donors, civil society, UN agencies and the private sector) to improve learning assessment globally. The GAML's work is organized around task teams that tackle the issues related to globally comparable data on learning outcomes.

The GAML is building global consensus on the international reporting and measurement of learning based on the best fit-to-purpose methodological approaches and practices. It is developing: (1) all of the SDG 4 indicators related to learning and skills as well as the methodological tools and standards needed to ensure global comparability; (2) standards, guidelines and tools to help countries strengthen the implementation of their assessments and evaluate their quality; and (3) capacity development tools to help countries collect, analyze and use learning assessment data. The methodological work of the GAML informs other global processes.

Source: Adapted from the UIS GAML website http://gaml.uis.unesco.org/.

A growing global trend that is also a cause for concern

LSLAs are widely accepted instruments to help countries achieve *equity* and *quality* in education. As a result, the observed rise and spread of LSLAs around the world (national and cross-national) is expected to continue. This is particularly so since over the past decade, global and regional education actors have also "reframed their identities and work programmers in education" to incorporate LSLAs (Addey, 2018) while development partners increasingly encourage countries to engage in such initiatives.

As a consequence of these trends, contemporary global education policy debate revolves largely around measurement and comparisons of educational outcomes. A number of analysts, however, have raised concerns that the increased focus on the measurement of learning outcomes may influence educational policy and practice in ways that may be detrimental to the broader SDG 4 agenda. Some of these concerns stem from characteristics inherent to the design of LSLAs while others result from the possible analysis, interpretation and (mis)use of the data collected. The following chapters examine both of these limitations with particular attention to the implications for progress towards the global commitment for equitable, effective and relevant learning for all.

> A number of analysts have raised concerns that the increased focus on the measurement of learning outcomes may influence educational policy and practice in ways that may be detrimental to the broader SDG4 agenda. Some of these concerns stem both from characteristics inherent to the design of LSLAs, while others results from the possible analysis, interpretation and (mis)use of the data collected.

3. Valuing more than what can be measured: Limits to large-scale learning assessment design

This chapter reflects on the unintended consequences of large-scale learning assessments (LSLAs) that stem directly from their design. The section addresses four main dimensions that risk being negatively affected by the very nature and design of LSLAs: the conceptualization of education and learning; the general understanding of education quality; the range of knowledge, skills, attitudes and values taken into consideration; and the attention to inclusiveness and respect for diversity.

Constraining the conceptualization of education and learning

This first section explores how an over-emphasis on LSLAs may result in constricting the conceptualization of education and learning. The section starts with an overview on the plural purposes of education and the multiple dimensions of learning. It then examines how a focus on quantitative measurement and system-level comparability often leads not only to overlooking alternative assessment modalities better suited to capture certain domains of learning but, more importantly, to diverting attention away from a holistic view of education and its larger purposes.

Recalling the plural purposes of education

It is important to recall that education serves multiple purposes for individuals and communities alike. Beyond the acquisition of foundational and vocational skills as well as general knowledge and competences, education is also about developing values of respect for life, human dignity and cultural diversity required for social harmony in a diverse world (UNESCO, 2015a). It is about being aware of the choices made on a daily basis to live healthy and fulfilling lives, and contributing to the social, economic and environmental components of sustainable development. Education is relevant and closely interlinked to the various dimensions of personal and collective human development. Thus, it implies balancing and integrating the multiple social, economic, cultural, ethical, environmental, spiritual and political purposes of education.

Reaffirming the relevance of the four pillars

Learning is a complex process of acquiring or developing knowledge, skills and attitudes all underpinned by values (UNESCO, 2017c). An integrated approach to learning is captured by the four pillars of learning (UNESCO, 1996) - learning to know, to do, to be, and to live together (see **Figure 1**). Learning to know means learning the foundational and transferable skills that are the basis for lifelong learning and a preparation for active participation in the learning society. Learning to do implies learning to apply acquired knowledge and skills in both familiar and unforeseeable situations in work and life. Learning to be means learning to develop

one's personality as well as the ability to make independent judgments and having a sense of responsibility. Learning to live together entails the ability to understand others and have a sense of respect for human dignity, diversity and cultures to ensure responsible and active citizenship. These four pillars of learning are arguably even more relevant to sustainable human and social development in today's rapidly changing world – one which calls for new modes of learning to foster the competencies that societies and economies need today and tomorrow (UNESCO, 2017c).

Figure 1. The four pillars of education

- Independent judgement
- Sense of personal responsibility
- Tapping the buried treasure of hidden talents

LEARNING TO BE

LEARNING TO KNOW

- Instrumental basic learning skills
- Foundational skills
- Presupposes 'learning to learn'
- Foundation for lifelong learning
- Preparation for the learning society

- Learning for work and life
- Competencies to deal with unforeseeable situations
- Alternating study and work

LEARNING TO DO

LEARNING TO LIVE TOGETHER

- Understanding others
- Respect for human dignity and diversity
- Learning for responsible and active citizenship

Source: Adapted from UNESCO, 1996.

Assessment beyond measurement and learning beyond assessment

Curricular content and the expected outcomes of learning practices are broad. However, LSLAs are by design generally limited to a small number of easily-measurable areas of learning that lend themselves to system-level or cross-country comparability. These measurable areas represent "foundational skills" and/or basic general knowledge, such as reading and mathematics (and sometimes sciences) that are considered key to fostering skills and competencies as well as enabling further learning in these domains and beyond. Thus, LSLAs tend to overlook a variety of learning dimensions that are important for the multiple purposes of education.[24]

> Curricular content and expected outcomes of learning practices are broad. However, LSLAs are by design generally limited to a small number of easily-measurable areas of learning that lend themselves to system-level or cross-country comparability, and/or to so-called "foundational skills" and/or basic general knowledge.

Arguably, this is not inherently problematic. However, in a context of over-reliance on LSLAs, we run the risk of valuing what we measure rather than engaging in measurement of the learning we value (Biesta, 2009).

---

24 This has been the case with assessments that focus on the utilitarian aspects of learning and especially, its contribution to economic development – most notably, the PISA program. Other assessment programs have historically been characterized by a more inclusive or comprehensive approach and, consequently, are attentive to a greater number of learning dimensions. For reasons beyond the scope of this piece, however, existing assessment programs display growing convergence in relation to the areas of measurement, which translates to a greater emphasis on numeracy and literacy.

This is particularly the case given that global education discourse increasingly tends to equate assessment with quantitative measurement oriented at system comparability, while neglecting other forms of assessment. Assessment is critical to learning. However, it is questionable whether it is possible or even desirable to capture important social, civic and personal learning through traditional quantitative methods. For instance, key competencies in communication, collaboration, creativity, ethics and social relations are largely culturally dependent and strongly linked to personality. As such, properly assessing such competencies requires that learners demonstrate their attitudes and perceptions – and not simply their knowledge and comprehension. These may be better captured using different types of assessments, such as portfolio assessments, observation or even varieties of formative assessment that are not designed for comparability. There is, however, little to no space for such alternative assessment methods to monitor learning at the system level.

> Assessment is critical to learning. However, it is questionable whether it is possible or even desirable to capture important social, civic and personal learning through traditional quantitative methods.

Finally, we should be mindful that although learning is ubiquitous, discussions on its assessment tend to focus on the portion of learning that is purposeful and organized as well as on the outcomes that can be measured. While learning can be deliberate, intentional, purposeful and organized, it is also often much less institutionalized and may take place both intentionally or unintentionally in various activities of daily life, in the work place or non-formal education programmes, in the local community and on a self-, family- or socially-directed basis (UNESCO, 2015a). With the heightened emphasis on LSLAs in education discourse, there is a risk that the notion of learning is being reduced to the learning that can be measured.

Limiting our understanding of good education

As a consequence of such dynamics, it is possible that through their increased involvement in and attention to LSLAs, international agencies and national governments alike are helping create a reality where such assessments largely shape our understanding of what is valuable in education and education's goals (Hamilton, 2017). Such dynamics ultimately narrow our general understanding of what constitutes good education (Biesta, 2009), and are likely to shift discussions away from important dimensions of education that cannot be ignored. Crucially, such shifts do not only occur at the discursive level. As discussed in the next chapter, they are indeed likely to have a concrete impact when they become institutionalized in policy texts with the power to transform teacher practices and identities, the formal and hidden curriculum taught at schools, and the priorities of a range of education stakeholders.

Narrowing our understanding of education quality

This section explores how the growing emphasis on learning outcomes as a key dimension of education quality can divert attention away from other relevant components (e.g. societal outcomes, school and classroom-based settings, teaching and learning procedures), and/or the role of macro-level factors that affect learner achievement. The section starts with a brief

presentation of the multi-dimensional nature of the education quality construct and follows with a discussion on the balanced approach required to encompass the different components and indicators of education quality.

Education quality: A multi-dimensional construct

Traditional approaches to gauge the quality of education have relied on proxy indicators of educational provision and inputs that affect the progress of teaching and learning, such as pupil-teacher ratios, percentage of trained teachers or expenditure per pupil as a percentage of the GDP.[25] More recently, learning outcomes have been gaining prominence as a crucial indicator to gauge the quality of education. The *Global Education Monitoring Report* (GEMR), for instance, has recently proposed a revised framework to understand education quality (UNESCO, 2017d) that includes an important focus on learning outcomes for individuals and societies as determined through performance on LSLAs.

As emphasized by the GEMR, however, individual outcomes cannot be understood in isolation or as the only valid measure of education quality. Dimensions other than learning outcomes need to be taken into consideration as well. These include the outcomes for society (e.g. the encouragement of behaviors linked to sustainable social and environmental development, global citizenship, living together and diversity) as well as school and classroom settings. The latter encompasses the teaching-learning process; the presence of well-trained and motivated teachers as well as a rigorous teaching process; strong school leadership; community participation and democratic governance; and adequate structures and material inputs.[26]

More generally, understanding learning outcome levels or how they change over time requires consideration of macro-level factors, such as the economic, political and social contexts or the structure and organization of the education system (i.e. policies, system financing, curriculum, standards and accountability, recruitment and incentives, professional development). Not doing so will inevitably result in a biased analysis and perspective. Thus, improved education quality does not necessarily lead to improved learning outcomes as variation in student learning is attributable not only to school-related factors but also to family and community characteristics (Ungerleider, 2006). While many LSLAs are precisely designed to gain a better understanding of the impact of such factors, their centrality may sometimes escape the attention of the public, politicians and education practitioners – particularly, if the dissemination of results focuses on comparative perspectives and, in the case of cross-national assessments, emphasizes the country rank vis-à-vis other countries. While LSLAs are needed to monitor the performance of education systems and improve learning and equity in schools, they have greater potential when considered alongside a range of other factors in and outside the school.

---

25   These approaches are summarized in Tawil et al. (2012). They include, for example, reference to the UNESCO Santiago Model (UNESCO Santiago, 2007), which proposes five dimensions of the quality of education that aim to capture the perspectives of various education stakeholders; the analytical framework by Tikly and Barret (2007), which calls for the analysis of historical, cultural, socio-economic and political contexts to understand how education systems operate; the Fabric Model of Quality Education by Nikel and Lowe (2010), which proposes seven conceptual dimensions arranged to emphasize that the quality of education is much like a "fabric" that is at its strongest when "stretched" or maintained in tension.

26   A similar point is made by Unterhalter and Brighouse (2007), who, drawing on the capability approach perspective, note that learning outcomes are inappropriate or insufficient to capture key dimensions of education quality other than instrumental purpose, including education's intrinsic and positional values.

Assessment: A necessary yet insufficient component to capture education quality

Understanding various other dimensions related to the learner, the teacher, the school, the education system and economic, social and political contexts is crucial to evaluating the quality of education (Colclough, 2012; UNESCO, 2016b; Zadja, 2014). Data on these dimensions are not always generated by LSLAs but through other sources, such as administrative or household surveys as data collected through background questionnaires is not always sufficient. Thus, LSLA data considered and analyzed in combination with a range of other data generated at country level would better capture the different dimensions of quality education.

To be sure, this kind of analysis is frequently conducted by researchers and specialists based in government agencies or assessment consortia. However, such relations are frequently de-emphasized (or directly overlooked) by less specialized reporting, such as those put forward by the media. Further, a balanced and multi-dimensional approach appears to be progressively rare as LSLAs continue to gain prominence. Their results are indeed increasingly accepted as a complete measure of quality education (Hanushek and Woessmann, 2015; Komatsu and Rappleye, 2017) and have also been referred to as the "major component" (ACER-UIS, 2017) or "a critical baseline" (Friedman et al., 2016) of quality education. While this generally demonstrates a concern for learning as an important outcome of schooling, there is a risk that education quality may get narrowly equated with improved learning outcomes as monitored through LSLAs (Sayed et al., 2018; Smith, 2016).

> Ultimately, levels of learning outcomes as measured by LSLAs are a necessary dimension but not the only one needed to measure the quality of education.

Using results from learning assessments as the sole means of evaluating the quality of education has multiple repercussions. It risks shifting focus away from important areas of quality education, such as striving towards equity; ensuring safe, adequate and effective learning environments; deploying qualified teachers and expanding their capacities; ensuring the relevance of the curriculum; and providing good school leadership. It may also result in the neglect of other social, economic and contextual dimensions of education quality. Ultimately, levels of learning outcomes as measured by LSLAs are a necessary dimension but not the *only* one needed to measure the quality of education. It is even questionable if quality can be adequately gauged without reference to these other social, economic and contextual dimensions.

Restricting what we value

As discussed above, LSLAs may entail a narrowing of the multiple purposes of education and dimensions of learning. This section elaborates on one of the processes through which this narrowing occurs – namely, the limited number of domains targeted by most LSLAs. The section first discusses the different modalities through which this curricular reduction operates and then provides examples of efforts to integrate a broader range of knowledge, skills, attitudes and values into LSLAs along with the challenges faced in such a pursuit. Finally, the section reflects on the desirability of such efforts and the possible risks they entail.

Limited domains and subdomains

As argued above, the limited number of targeted domains and abilities can be considered intrinsic to the design of LSLAs. These assessments necessarily focus on a limited number of subjects and abilities – more specifically, easily-measurable areas of learning that lend themselves to system-level or cross-country comparability, or are considered foundational skills that enable further learning in other areas.

Consequently, some have argued that LSLAs are too narrowly focused in two ways. First, that they do not adequately assess the broad range of knowledge across a wide spectrum of domains, such as the arts or media and information literacy in contemporary cyber-societies, nor do they assess broad competencies that schools are meant to develop, such as citizenship or environmental responsibility. Second, within the domains assessed, LSLAs are insufficiently attentive to the breadth of knowledge available. For example, in assessing language, subdomains, such as the appreciation of literature, writing skills and listening comprehension, are often missing from these assessments even if they are important goals of the discipline (Ungerleider, 2006).

Beyond numeracy and literacy?

The design of LSLAs thus tends to overlook the plural purposes of education and the dimensions of learning and, by implication, the principles at the core of the Education 2030 commitments. This is a consequence of the content of the assessments as well as their implementation process and the question formats used in their design. Multiple choice or short answer test items are designed in a way that can be easily scored against a defined set of rubrics, limiting the possibilities of assessing perceptions, attitudes and judgments. Similarly, most assessments do not necessarily aim to capture data relevant to healthy lifestyles or values and attitudes that support democracy, the protection of human rights, respect for cultural diversity, active responsible citizenship and environmental sustainability.[27]

In this section, we reflect on some of the knowledge domains, skills, attitudes and values which, while constituting key elements of the Education 2030 agenda, are only partially (and with great difficulty) captured by LSLAs. A review of the test items used by these assessment instruments

---

27 Difficulty evaluating these domains cannot be dissociated from the fact that they rarely constitute standalone subjects within the national curriculum but are typically embedded in the curriculum as cross-cutting themes.

will help illustrate the difficulty of incorporating a wide range of competences (i.e. learning domains, knowledge, skills, attitudes, values) into ordinary testing practices.

Healthy lifestyles: Insufficiently addressed domains and regressing

Education is considered one of the most powerful ways to improve an individual's health and ensure that the benefits are passed on to future generations. Reciprocally, health is a key enabler and determinant of learning (UNESCO, 2016b; 2016c). Yet, most LSLAs do not explicitly support education for healthy lifestyles. In terms of national assessments, health, nutrition and the well-being of children – as well as of their parents or guardians – are most frequently measured in early childhood development as it is deemed to be a key dimension to ensure that children are learning and remain on track for primary education (UNESCO et al., 2017). There is little to no evidence that the notion of healthy lifestyles is a core component of LSLAs targeting children of primary or secondary school age – or even adults. This is particularly true for nationally developed assessments. To be sure, background questionnaires frequently collect data on student practices and habits in relation to health or nutrition. However, these generally only contain questions to determine whether or not learners are *in conditions* to learn – that is, whether they have access to proper healthcare services, enjoy appropriate and sufficient nutrition at school, and so on. No information is generally collected on their approaches to disease prevention or their opinions on behaviors, such as health and dietary habits or exposure to sexual risk. Likewise, while some items may be designed to address health issues, the tasks are generally aimed at measuring specific skills, such as reading comprehension or solving a multistep problem.

For example, the PISA 2015 Science item Running in Hot Weather (OECD 2015)[28], includes five questions on the interaction between health and the environment. This item aims to assess a range of abilities: to interpret the display; manipulate one variable and compare the outcomes of two trials; infer the function of sweat from the written text and the simulation; draw on procedural knowledge to explain how the data they have collected supports their answer; and develop a hypothesis, extrapolating beyond the data that can be directly collected through the simulation. The learner is not expected to have any prior knowledge on the topic to answer the questions as the responses can be deduced from the stimulus. Similarly, it is not intended to assess decisions/actions/judgements of exercise in hot weather. Hence, PISA test questionnaires rely on health-related materials in order to assess science and problem-solving abilities – but all cues are given in the stimulus, which precludes the element of judgment from the assessment.

Some efforts have nevertheless been made to place a greater emphasis on the notion of healthy lifestyles. For instance, in response to the high rates of HIV/AIDS in sub-Saharan Africa, SACMEQ 2007 included a module on HIV and AIDS to ascertain children's attitudes to people living with the disease, their perceptions of courses on HIV and AIDS, and their preferred source of information on the disease.[29] The extent to which teacher training covers issues related to HIV and AIDS, which are central to any effective education strategy to promote prevention, were also studied (Dolata, 2011).

---

28 The PISA 2015 Science item Running in Hot Water can be accessed online at the following link: http://www.oecd.org/pisa/PISA2015Questions/platform/index.html?user=&domain=SCI&unit=S623-RunningInHotWeather&lang=eng-ZZZ
29 Please note, however, that the module on HIV and AIDS was not included in the last cycle of SACMEQ (2013-14).

Global citizenship education and Education for Sustainable Development: Increasingly defined and assessed domains

Education for Sustainable Development (ESD) and global citizenship constitutes one of the key dimensions of the Education 2030 agenda. As early as 1974, the General Conference of UNESCO adopted the *Recommendation concerning Education for International Understanding, Co-operation and Peace and Education relating to Human Rights and Fundamental Freedoms.* The 1974 Recommendation encouraged states to take the necessary institutional and pedagogical measures to incorporate a series of principles (i.e. respect for cultural diversity, the promotion of peace and non-violent conflict resolution, fairness, caring for the planet, etc.) into their education programs (UNESCO, 1974).

As important as such endeavors may be, designing measures to assess related knowledge, skills, values and attitudes is a complex task. This is particularly true as existing tools are often limited to self-reporting questionnaires, which are typically affected by a social desirability bias as participants commonly attempt to respond in ways they perceive as socially acceptable. Compounding the problem further, in cross-national assessments it can be difficult to capture culturally relevant norms specific to particular countries or regions and preserve the cross-country comparability of collected data at the same time. Currently, technical expertise to remedy this problem seems to be at a developmental stage. Furthermore, it is unclear whether ESD will be included in national learning assessments in the near future. In Europe, although the knowledge dimension of global citizenship is generally included in national assessments, students' attitudes and skills are less likely to be covered (De Coster and Sigalas, 2017).

Capturing citizenship education in national assessments is particularly challenging as it encompasses a variety of teaching/learning approaches, from subject-based or cross-curricular delivery to extra-curricular activities and participation in school life (De Coster and Sigalas, 2017). Nevertheless, efforts are being made to do so. Cross-national learning assessments, such as OECD's PISA Global Competencies assessment and IEA's International Civic and Citizenship Education Study (ICCS), are making an effort to assess components of ESD and citizenship education. Similarly, the ICCS, which has been conducted in 38 countries, is designed as a non-curriculum-based assessment to capture how young people prepare to undertake their roles as citizens (IEA, 2016b). The study reports on students' knowledge and understanding of concepts and issues related to civics and citizenship as well as their beliefs, attitudes and behaviors. **Box 4** provides an example of the approach adopted by ICCS to assess knowledge on religious diversity as well as attitudes and dispositions towards tolerance. IEA has also encouraged the development of regional reports to analyze the contexts for civic and citizenship education as well as students' perceptions of public institutions and government, their views on peaceful coexistence and perceptions of social cohesion and diversity.[30]

---

30   To capture these dimensions, pupils are provided the space to write and express their opinions, adding another challenge to scoring the assessment.

Box 4. **Respectful co-existence – Test item from the International Civic and Citizenship Education Study 2009**

Below is a sticker that people can buy on the internet.

The sticker is made up of symbols that represent different ways of thinking about the world. The symbols have been put together to look like the English word 'coexist' which means 'live together'.

Q What is the **most likely** purpose of this sticker?

☐ to show that different ways of thinking are all the same

☐ to show that people should think carefully about what they believe

☐ to show that people can accept others even if they have different beliefs

☐ to show that people with different ways of thinking about the world can never happily live together

Source: Brese et al. 2011, p. 7 (reproduced with permission).

In the global South, the SEA-PLM is currently the only cross-national assessment with a module to measure global citizenship attitudes and behaviors. This module focuses on attitudes and values – rather than cognitive outcomes, behaviors and skills – and is expected to reflect universal values as well as those specific to the South-East Asian nations. Thus, the recently developed assessment framework acknowledges the challenges posed by an innovative assessment area yet makes it explicit that the assessment must be designed in a culturally appropriate way that is coherent with the idiosyncrasies in the South-East Asian region while remaining cognizant of its heterogeneity (Parker and Fraillon, 2016).

Writing skills open doors to measuring creativity, curiosity and the appreciation of culture

It might be worth noting that LSLAs do not assess writing, which is a foundational skill required for communication, future learning and full participation in economic, political and social life as well as in many aspects of daily life.[31] In a digital age and in the context of a knowledge economy, personal and social communication is increasingly conducted in written text, including through mobile phones and social media. Assessing writing skills or the use of them to measure domains, such as creativity, curiosity and the appreciation of culture, also generally lies beyond the scope of LSLAs. This is due to the fact that a written text is particularly difficult to measure in a way that preserves system-level comparability – as a consequence of its subjective nature.

Recognizing the importance of writing skills and the fact that they are measured in examinations, efforts are being made to include this dimension in LSLAs. While at a developmental stage, SEA-PLM marks the first cross-national initiative to measure *writing literacy*, understood as the ability to construct meaning by generating a range of written texts to express oneself and communicate with others to meet personal, societal, economic and civic needs. This definition acknowledges that writing goes beyond the ability to spell words in legible handwriting. This understanding of 'writing' "conveys the idea that it is done in a context, for an audience and with a purpose. It includes a range of cognitive skills such as generating and organizing ideas, applying vocabulary and drawing on knowledge of linguistic structures and textual features" (SEA-PLM et al., 2017, pp. 30-31).

Moving towards measuring collaborative thinking and abilities as well as socio-emotional skills

In general, LSLAs exclude the measurement of collaborative thinking and skills. Yet, these are critical to the generation and incubation of ideas (Hannon et al., 2011), which are increasingly the product of cooperation among a group or groups of individuals thinking and working together and drawing on existing knowledge to create new knowledge (OECD, 2018a). Collaborative skills and thinking (i.e. the ability to relate well to others, to cooperate, work in teams, manage and resolve conflict, etc.) are difficult to measure in individually undertaken assessments. Such skills are increasingly required by young people to adapt to the rapidly shifting demands of the labour market. While measuring group decision-making, coordination or production through individual assessments has proven to be particularly challenging, some efforts are being made to capture individual capacities within collaborative situations. PISA, for example, employs computer-based tests to assess students' collaborative skills using digitally simulated agents to represent team members with a range of skills and abilities. This approach affords the high degree of control and standardization required for measurement. It also enables the observation of students in a number of collaborative situations and allows measurement within the time constraints of the test administration (OECD, 2017). More generally speaking, however, such efforts remain at an essentially nascent stage as there is still a limited understanding of the possible challenges.

---

31   This cannot be dissociated from the fact that instruction and mass literacy efforts were historically focused on reading rather than on writing.

Given their intrinsic value, socio-emotional skills is another area of measurement that is receiving more attention as a key determinant of learning progress. For instance, LLECE is currently working on a socio-emotional module for the next regional assessment – ERCE 2019. The module was piloted in 2018 and aims to reflect the broad understanding of education quality advanced by Education 2030 and to examine the effect of such skills on learning outcomes and the factors determining such abilities.

Increased relevance of digital skills

The debate on the measurement of digital skills has only recently entered the global education arena. Remarkably, this is largely due to the selection of digital, media and information literacy skills for SDG target 4.4 which, in turn, focuses on skills for work among youth and adults. This explains why conceptual and technical work conducted at the global level has been somewhat disconnected from debates on the assessment of school-aged children.

While there is a burgeoning number of initiatives oriented at assessing digital and ICT skills, they are highly heterogeneous in nature, in content and in terms of responsible authorities (UIS, 2018c). Moreover, experiences with digital assessments appear to be largely restricted to high-income countries and cross-national initiatives targeting basic education students are rare. An exception to this is IEA's International Computer and Information Literacy Study (ICILS). Other assessments in place, however, typically target the adult population.

In any case, a common trend across these initiatives is that most assessments focus on a relatively limited number of skills, such as digital information use and technical skills, while other areas, including online collaboration and problem solving with ICT, are less frequently addressed (Siddiq et al., 2016). Again, such trends suggest that, by design, LSLAs can only help evaluate a relatively limited number of skills. This is particularly problematic as increasingly complex and sophisticated digital skills constitute the basis for further learning. In other words, learning in a variety of domains is largely conditioned by one's ability not only to perform technical operations related to ICT but to engage in complex problem solving, critically evaluating information and so on (UNESCO, 2018). Despite measurement difficulties, some incipient initiatives are precisely attempting to capture a more comprehensive set of ICT-related competencies. This is the case for UNESCO Bangkok's *Digital Kids Asia Pacific* project, which has recently developed the regional Digital Citizenship Competency – a framework covering a wide range of domains, including digital safety, emotional intelligence and creativity. The assessment framework has been piloted in four countries and is expected to provide an evidence-based understanding of children's use of ICT beyond basic skills.[32]

Room to manoeuver: Capturing other areas of competence

Summing up, and as the previous examples suggest, organizations and national governments are attempting to capture other areas of competence, such as transversal, transferable, or so-called 21st century skills, including "non-cognitive" or cross-cutting skills and competencies

---

32 For more information on this initiative, consult https://teams.unesco.org/ORG/fu/bangkok/public_events/Shared%20Documents/EISD/2017/
Oct2017%20-%20KFIT%203%20Launch%20-%20Dig%20Citizenship/DKAP-project-brief.pdf

related to sustainable development, citizenship in a global world and health as well as those related to digital literacy. As these areas are understood to represent the core of the skills and competencies required to be effective workers and citizens in a knowledge-based society, the OECD, the IEA, as well as regional assessments, have launched a range of initiatives to assess them. This reflects a broader acceptance of a holistic view of learning where not only basic general knowledge and "foundational skills," but also more specific competencies related to global citizenship, sustainable development and health, are deemed central and relevant in today's world. Likewise, a range of overarching knowledge, skills, attitudes and values necessary for effective communication, collaboration, problem-solving and social relations are favored.

> The response to the diverting effect of LSLAs cannot rely exclusively on more testing. Rather, the solution requires exploring ways to better integrate LSLAs into broader assessment systems, combining them with other forms of evaluation.

At the same time, awareness and competency in areas related to socialization and acquisition of values, which appear to be culturally dependent, may be better captured in non-traditional forms of assessment that are incorporated in the teaching and learning process. Thus, despite interest in capturing such competencies, the desirability for global measures remains questionable. In addition, such efforts could translate into greater testing frequency and a considerable increase in the resources devoted to assessment – in both financial costs and time.

In sum, it is possible to recognize the tension between the need to broaden and diversify the scope of LSLAs and the impossibility and undesirability to assess every existing domain of learning through LSLAs. A corollary of this tension could be that the response to the diverting effect of LSLAs cannot rely exclusively on more testing. Rather, the solution requires exploring ways to better integrate LSLAs into broader assessment systems, combining them with other forms of evaluation so that they do not become the main and single focus of attention.

Limiting who we value

This section reflects on the limitations presented by LSLAs in terms of inclusion. After discussing the potential of LSLAs as equity instruments, the section explores different ways in which certain groups are excluded from assessment exercises – or participate in comparatively unfavorable conditions. Finally, the section reviews a variety of efforts recently made to increase assessment inclusiveness – and the challenges and risks that these entail.

Assessment at the service of equity

The Education 2030 agenda is unambiguous in its commitment to equity and inclusion – both considered cornerstones of a truly transformative agenda. Equity entails including everyone in the educational process while making sure that everybody, regardless of their background, is learning. Calls for a *learning equity agenda* (Wagner, 2018) are precisely an attempt to capture the need to increase learning levels among the most disadvantaged segments of the population.

This approach has key implications for measurement efforts. Assessment design is expected to take into consideration policy priorities with regard to equity at the beginning of the assessment process and collect the right kind of information to help link assessment results with policies to reduce inequality. More specifically, this renewed vision for education requires data to be disaggregated by a wide range of dimensions when possible (UNESCO, 2016c). LSLAs do hold the potential to help us understand different dimensions of equity in the education system, including regional equity, school equity, gender equality, linguistic equity, socio-economic status and sub-group equity (e.g. students with disabilities or recent immigrants/ refugees) (Brink, 2018).

> Assessment should be inclusive, designed to be relevant for as many members of the target population as possible and to respond effectively to all learners.

Assessment is therefore expected to be inclusive – i.e. designed to be relevant for as many members of the target population as possible and to respond effectively to all learners (UNESCO, 2017b). In other words, it is expected that the sample of learners will not purposefully exclude certain groups – for instance those with disabilities.[33] It is also expected that the test takers in different parts of the country or across countries perceive and respond to the questions in the same way, regardless of their ethnic background, immigrant status, gender or residence.

Ensuring these principles of equity and inclusion hold important implications for measurement efforts. For instance, it requires the translation and adaptation of test materials to suit different population groups or to accommodate children with special needs. It also requires additional efforts in over-sampling vulnerable or minority populations and ensuring that the number of students from these groups fulfills the minimum requirement for a meaningful analysis. In addition, instruments may need to be adapted and conditions of testing adjusted to ensure that children in these groups have an equal opportunity to perform the test. Likewise, background questionnaires have to be adjusted to ensure that relevant information is adequately captured.

Who is excluded from assessment?

The inclusiveness imperative is not always observed. The reality is that many vulnerable and marginalized groups (individual students or schools) are excluded from the assessment design and sampling for various reasons. To start, out-of-school children (or children enrolled in unregistered schools) are generally not included in the sample – as LSLAs focus on the formal education system and are administered only to children enrolled in authorized and officially recognized schools.[34]

Entire schools are sometimes excluded because they are located in remote hard-to-reach areas or when they provide education to groups of children excluded from the test (e.g. school for the blind). Exclusions also occur at the individual child level within schools. Children with moderate to severe permanent physical disabilities, those with cognitive, behavioral or emotional disabilities

---

33   See Rowe, E. 2017. Australia Country Case Study. Background paper for *Global Education Monitoring Report* 2017/8, on the practice of selective sampling in Australia.

34   There is growing awareness on the need to incorporate such populations into LSLAs. Recent attempts to incorporate out-of-school children and young people include PISA for Development as well as EGRA and certain citizen-led assessments (cf. Yasunaga, 2014).

as well as those with insufficient language experience tend to be excluded. Compounding the problem, there is limited information on and understanding of the measures taken by many national assessments to tackle exclusion and accommodate all learners. Although these are crucial questions, technical guidance is limited and frequently restricted to high-income contexts.

Exclusion and misrepresentation dynamics are particularly visible in countries with significant levels of linguistic diversity. In such contexts, language of instruction is a challenge and can often hold the key to making education more inclusive for disadvantaged groups (UNESCO, 2016b). While it is difficult to estimate the share of students who are taught in their home language (UNESCO, 2017d), one estimate indicates that as much as 40% of students are not taught in a language they speak or fully understand (Walter and Benson, 2012). LSLAs as well as other assessment modalities are generally administered in the language of instruction, regardless of whether or not the child fully understands it. This can result in inadvertent exclusion or increased non-participation of marginalized groups. It may also be the case that children from these groups particpate in LSLAs but that their real learning levels remain invisible in education data at the global and national levels (UIS, 2018a). These trends conceal the challenges countries face regarding equity issues and the deprivation of vulnerable groups of their right to education. Of course, assessment also has the potential to draw attention to such issues. For instance, the low rates of achievement help raise awareness to government officials and development professionals that at least initial reading and writing should be taught in a learner's own languages (Benson and Wong, 2015). Interpretation remains thus a key determinant of the ultimate impact of LSLAs – and their potential to contribute to an equity agenda.

Potential risks in over-accommodating learners

There have been recent efforts made to extend assessments to marginalized groups. ASER Pakistan has been testing children with moderate to severe difficulties in seeing, hearing, walking, caring, understanding or remembering (Singal and Sabates, 2016). Similarly, RTI international has adapted the EGRA instruments to accommodate children with vision and hearing impairment, and to improve the classroom experience for them. PISA 2015 and beyond has also made efforts to broaden access to their assessment. However, as illustrated in **Box 5**, the intention to increase equity and inclusion can be problematic in its own way.

Efforts are increasingly made to translate and adapt assessment instruments in multilingual contexts. Likewise, all cross-national initiatives have elaborate translation and adaptation guidelines and attempt to make the test as comparable as possible across countries and languages. Although very few LSLAs in developing countries are conducted in most or all languages of instructions, some commendable efforts are nevertheless being made to include more languages in the implementation of the assessment. For instance, PASEC 2019 will also be administering its tests in the early grades in mother tongue languages after only administering the test in the main language of instruction in its 2013-14 cycle. Similarly, South Africa administered its Annual National Assessment in 11 other mother tongue languages until its abolishment in 2014 (UIS, 2019a).

Box 5. Test accommodations – a double-edged sword?

In 2011, PISA's governing board expressed interest in seeking ways to widen access to those assessments for students who are intellectually disabled, functionally (physically) disabled and those with insufficient language experience. Until then, accommodations had been limited to small-group or one-on-one testing. Other forms of accommodations (e.g. graphic modifications, audio presentation, bilingual dictionaries, dictation of answers, word processors, adaptive denatured or tools, etc.) have not been widely allowed in PISA. Without these types of accommodations, students who could otherwise participate in a meaningful way must be excluded from the assessment. Yet, it is shown that the accommodation most frequently required by students with special education needs is unlimited or extended time, which is not permitted for students taking the standard PISA test booklets.

However, providing accommodations to include marginalized or vulnerable groups as part of the sampling processes does not come without risks. First, there is a concern that student needs may vary to an extent that makes it difficult to guarantee comparability within and between countries. A second concern is that some accommodations may appear to place students at a comparative advantage to those not using the accommodations. Finally, accommodations could provide a perverse incentive to identify students as special needs even if they are not simply to improve performance.

Source: Based on OECD, 2011.

Floor effects

So-called 'floor effects' constitute another design-related limitation that may reduce the potential of assessment to support an equity agenda. Floor effects occur when most of the participants in a given assessment attain scores close to the baseline. As discussed by Wagner et al. (2018), an excessive emphasis on comparability at the expense of local validity (as entailed by LSLAs) might render assessments inadequate to provide a proper understanding of the situation of disadvantaged populations – which are more likely to be found at the tail end of the score distribution. A similar point is raised by Winthrop and Anderson (2013) who note that, although assessments with considerable floor effects can draw attention to some relevant issues, their analytical potential is likely to suffer. In short, limited local validity has a particularly negative impact on the potential of LSLAs to support the equity agenda.

> An excessive emphasis on comparability at the expense of local validity might render assessments inadequate to get a proper understanding of the situation of disadvantaged populations.

Do limits in design lead to limits in use?

Overall, the trends described in this chapter are indicative of the necessary partial and imperfect nature of the data produced by LSLAs. This should not be read as a flaw or a limitation exclusive to standardized testing – as any form of data collection is likely to prioritize certain competencies over others and will inevitably rely on a selective definition of education quality. However, the trends described here invite a note of caution on the risks of unbalanced and excessive credence or attention to LSLA data. If not appropriately combined with other data sources (qualitative or quantitative), the inherent limitations of LSLAs are in fact setting the stage for their misuse. Thus, because assessment data is by definition partial and imperfect, both unbalanced attention to LSLAs or failure to integrate them into other system components, are likely to hinder their ability to contribute to a meaningful advancement towards greater education quality and equity.

4. Beyond data as an end in itself: Limits to the use of large-scale learning assessments

The previous chapter revealed that the design of large-scale learning assessments (LSLAs) is at the heart of some of their unintended consequences. However, most of the limitations associated with such tools stem directly from the use of the resulting data – or lack thereof. This chapter first provides some conceptual clarifications regarding the purposes of LSLAs and reflects then on three main groups of concerns over the use of resulting data – namely, its under-use, over-use, and combination with (or subordination to) accountability measures.

Uses of large-scale learning assessments: Clarifying terms and concepts

A discussion on the unintended consequences of LSLAs must necessarily be preceded by an overview on the purposes and intended uses of such tools. To this end, this section starts by providing some conceptual clarification on the type of information generated by LSLAs. Building on this, it then discusses the multiple uses and purposes of LSLAs, and the different stakeholders that can benefit from them.

Differentiating 'results' from 'evidence'

When using LSLA data, a 'result' is a finding that stems from a single analysis. For example, the average scores of boys and girls in a given subject or learning domain; or the distribution of a given group across different proficiency levels. 'Evidence,' on the other hand, is a package of findings used to provide a data-based response to a policy (or practice-related question) usually by describing trends, identifying patterns, pointing to key factors to explain certain regularities and highlighting potential areas of intervention. The policy or practice response should be based on evidence or results from several analyses or could be gleaned from multiple data sources and, most importantly, must include an element of judgment and reflection (Brink, 2018).

For example, if the policy is intended to ensure equity between boys and girls and to reduce the gap in performance between them, then the evidence could comprise an examination of the way in which the gaps have widened or narrowed over time and/or if there is a persistent trend. In combination with a historical review of the policy efforts put in place to address the issue as well as the research on key policy sensitive factors that affect performance, this information would then be instrumental to policy design and formulation (Brink, 2018). Thus, results do not speak for themselves and evidence-based policy advice does not *naturally* stem from raw data.

Clarifying the difference between analysis and use

Data from LSLAs – at a single point in time or across assessment cycles – give governments evidence to address system inefficiencies by providing answers to key questions, such as who is learning what and who is not, where, when and why (Montoya, 2016). Learning achievement scores and information from the background questionnaires are generally used by experts and researchers[35] to describe the knowledge and skills of a target population. This generally involves several types of analyses. First, understanding the factors that influence learning achievement (e.g. home and school context and practices), and if these are changing over time. Second, identifying general trends in learning achievements and evaluating progress towards specific targets using a set of indicators. Third, highlighting disparities in cognitive abilities among sub-populations of learners by relevant dimension, including socio-economic, regional, gender, migration status and mother tongue.

Policy-makers, in turn, use the results or evidence from LSLAs for a variety of purposes:[36]

Monitoring and evaluation: to monitor a set of issues and trends towards national, regional or global goals and targets; and to evaluate education policies.

Inform policy formulation: to orient policy formulation, design and implementation or improve the effectiveness of policy measures already in place (**see Table 5**).

Agenda setting: to set agendas by raising awareness on certain issues, stimulating debate within civil society and government circles, identifying priority issues for reform and generating 'positive pressure' to incite governments to undertake reforms or advance policy change – particularly when the assessment results are surprising or worrisome.

Accountability: to hold a range of agents accountable, including immediate education stakeholders (administrators, principals, teachers, school boards, private providers) as well as those arguably external to the education system (governments, local authorities, etc.).

In-depth analysis: to add depth and perspective to the analysis of education systems, and direct policy-makers and practitioners to further focused studies that could more adequately reflect their local and contextual realities.

While the potential use for LSLA data by policy-makers, experts and researchers is more or less known, there are other groups who may also be interested in the data and findings of the assessment. These include teachers and school leaders; parents and students; interest groups, such as NGOs, curriculum developers, teacher trainers or teacher unions; and researchers and academics conducting further analysis. The general public can also benefit from learning assessment results so long as they have been adapted to non-technical audiences – for instance,

---

35 Analysts typically come from government agencies, public institutes of evaluation, university departments and research groups; specialized private organizations; cross-national assessment consortia; and international organizations. Their exact affiliation, however, varies greatly with administrative, educational and cultural traditions.

36 The uses and purposes of LSLAs have been reviewed by a wide range of scholars. For a discussion of the listed items, refer to Tobin et al. (2015), Lingard and Grek (2007), Goldstein and Thomas (2008), Fullan (2010) and Kellaghan et al. (2011).

through the use of simple infographics on social media (for an overview of the range of potential users and target audiences, see UIS 2018c).

Each one of these groups can in fact use and benefit from LSLAs in different ways. For example, teachers can rely on assessment results to adapt teaching methods and provide adequate support to students while parents and the public can use them to hold schools or administrations accountable (UIS, 2018c). There is, however, limited knowledge on how and under what circumstances different education stakeholders are more likely to make effective and appropriate use of available data, and about their ultimate impact and the possible risks and challenges this may entail.

Table 5. **Examples of policy initiative that large-scale learning assessment data can inform, by level**

SYSTEM	CURRICULUM	SCHOOL	TEACHER	HOME
• Allocating resources • Implementing programatic reforms • Outlining goals for curricular achievement	• Curriculum development • Curricular content and methods • Curriculum design	• Setting faculty priorities • Improving student support services • Enriching school environment	• Securing resources for professional development or improving pedagogical practices • Revising courses and assignments	• Informing programmes to encourage parental involvement

Source: Authors.

Under-using assessment data

Evidence from LSLAs is assumed to be a key ingredient for the improvement of education systems and learner achievement. However, this change can only occur if key stakeholders act upon the available evidence – analyzing the data and communicating the results to key audiences and using it to guide policy action. In this section, we address less commonly acknowledged factors that frequently constrain an effective and sufficient use of data – particularly, the limited ownership in funding, design, management and dissemination.

Data collection will not improve quality

As previously noted, LSLAs have become essential components that provide key insights to multiple dimensions of education systems. As such, understanding and monitoring the results as well as why and how they may have changed is necessary for the adoption or adjustment of any policies and strategies to improve learning. If the resulting data is not carefully analyzed and disseminated to the right stakeholders who can then use it for further action, the process becomes ineffective (Kellaghan and Greaney, 2001).

However, the use and dissemination of results cannot be taken for granted (Tobin et al., 2015; UIS, 2018c). The learning assessment process typically culminates with the preparation of a report that summarizes the findings and is perhaps announced at a launch event – but

discussion of the results tends to die out after that. Governments, regional or international organizations and other users too frequently remain focused on explaining the current state of education and trends, or on reporting only on a set of pre-determined indicators rather than figuring out how to best use the findings to improve the context and process that enables learning. Another roadblock to the productive use of results is that LSLAs are not always perceived as useful by teachers and educators and consequently, they might end up having little impact on teaching-learning processes. While reporting is a necessary step (as it requires understanding the results and complex relationships between learning and the variables that influence it), capturing data does not automatically translate into the improvement of learning. When assessment is not used to drive change, the measurement of learning outcomes risks becoming an end in itself.

> The use and dissemination of results cannot be taken for granted. While reporting is a necessary step, capturing data does not automatically translate into the improvement of learning.

As previously noted, the development and administration of LSLAs is resource heavy as it requires substantial funding, infrastructure and capacity. The total cost may be significant for low- and lower-middle income countries.[37] Thus, making use of the resulting data is crucial to maximize the utility of the investment and ensure that assessment is not perceived as a waste of resources that places an unnecessary strain on public finances.[38]

Acknowledging barriers to enable effective use

Despite the many purposes that LSLAs are expected to serve, there is limited understanding of how and under what conditions assessment results can be used effectively. Some of these determinants have been discussed extensively and concerns can be found in seminal works on LSLAs, including Kellaghan and Greaney's *Using assessment to improve the quality of education* (2001) (see also UIS, 2018c and Raudonyte, 2019 for a more recent account on barriers constraining the use of LSLA data). Frequently acknowledged constraints are manifold and include financial and technical barriers (e.g. limited or fragmented expertise among concerned stakeholders or the poor quality of assessments in place); lack of political engagement or will; poor integration into policy processes; limited alignment between assessment and other components of the education system; limited reliability or relevance of the information; insufficient depth of analysis; inadequate presentation formats; and poor or inappropriate dissemination (UIS, 2018a, 2018c; Lockheed, 2015; Tobin et al., 2015).

Meanwhile, other constraints remain under-theorized or under-explored and, more importantly, discussions on the many barriers limiting use remain relatively de-contextualized. Consequently,

---

37 Overt financial costs include staffing, training and capacity development, travel and equipment, reporting and dissemination (Wagner, 2011; Lockheed, 2015). In addition, participation in some cross-national initiatives also includes a participation fee. Total direct costs range from $400,000 USD to $2 million USD (Clarke, 2016).

38 There has been some debate on whether the benefits of assessments are worth their cost. Some have argued that they are among the least expensive innovations in education reform (Clarke, 2012) or that their cost is relatively small compared to the cost of policy implementation or of making uninformed or incorrect decisions (UIS, 2016a). However, other scholars argue that any analysis of the cost-benefit of such exercises entails an element of normative judgment and that such issues have been insufficiently addressed in public discourse (Engel and Rutkowski, 2018).

The limited use of assessment results can never be exclusively attributed to a lack of political will. Rather, limited opportunities, support, resources or capacity of key education stakeholders to access, interpret and act upon the available data always play an important role.

there is a risk of overlooking the specific dynamics that affect low-income countries or late-adopters of LSLAs. This section elaborates on one particular group of constraints specific to developing countries – namely, limited ownership. The limited use of assessment results can never be exclusively attributed to a lack of political will. Rather, limited *opportunities, support, resources or capacity* of key education stakeholders to access, interpret and act upon the available data always play an important role. This is particularly true for less-wealthy countries where the current capacity to integrate national and cross-national initiatives to education systems tends to lag behind due to limited institutionalization (UIS, 2018c).

Consequences of limited ownership

In many countries, the lack of alignment between data production and data use is also a direct consequence of the limited ownership of assessment data. It is feasible to venture that, if countries do not sense that they hold some power over the assessment process, then they will feel less inclined to use the data and act upon the evidence. This lack of ownership is in turn the result of countries' limited participation and influence on different parts of the assessment process, including funding, design, data management and the dissemination of results.

Externally financed assessments. A general trend of *externally* financed learning assessments is observed for both national and cross-national assessment, and particularly, in developing countries. For example, at least two thirds of national learning assessments are funded with the support (in part or fully) by foreign or external donors (UIS, 2019a), such as the GPE, UNICEF, USAID and the World Bank.[39] For PASEC 2014 and beyond, the CONFEMEN funds half of the cross-national assessment while the government will need to finance the rest to participate. Repercussions of externally financed assessments could account for the uneven and sporadic nature of assessments in many countries. More worrisome still is that externally funded assessments could contribute to the lack of alignment between national assessment and other components of the education system, which paves the way to limited national ownership of the assessment results.

Controlled assessment design and management process. When assessments are outsourced to non-government organizations or international agencies, countries bear limited responsibility and exert little control over the design of assessment instruments and the management of the resulting data –particularly as country representatives and experts provide limited input on the development of assessment instruments. This is the case for cross-national assessment initiatives where the lead agency along with a data contractor prepare the participating country datasets in

---

39 In North America and Western Europe, no assessment is supported by external funds. A small proportion of Latin American countries receive external support for their assessments (11%) but rates are higher in other regions: Asia-Pacific (30%), Eastern Europe (30%), sub-Saharan Africa (36%), and the Arab States (50%) (Clarke, 2017).

a standardized way before they are shared with national teams for their own analysis. While such arrangements help to ensure that the data have been prepared in a standardized way, managing organizations could be perceived as 'controlling' and as the 'owners' of the data.

Strict dissemination schedules. Also common to cross-national initiatives, countries may have little say in the data or the analysis that is disseminated about their country and the learning levels of the target population. They may also only be allowed to disseminate their national report after the release of the international comparative report (OECD, 2009; IEA, 2019). This is due to customary agreements between governments and agencies leading cross-national initiatives, which give the agencies the power to disseminate the results and interpretations of these results (regardless of whether countries are satisfied with them) to avoid a conflict of interest.[40]

> Limited ownership is not only likely to result in data under-use and policy inaction but might also prompt governments, education authorities and other actors to interpret or use these assessments in a misguided way.

Data ownership is less of an issue in national assessments as countries usually have control over the assessment strategy, design and analysis. However, even the design and development of testing instruments can be conducted (or largely influenced) by donors funding the assessment scheme and particularly for low-and middle-income countries. There is evidence that in some countries, the results of national assessments are not validated or disseminated appropriately, which diminishes their policy value (European Union, 2014). Regardless of the type of assessment, it is worth noting that limited ownership is not only likely to result in data under-use and policy inaction but might also prompt governments, education authorities and other actors to interpret or use these assessments in a misguided way. Lack of ownership might thus have a misleading effect as it has the potential to distort policy priorities and suggest ill-suited courses of policy action.

Over-using assessment data

In this section, we discuss risks related to the over-use and over-interpretation of assessment data – i.e. when there is an excessive focus, attention or credence given to assessment results. We address the potential of over-use dynamics to generate a misleading or diverting effect on the decisions and behaviors of a wide range of stakeholders and users of LSLAs – most notably, policy-makers, teachers and researchers. This section reflects first on the different formats, channels and agents of data dissemination and continues with a discussion on four different mechanisms through which assessment data encourages over-use dynamics – namely, incentivizing policy action oriented at massaging figures, encouraging uninformed policy borrowing, stimulating misleading statements of causality, and promoting the use of results-based funding schemes.

---

40 While countries may be given the option to suppress some variables before dissemination or publication of the data, this will only be done if it can be shown that this data "presents a risk to student, school, and/or teacher anonymity, or for technical errors that could not be resolved by data contractors" (OECD, 2018b, p. 1).

Multiple forms of dissemination

Once the LSLA data have been analyzed, the results or findings are summarized in the form of one or more dissemination products. These include analytical reports, policy briefs, media reports, press releases, assessment databases and even blogs and social media posts to reach wider non-technical audiences. These products are then disseminated in various ways. Of course, not all types of products are needed for the dissemination of assessment results.

For cross-national learning assessments, dissemination and communication of information is done in a systematic and relatively homogeneous way. The date of the release of results is set and the findings are kept under embargo until the day of release. Following that, new findings and thematic reports are released periodically, the database is made public and articles are featured in national newspapers and international blogs. In contrast, dissemination of results from national assessments is more heterogeneous in nature. Only one in four assessments will release a public database while one in three will be featured in the media (UIS, 2019a).

The role of the media

Regardless of the dissemination choices made by the concerned authorities, the influence of LSLAs is also mediated by their impact on public opinion and sentiment (Feuer, 2012; Addey, 2018; Yasukawa et al., 2017). This is largely the product of the media attention that LSLAs receive. Media in this instance refers to traditional print and broadcast media as well as digital and interactive media. High levels of media resonance and exposure allow LSLAs to shape the common understanding of what is valuable in education (Hamilton, 2017; see examples of such dynamics in Dixon et al., 2013; Steiner-Khamsi et al., 2017; and Waldow et al., 2014). As a result, what is valued (and funded) in an education system becomes shaped by the prevailing discourses of the most widespread and extensively disseminated cross-national learning assessments – i.e. PISA, TIMSS and PIRLS. While these dynamics are not inherently problematic, they may reinforce an over-reliance on assessment data – diverting attention from other relevant issues and potentially resulting in misleading inferences and statements of causality. However, further exploration is required to understand how different media contribute to the formation of public discourse around LSLAs (Hamilton, 2017) – particularly as the media are highly diverse in terms of target audience, who engages with them, their accuracy in reporting, their use of emotional language, etc.

> High levels of media resonance and exposure allow LSLAs to shape the common understanding of what is valuable in education.

Spurring policy change – with a focus on figures

As noted above, LSLAs are expected to contribute to the educational policy discourse by informing policy-makers' priorities and agendas. However, such policy-shaping effects pose significant risks if policy action is exclusively oriented to impact and improve the figures rather than engage in substantive, structural change.

Such dynamics are particularly likely to occur in the case of cross-national assessments. This is largely the product of the league table format in which results of these exercises are often disseminated. League tables have become a common mechanism in the media to compare educational conditions and performances, illustrate progress and highlight setbacks across and within countries. The presentation of results – with the highest-performing countries on the top and lowest-performing on the bottom – allows policy-makers to gauge the performance of their school-attending children in comparison with those in other countries. Thus, the league table rank serves as a proxy for national progress. Illustrative examples of the trend to improve figures include Australia's goal to be ranked in the top five PISA countries by 2025, Thailand's objective to reach the OECD average PISA performance by 2021 (Breakspear, 2014) and the Gulf States' goal to improve their rank in the PISA league table (UNESCO, 2017f).

By showcasing the leaders and those who are falling behind, this type of comparative data may increase pressure on governments to raise standards in the various dimensions that are associated with learning outcomes – even among the lower performing countries. Hence, cross-country comparisons can encourage the undertaking of policy initiatives aimed at improving outcomes. Empirical examples of such dynamics are manifold. For example, publicizing the results from TIMSS (1999) generated reform pressures in the United States and the United Kingdom (Gorard, 2001), and those from PISA (2000) highlighted large disparities in Germany which triggered major public debate on the need for reform.

> League tables encourage countries to set targets against measurable goals achieved by other education systems and are consequently likely to incite them to emulate or learn from policies and practices applied elsewhere.

While such trends are not inherently problematic, the analyses unfortunately tend to focus excessively on rankings and average scores instead of on the determinants that explain the differences in country performance (UNESCO, 2017a; Rutkowski and Rutkowski, 2016; Torney-Purta and Amadeo, 2013). The focus on rankings and scores can result in potentially diverting or misleading effects. This is particularly the case for policy changes that stem from (ill-guided) policy-borrowing dynamics that rely on a highly superficial and decontextualized understanding of the impact of policy interventions and their capacity to affect test scores.

Encouraging policy borrowing

There is growing evidence that, in recent years, some countries have implemented education reforms or policy initiatives based on their performances in the assessments (Fischman and Topper, 2017; Baird et al., 2016).[41] League tables encourage countries to set targets against measurable goals achieved by other education systems and are consequently likely to incite them to emulate or learn from policies and practices applied elsewhere (OECD, 2016a). In this sense, rankings and league tables have become important policy tools to accelerate change and innovation

---

41   There is also increasing evidence that governments use their participation in cross-national initiatives to legitimize or accelerate policy reform, including pre-established and ideologically-driven policy agendas that would otherwise be difficult to justify on purely technical grounds (see Raudonyte, 2019).

(Steiner-Khamsi, 2003). Thus, LSLAs may not only spur policy changes but also encourage policy borrowing practices, inciting countries to adopt practices and policies from top-performers or so-called league leaders, such as Finland or Shanghai, and resulting in 'educational pilgrimage' where these league leaders become popular travel destinations for educational policy-makers, teacher researchers or other players in the education system (Sahlberg, 2007; Addey, 2014).

As a result of this trend, concerns have been raised that while assessments are vital to drive educational development, there is a risk that educational systems and their partners will overly rely on tests to drive educational reforms. Some researchers have documented how such dynamics can have misleading effects among policy circles. Ultimately, they can prompt policy-makers to pick individual policies from high-performing countries with little regard for cultural context or overall policy coherence (Breakspear, 2014) or adopt politically attractive policies and quick solutions (Lingard and Sellar, 2013) rather than promote meaningful and sustainable policy learning with weighted considerations on the purposes of schooling. For instance, Gorur and Wu (2014) have documented how Australia's ambition to be featured among the top five countries in PISA has led the country to try to emulate East Asian top performers – even though knowledge on the specificities of such systems is limited and largely restricted to their averages score and ranking position.
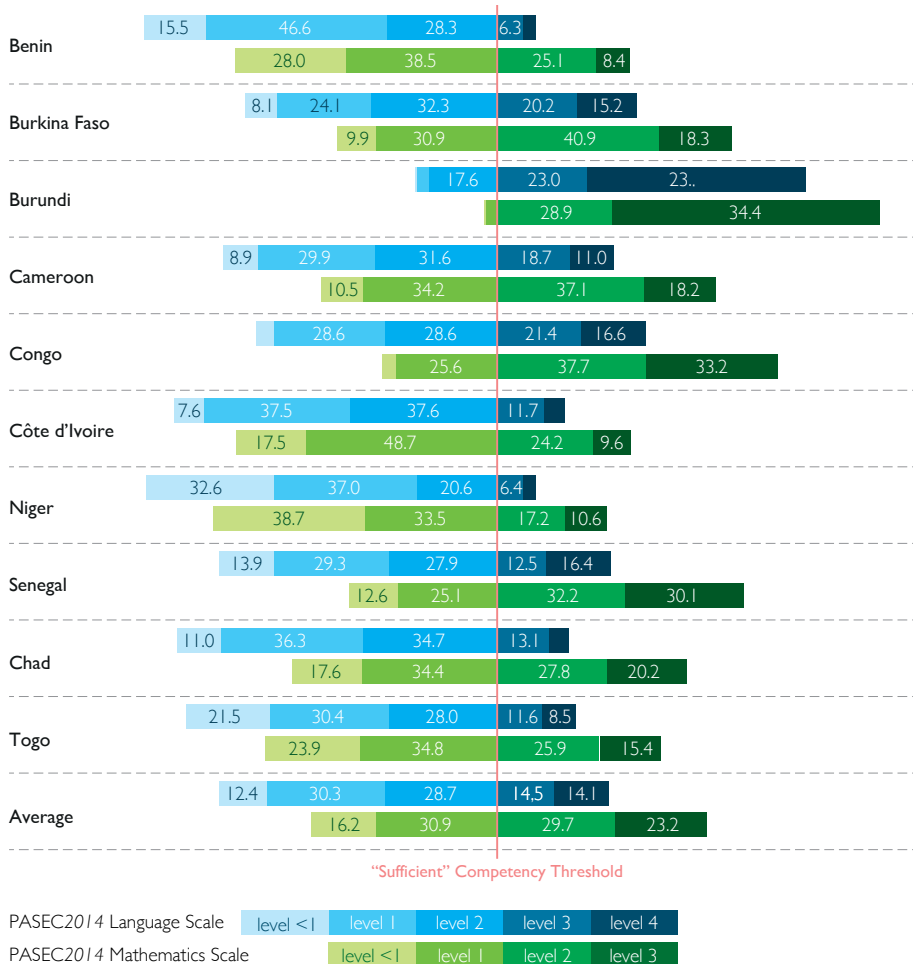
These dynamics are compounded by the fact that the use of education-related "big data" for evidence-based policy-making remains limited and uneven – partly due to the insufficient institutional capacity of some countries to analyze such data and link results to policies. As a result, these countries may rely on research findings from other countries and regions, even though they may not be relevant to their own education systems (UNESCO, 2017e). In this sense, it is heartening to see that some agencies, such as CONFEMEN, have recently started to present country results by alphabetical order – rather than by ranking them on the basis of their performance – a small change that can help neutralize these dynamics. For an example of alphabetical listing of learning achievement findings, please see **Figure 2**.

Research and causation

There has been growing interest in LSLAs within the academic community. This is reflected in the increased number of published opinion-based commentaries and peer-reviewed scholarly articles on LSLAs. The publications have been diverse, focusing on the implications of LSLAs, the use of data for national policy and further analyses based on empirical research (Fischman and Topper, 2017).

One area of study that has recently become more visible in international education policy discourse is research on the impact of education quality on economic growth as measured by average test scores on cross-national learning assessments. The topic is at the heart of heated debates among academics and international organizations. For example, Hanushek and Woessmann (2015) have argued that improvements in the skills of a population – as measured through improved learning outcomes in cross-national learning assessments from one cycle to the next – contributes to substantial GDP growth. This could be an attractive message with high impact that could lead to the belief that increased test scores will lead to economic gains. Other analysts, however, have noted that such a message can be considered excessively simplistic for two reasons. First, learning does not end at school – it is a ubiquitous process that keeps evolving

Figure 2. **Percentage of pupils, by competency level achieved in language and mathematics – early primary**

Benin
- 15.5 | 46.6 | 28.3 | 6.3
- 28.0 | 38.5 | 25.1 | 8.4

Burkina Faso
- 8.1 | 24.1 | 32.3 | 20.2 | 15.2
- 9.9 | 30.9 | 40.9 | 18.3

Burundi
- 17.6 | 23.0 | 23..
- 28.9 | 34.4

Cameroon
- 8.9 | 29.9 | 31.6 | 18.7 | 11.0
- 10.5 | 34.2 | 37.1 | 18.2

Congo
- 28.6 | 28.6 | 21.4 | 16.6
- 25.6 | 37.7 | 33.2

Côte d'Ivoire
- 7.6 | 37.5 | 37.6 | 11.7
- 17.5 | 48.7 | 24.2 | 9.6

Niger
- 32.6 | 37.0 | 20.6 | 6.4
- 38.7 | 33.5 | 17.2 | 10.6

Senegal
- 13.9 | 29.3 | 27.9 | 12.5 | 16.4
- 12.6 | 25.1 | 32.2 | 30.1

Chad
- 11.0 | 36.3 | 34.7 | 13.1
- 17.6 | 34.4 | 27.8 | 20.2

Togo
- 21.5 | 30.4 | 28.0 | 11.6 | 8.5
- 23.9 | 34.8 | 25.9 | 15.4

Average
- 12.4 | 30.3 | 28.7 | 14.5 | 14.1
- 16.2 | 30.9 | 29.7 | 23.2

"Sufficient" Competency Threshold

PASEC*2014* Language Scale: level <1 | level 1 | level 2 | level 3 | level 4
PASEC*2014* Mathematics Scale: level <1 | level 1 | level 2 | level 3

Source: PASEC, 2015.

through life experiences (OECD, 2010). Second, a country's economic outcomes depend on a host of variables outside the realm of academic achievement and educational attainment (Feuer, 2012). Thus, the claim that a causal relationship exists between academic achievement and a country's economic performance can be misleading. Researchers like Komatsu and Rappleye (2017) demonstrate that the correlation between the two is weak, suggesting that a causal relationship hardly exists. Similarly, Feuer (2012) has shown in the past that the claims of a relationship between educational outcomes and economic outcomes of countries lack sufficient empirical support. Further, they argue that assessment scores are not adequately valid data sources to forecast a nation's overall economic well-being.

One of the main roadblocks for research on the predictive powers of LSLAs lies in the difficulty to adjust for a number of confounding variables. For instance, results from LSLAs may be considered a good predictor to forecast enrolment and completion of higher education – as evidence from longitudinal surveys in different countries indicate, performance in PISA is a strong and consistent predictor of subsequent educational experiences (OECD, 2010). However, it also could be that better economic performance leads to an increase in educational participation and achievement. Alternatively, it could be that having more people with higher education leads to improved economic performance (Earle, 2010).

In sum, there is good reason to exercise more caution in the usage of causality language. In addition, as highlighted by Komatsu and Rappleye (2017), there is also a need for more rigorous global discussion on education policy and a more careful consideration for how statistics and data are used to inform policy-making. Similarly, the relative importance of assessment data vis-à-vis other sources of information constitutes another area that requires further reflection so that research is informed by a diverse corpus of evidence.

Results-based financing

The over-use of learning assessments can also take place within the development community. This is particularly the case when the quality and results of LSLAs are attached to the level of funding received at the national level from external donors. For instance, the Global Partnership for Education's (GPE's) 2016-2020 strategic plan, which is focused on improving learning and equity through stronger education systems, encourages the development of national assessment systems. GPE retains a concomitant focus on data and results and use of a results-based framework to measure progress and hold all members of the partnership to account (GPE, 2016). As such, GPE may stall funding if assessments are not conducted or if student performance is not shown to be improving. While such schemes hold some potential in terms of encouraging policy change, they also entail important risks. Most notably, results-based funding schemes can end in practices of "gaming the system" (UNESCO, 2017d). These practices include selective sampling, shaping of the testing pool or choosing the highest performing students to undertake the test to secure funding.

> Results-based funding schemes can end in practices of "gaming the system" [such as], selective sampling or shaping of the testing pool.

Coupling assessment with accountability

A number of the risks typically associated with LSLAs are indeed a product of their combination with (or subordination to) high-stakes accountability frameworks. This section explores the nature of the arrangements combining both these policy tools as well as the main risks and challenges associated with such schemes. The section ends with a note of caution on the relative importance of high-stakes accountability frameworks at the global level and is followed by a discussion on the need to gain a richer understanding of other varieties of LSLAs – most notably, sample-based and low-stakes schemes.

LSLAs as a component of accountability policies

LSLAs are generally conceived as low-stakes for the assessment takers. They differ in this way from national examinations that are used for selection, streaming or end-of-cycle certification purposes (UIS, 2019b). However, even if LSLAs are not high-stakes for students, they may have consequences that amount to higher stakes for other stakeholders, including teachers, principals, schools, or education administrations.

This is the case for so-called managerial or test-based accountability frameworks. This particular variety of accountability is characterized by a reliance on standardized-test data and a clear focus on results that generally equate to learning outcomes. In test-based accountability frameworks, certain stakeholders – teachers, principals and schools in particular – are held accountable for the performance of students relative to a set of pre-established standards and as measured by results in LSLAs (Verger and Parcerisa, 2017; Anderson, 2005; Hamilton et al., 2002). Stakeholders may be rewarded when the results are good yet penalized or sanctioned when the results are deemed unsatisfactory. The nature of the consequences attached to results varies greatly. Failure to attain minimal acceptable standards may result in the closing of a school, its transfer to new management and the firing or re-assignment of teachers and directors. High-performance, in turn, is encouraged through a number of material or organizational incentives, including monetary bonuses, promotion opportunities or an increase in school autonomy.

Perverse incentives when test scores become the central focus

The implications and unintended consequences of accountability dispositions attached to learning assessments have been extensively discussed in specialized literature – particularly in relation to high-stakes arrangements where assessment results have direct and clear consequences for teachers, principals and schools.

As suggested above, with explicit sanctions and rewards, test scores may become the central focus of schooling to the point that teaching time is focused on practicing for the test rather than on the standards-based content (Weiss et al., 2002). This may also represent a perverse incentive for schools and teachers who may engage in questionable practices to maximize test scores. For instance, schools may selectively sample 'good students' to bias the testing pool, narrow the curriculum, promote teaching to the test, only focus on students who are most likely to pass or even encourage explicit cheating (UNESCO, 2017d).

> With explicit sanctions and rewards, test scores may become the central focus of schooling to the point that teaching time is focused on practicing for the test rather than on the standards-based content. This may also represent a perverse incentive for schools and teachers who may engage in questionable practices to maximize test scores.

These practices are problematic for a number of reasons. First, they constitute a form of gaming the system that can have a misleading effect over monitoring efforts and may also generate inflated scores that are not representative of what students actually know (Koretz, 2017). Secondly, they are likely to end up undermining overall education quality and equity. In fact, even their effects on learning outcomes remain subject to debate. For instance, the 2017/18

Global Education Monitoring Report findings show that learning outcomes, as evaluated by performance on PISA between 2003 and 2015, have not improved in test-based accountability systems. Lastly, they may entail a de-facto or informal narrowing of the curriculum as knowledge, skills and competences as well as attitudes and values that are not the subject of testing tend to receive less attention.[42]

The case of merit-based pay

The dynamics of employing sanctions and rewards are particularly visible in the case of performance-based pay arrangements – a policy option that has indeed been gaining some traction over the last decade. In many countries, student test scores are increasingly incorporated to teacher evaluations (UNESCO, 2017d) and therefore, attached to their pay, promotion and employment status (Larsen, 2005). This has been the case, for instance, in the United States, Portugal, Chile and England, which have all experimented with different performance-related pay programs (UNESCO, 2014).

While this may inspire teachers to work hard to improve their students' results, teacher sanctions based on student test scores or evaluations can have multiple negative consequences for instruction, learning and equity (UNESCO, 2017d). According to the 2017/18 *Global Education Monitoring Report*, performance-based pay can promote an unhealthy competitive environment, reduce teacher motivation and encourage teaching to the test (UNESCO, 2017d). Also, it should be noted that the quality of teaching should not only be defined by achieved test results but also by teachers' values and principles; their professionalism; the extent to which they use the classroom and other available materials; or their ability to provide a positive learning environment and to teach according to individual student needs. As such, linking student performance to teacher salaries may lead to a very narrow set of requirements of what constitutes a 'good teacher.' Frustrated and demotivated teachers may then choose to leave their job, making it harder to recruit and retain talent in the teaching force. Such dynamics have been documented in the United States where test-based accountability is particularly well-established. As shown by Ryan et al. (2017), high-stakes assessments have a negative effect over teacher retention, mediated by stress and burnout levels.

A global phenomenon on the rise?

The use of LSLAs for accountability purposes can be found in countries around the globe and constitute a widespread and well-established trend. Some of the best-known and more consolidated examples include Australia, where LSLAs are used to publicly compare schools and reward teachers (Hardy and Boyle, 2011); the United States, where test results are used to close down schools whose students fall below state proficiency standards (Darling-Hammond, 2004); Chile, where school funding is affected by performance in learning assessments (UNESCO,

---

42 Evidence on these questions remains highly contested. Some argue that there is no evidence on the narrowing of the curriculum according to an analysis of textbook content (Ramirez et al., 2018). Others have noted that the effects of the new testing culture could be reflected in less obvious ways, including the distribution of instructional time among different areas of competence (Smith, 2019).

2017d); or, more recently, Tanzania, where a pilot project established performance-based pay for teachers and schools on the basis of test results (Mbiti et al., 2018).

Available data suggests, however, that in developing countries, most LSLAs are sample-based (UIS, 2019a) and are consequently ill-suited to inform high-stakes accountability frameworks. In addition, the spread of accountability arrangements, does not appear to advance in a linear or irreversible way (Verger et al., 2019). There is evidence that some countries have recently made an effort to decouple accountability efforts from LSLAs. This is the case for Denmark, which has taken steps to protect their students, teachers and schools from punitive measures linked to student performance in assessments (see **Box 6**). Together, these trends suggest that, although test-based accountability constitutes a globalizing phenomenon, it is not a totalizing one.

Box 6. Protecting students, teacher and schools from sanctions linked to performance

Denmark conducts a census-based large-scale assessment in various subjects between Grades 2 and 8. The test, which is computer-based and adaptive (therefore, no two students are taking the same test) is also one of a number of pedagogical tools available to teachers to provide diagnostic information, help evaluate student performance and design individual student plans.

The results are analyzed at individual student levels, for groups of students (aggregated by sex, age, socio-economic status or other), and at classroom, school, municipality and regional levels. They are available publicly only by subgroup at the national level. The data remains confidential and protected under a national act. By ensuring the confidentiality and anonymity of students and other types of identifying information, the act protects schools, teachers and students from potential punitive measures linked with students' assessment results.

Source: Based on Shewbridge et al., 2011; OECD, 2014.

Such observations have important implications for the debate around LSLAs. In a number of high-income countries, the spread and consolidation of standardized testing has been paralleled (or followed) by the adoption of high-stakes accountability instruments. In these cases, the appropriateness and risks of LSLAs have usually been discussed as a component of the accountability frameworks. Such an approach, however, is ill-suited to study and discussion on the role and impact of LSLAs in low- and middle-income countries. In addition, subordinating the debate on assessment to the debate on high-stakes accountability entails important risks when it comes to high-income countries. This holds true as some of the consequences entailed by the adoption of assessment instruments – rather than their use for accountability purposes – can be easily overlooked.

> Subordinating the debate on assessment to the debate on high-stakes accountability entails important risks when it comes to high-income countries

Low-stakes assessments and teaching to the test: A new debate

One of the implications of the increasing attention to high-stakes accountability frameworks is that it highlights our limited understanding and recognition of the consequences of low-stakes assessments. For instance, teaching to the test and curriculum narrowing have generally been studied as an unintended consequence powered by high-stakes examinations and not LSLAs with the primary purpose of providing feedback on the overall health of the system. Some researchers recognize that the drawbacks of teaching to the test inherent in the impact of LSLAs on policy may be exaggerated (Ritzen, 2013). However, others have argued that even when national assessments are low-stakes for the learner, they may still compel teachers to teach to the test due to the symbolic consequences of LSLAs and the perceived harm to school and teacher reputation. As mentioned, countries are increasingly using their overall rank in international assessment league tables such as PISA, as a national indicator of progress and improvement – or lack thereof.[43] This may create an incentive for education authorities to re-shape the curriculum or adapt teacher practices to the content or style of cross-national assessments in an effort to improve student performance. Similarly, national assessments not formally attached to material incentives for teachers and schools are sometimes purposely used to orient school choice and to encourage school competition. This is particularly likely to occur when school-level results are made available to the general public, sometimes under the form of league tables and

> Even assessments designed as low-stakes might eventually acquire higher stakes at a symbolical level, encouraging competition among schools or leading them to try to shape their testing pool.

rankings.[44] Thus, even assessments designed as low-stakes might eventually acquire higher stakes at a symbolic level, encouraging competition among schools or leading them to try to shape their testing pool.[45] In the end, a new debate emerges on whether sample-based cross-national assessments or low-stakes national assessments could also be influencing teaching and learning strategies.

---

43   See for instance Takayama (2008); Niemann et al. (2017); or Baroutsis and Lingard (2017) for an overview of the Japanese, German and Australian cases, respectively.
44   In fact, even when these arrangements are not in place, schools might use their results for promotional purposes.
45   These dynamics are more likely to unfold in school systems with high levels of school choice and organized along market dynamics. In such contexts, schools have a greater incentive to focus their efforts on improving the school's performance in standardized tests as such a reputation is likely to have a determinant effect on their enrolment (see for instance Falabella, 2014 for the Chilean case).

Concluding thoughts

The potential of large-scale learning assessments (LSLAs) as drivers of educational progress has been well-established by a growing corpus of specialized literature. Over the last decades, interest in LSLAs at national, regional and international levels has only grown – arguably, to the point of no-return. No matter what education systems may look like in the future, LSLAs are here to stay. Not only are an increasing number of countries undertaking LSLA initiatives, but the assessment market has also experienced a substantive transformation. An array of initiatives committed to supporting the implementation of assessments has flourished – with assistance from international and regional development agencies. The assessment landscape has thus become more complex and diverse in nature, and with such transformation, new challenges have emerged.

As argued here, LSLAs hold great potential but the benefits need to be weighed against the potential risks. Some of these risks stem directly from their design while others are the consequence of inappropriate uses of the results. While these limitations do not render assessments useless or inherently problematic, there is cause for more vigilance to avoid potential misuses. There is a need for a more careful and balanced design of tools and more tempered use of the results generated to appropriately inform policy actions.

In addition, it appears that LSLAs are increasingly taken for granted in most education systems. However, there is growing awareness that the ultimate impact of LSLAs essentially depends on *how* we use them – at least as much as on design and data collection methods employed. Yet, a number of questions concerning data use remain under-researched and insufficiently debated. In this section, we advance some of the areas of enquiry that would benefit from increased attention, reflection and research.

Living up to the assessment promise

Assessments in context. The purposes and uses expected to be served by LSLAs are manifold. However, understanding is limited on *how* and under *what conditions* they can effectively live up to their promise – i.e. how they can effectively inform policy change and/or practice to ultimately improve education quality and equity. It is thus necessary to refine our understanding of how contextual economic, institutional and political factors impact LSLAs.

Gaining this understanding is of crucial importance in light of the emerging evidence on the possible side-effects and unintended consequences of assessment. A growing body of research documents how LSLAs can encourage undesirable or even opportunistic behaviors among a range of stakeholders. Additional research is therefore needed to understand which assessments and conditions are likely to drive greater education quality and equity, and which ones are likely to backfire.

Beyond large-scale learning assessments. As noted throughout the piece, LSLAs may require some adjustment if they are to serve a quality and equity agenda. However, LSLAs cannot possibly serve *all* the purposes and objectives that an assessment system is expected to satisfy. Further reflection is needed on the role other modalities of assessment can play, including classroom-based assessments with formative and diagnostic purposes, portfolio evaluations and observation or peer evaluations. These assessments can indeed prove to be a better fit when it comes to measuring relevant learning areas that do not lend themselves to system-wide or cross-country comparisons. Greater attention to other forms of assessment is thus needed to identify the most appropriate ways to capture all the competencies that we deem valuable.

Similarly, while government- or inter-governmental-led forms of assessment have received considerable attention over the last decades, other emerging modalities of LSLAs remain less well understood. This is particularly the case for citizen-led assessments and household surveys that contain an assessment component. Interest in such topics remains largely confined to the specialized communities of practice in charge of them. Hence, the opportunities and risks these assessments entail rarely constitute the subject of mainstream scholarship. However, studying these alternative forms of assessment may offer ways to address or compensate for some of the limitations of conventional LSLAs. More precisely, they have the potential to ensure greater levels of equity and inclusion by incorporating additional languages and reaching remote populations or out-of-school children. Well-informed reflection on their potentials and limits could therefore be of use to advance the design of effective assessment frameworks.

The future of assessment

The potential limitations posed in the previous section must be addressed before LSLAs can be optimally employed in the service of education quality and equity. However, most of the challenges associated with LSLAs lie well ahead of us. Possible issues that deserve further reflection include:

The role of the private sector. The assessment landscape is populated by an increasingly diverse number of organizations – some of which remain scarcely understood and scrutinized. There is indeed an array of invisible actors playing an instrumental role in the assessment process. This is particularly the case for most contractors who undertake a large share of the assessment effort – yet are subject to relatively low levels of visibility and public scrutiny. Given the private status of many of them, such intermediary actors frequently remain a blind spot. The new assessment landscape thus needs to be charted with special attention to the role of the private sector.

A change in the learning environment. There is currently a limited understanding of how the nature of learning has evolved over recent years – and particularly of how learning has been affected by technological change. Background questionnaires that usually accompany assessment have paid limited and uneven attention to the learning practices of children and youth as well as their attitudes towards learning within and outside schools. Additional efforts are needed to incorporate these dimensions into background questionnaires so that assessment can provide a better understanding of the new and changing learning environments in which 21st century learners are embedded.

Ethics of assessment. The availability of an ever-expanding body of assessment data raises a series of ethical issues related to consent and transparency; data ownership; and safety and anonymity. While there is growing interest in the need for ethical standards for data collection and use, these remain at a nascent stage. Further research and reflection are needed to gain a better understanding of where the possible risks and challenges lie as well as clarify different levels of responsibility on these matters. It is also necessary to devise a shared understanding of good data-management practices that can adequately address the new challenges posed by the emergence of big data and artificial intelligence.

Assessment data as big data. Big data is indeed one of the most over-used yet least-understood terms of the current time. Even if it features prominently in public and specialized debates, it remains unclear whether assessment data falls within this category. The rise of computer-based assessments and adaptive testing as well as the growing reliance on algorithms to analyze and draw conclusions from assessment data has not been accompanied by reflection or debate on the implications of such trends for the assessed nor for education systems at large – including for our conceptions of education quality and equality. Gaining a better understanding of the big data/assessment data interface is thus a key step to ensuring that data remains at the service of education stakeholders – and not the other way around.

A broader and better understanding of such emerging challenges becomes all the more urgent in light of the rapid transformation underway in education and learning. We must heed the call for critical reflection and in-depth dialogue if we are to meet our collective commitment to ensure effective and relevant learning for all.

References

ACER-UIS. 2017. *The Principles of Good Practice in Learning Assessment.* Melbourne: ACER.

Addey, C. 2014. *Why do countries join international literacy assessments? An Actor-Network Theory analysis with case studies from Lao PDR and Mongolia*. UK School of Education and Lifelong Learning: University of East Anglia.

Addey, C. 2018. The assessment culture of international organizations: from philosophical doubt to statistical certainty through the appearance and growth of international large-scale assessments in C. Alarcón and M. Lawn (ed.), *Assessment Cultures: Historical perspectives.* Frankfurt am Main: Peter Lang.

Addey, C., and Sellar, S. 2019. Is it Worth it? Rationales for (non)participation in large-scale learning assessments. ERF Working Paper 24. *Education Research and Foresight (ERF) Working Papers 24.* Paris: UNESCO.

Addey, C., Sellar, S., Steiner-Khamsi, G., Lingard, B. and Verger, T. 2017. The rise of international large-scale assessments and rationales for participation. *Compare: A Journal of Comparative and International Education*.

Anderson, J. A. 2005. *Accountability in education*. Paris: UNESCO-International Institute for Educational Planning.

Baird, J. A., Johnson, S., Hopfenbeck, T. N., Isaacs, T., Sprague, T., Stobart, G. and Yu, G. 2016. On the supranational spell of PISA in policy. *Educational Research*, Vol. 58, No. 2, pp. 121-138.

Barnes, S., Lyonette, C., Atfield, G. and Owen, D. 2016. *Teachers' Pay and Equality: A Literature Review – Longitudinal Research into the Impact of Changes to Teachers' Pay on Equality in Schools in England*. Warwickshire, UK: Warwick Institute for Employment Research.

Baroutsis, A. and Lingard, B. 2017. Counting and comparing school performance: an analysis of media coverage of PISA in Australia, 2000–2014. *Journal of Education Policy,* Vol. 32, No. 4, pp. 432-449.

Barrett, A. M. 2016. Measuring Learning Outcomes and Education for Sustainable Development: the new education development goal in W. C. Smith, and W. C. Smith (ed.), *The Global Testing Culture; shaping education policy, perceptions, and practice.* United Kingdom: Symposium Books.

Beggs, C. 2016. Overview of EGRA and EGMA. Global Alliance to Monitor Learning. http://gaml.uis. unesco.org/wp-content/uploads/sites/2/2018/08/4_USAID_20160511.pdf (Accessed February 2019).

Benavot, A. and Köseleci, N. 2015. *Seeking Quality in Education: The Growth of National Learning Assessments, 1990-2013*. Paper commissioned for the EFA Global Monitoring Report 2015.

Benson, C. and Wong, K. 2015. Development discourse on language of instruction and literacy: Sound policy and Ubuntu, or lip service? *Reconsidering Development,* Vol. 4, No. 1, pp. 1-16.

Best, M., Knight, P, Lietz, P., Lockwood, C., Nugroho, D. and Tobin, M. 2013. *The impact of national and international assessment programmes on education policy, particularly policies regarding resource allocation and teaching and learning practices in developing countries*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Biesta, G. 2009. Good education in an age of measurement: On the need to reconnect with the question of purpose in education. *Educational Assessment, Evaluation and Accountability*, Vol. 21, No. 1, pp. 33-46.

Breakspear, S. 2014. *How does PISA shape education policy making? Why how we measure learning determines what counts in education*. Australia: Center for Strategic Education.

Brese, F., Jung, M., Mirazchiyski, P., Schulz, W. and Zuehlke, O. 2011. *ICCS 2009 User Guide for the International Database – Supplement 5*. Amsterdam: IEA Secretariat.

Brink. S. 2018. Learning from promising practices in ensuring effective use of large-scale learning assessment data for education policy and practice: Developing evidence from national and international LSLA for policy and practice. Presentation given at the TALENT Workshop, November 2018, Dakar, Senegal. http://www.education2030-africa.org/images/Developing_evidence_from_national_and_international_LSLA_for_policy_and_practice.pdf (Accessed February 2019).

Colclough, C. 2012. Education, poverty and development–mapping their interconnections. *Comparative Education,* Vol. 48, No. 2, pp. 135-148.

Clarke, M. 2012. *What Matters Most for Student Assessment Systems: A Framework Paper*. SABER Working Paper No. 1. Washington D.C.: the World Bank Group.

Clarke, M. 2016. Value for money: What do developing countries really get from participating in international assessment programs? Presentation 13 January 2016, Washington D.C. http://siteresources.worldbank.org/1818SOCIETY/Resources/Value_for_Money_Clarke_13Jan2016.pdf (Accessed May 2019).

Clarke, P. 2017. Making use of assessments for creating stronger education systems and improving teaching and learning. Background paper for the Global Monitoring Report 2017/18.

Cresswell, J., Schwantner, U. and Waters, C. 2015. *A Review of International Large-Scale Assessments in Education: Assessing Component Skills and Collecting Contextual Data*, PISA, The World Bank, Washington, D.C./OECD Publishing, Paris: OECD.

Darling-Hammond, L. 2004. Standards, accountability and school reform. *Teachers College Record*, Vol. 106, No. 6, pp. 1047-1085.

De Coster, I. and Sigalas, E. 2017. *Eurydice Brief: Citizenship Education at School in Europe 2017*. European Commission: Education, Audiovisual and Culture Executive Agency.

DFAT Australia. 2015. *Strategy for Australia's aid investments in education 2015–2020*. Australia: Department of Foreign Affairs and Trade.

DFAT Canada. 2014. *Canada's International Education Strategy: Harnessing our knowledge advantage to drive innovation and prosperity*. Ottawa: Department of Foreign Affairs, Trade and Development Canada.

DFID UK. 2018. *DFID Education Policy: Get Children Learning*. London: DFID UK.

Diaz, M. and Flores, G. 2008. SERCE, Resultados Nacionales [SERCE, National Results]. Mexico D.F.: Instituto Nacional para la Evaluación de la Educación.

Dixon, R., Arndt, C., Mullers, M., Vakkuri, J., Engblom-Pelkkala, K. and Hood, C. 2013. A Lever for Improvement or a Magnet for Blame? Press and Political Responses to International Educational Rankings in Four EU Countries. *Public Administration,* Vol. 91, No. 2, pp. 484-505.

Dolata, S. 2011. Are there any disparities between girls and boys in the response of the education sector of HIV and aids? Assessment of educational HIV/AIDS prevention programmes applied by SACMEQ III countries. Paper presented at the IIEP Policy Forum on Gender Equality in Education, 3-4 October, 2011, Paris, France.

Earle, D. 2010. *How can tertiary education deliver better value to the economy?* Tertiary education occasional paper 2010/08. New Zealand: New Zealand Government.

Engel, L. C. and Rutkowski, D. 2018. *Pay to play: what does PISA participation cost in the US?* Discourse: Studies in the Cultural Politics of Education. Routledge.

European Union. 2014. *Learning Assessments*. DEVCO B4 Education discussion paper. Brussels: European Union.

Falabella, A. 2014. The Performing School: The Effects of Market & Accountability Policies. *Education Policy Analysis Archives*, Vol. 22, No. 70.

Fenwick, T. and Edwards, R. 2014. Networks of knowledge, matters of learning, and criticality in higher education. *Higher Education*, Vol. 67, No. 1, pp. 35-50.

Ferrer, G. and Fiszbein, A. 2015. *What has happened with learning assessment systems in Latin America? Lessons from the Last Decade of Experience.* The Dialogue: Leadership for the Americas. Commission for quality education for all: background paper. Washington D.C.: World Bank Group.

Feuer, M. 2012. *No country left behind: rhetoric and reality of international large-scale assessment*. William Angoff Memorial Lecture Series. New Jersey: Educational Testing Service.

Fischman, G. and Topper, A. M. 2017. *An Examination of the Influence of International Large Scale Assessments and Global Learning Metrics on National School Reform Policies*. Mary Lou Foulton Teachers College: Arizona States University.

Fiske, E.B. 2000. *Assessing Learning Achievement. Education for All – Status and Trends 2000*. Paris: UNESCO, Paris. Education for All Forum Secretariat.

Friedman, T., Schwantner, U., Spink, J., Tabata, N. and Waters, C. 2016. *Improving Quality Education and Children's Learning Outcomes and Effective Practices in the Eastern and Southern Africa Region: Main Report*. Report for UNICEF ESARO. Camberwell, Victoria: ACER Press.

Fullan, M. 2010. Positive pressure. A. Hargreaves, A. Lieberman, M. Fullan and D. Hopkins (eds). In *Second international handbook of educational change* (pp. 119-130). Springer, Dordrecht.

Global Partnership for Education. 2016. *GPE 2020 Strategic Plan*. Washington D.C.: GPE.

Global Partnership for Education. 2018. *GPE 2020 Strategic Plan*. Washington D.C.: GPE.

Goldstein, H. and Thomas, S. 2008. Reflections on the international comparative surveys debate. *Assessment in Education: Principles, Policy & Practice*, Vol. 15, No. 3, pp. 215-222.

Gorard, S. 2001. International Comparisons of School Effectiveness: the second component of the 'crisis account' in England? *Comparative Education*, Vol. 37, No. 3, pp. 279-296.

Gorur, R., and Wu, M. 2014. Leaning too far? PISA, policy and Australia's "top five" ambitions. *Discourse: Studies in the Cultural Politics of Education,* Vol. 36, No. 5, pp. 647-664.

Greeney, V. 2008. *Assessing National Achievement Levels in Education*. Washington DC: World Bank.

Hamilton, L.S., Stecher, B.M. and Klein, S.P. 2002. *Making Sense of TBA in Education*. Santa Monica, CA: Rand Corporation.

Hamilton, M. 2017. How International Large-Scale Skills Assessments engage with national actors: mobilising networks through policy, media and public knowledge. *Critical Studies in Education*, Vol. 58, No. 3, pp. 280-294.

Hannon, V., Patton, A. and Temperley, J. 2011. Developing an Innovation - Ecosystem for Education. Indianapolis, CISCO.

Hanushek, E. A., and Woessmann, L. 2015. *The Knowledge Capital of Nations: Education and the Economics of Growth*. Cambridge, MA: MIT Press.

Hardy, I. and Boyle, C. 2011. My school? Critiquing the abstraction and quantification of education. *Asia Pacific Journal of Teacher Education*, Vol. 39, No. 3, pp. 211-222.

IEA. 2015. TIMSS 2015 Context Questionnaires. https://timssandpirls.bc.edu/timss2015/questionnaires/index.html (Accessed May 2019).

IEA. 2016. PIRLS 2016 Context Questionnaires. https://timssandpirls.bc.edu/pirls2016/questionnaires/index.html (Accessed May 2019).

IEA. 2019. National Study Reports. http://pub.iea.nl/national_reports.html. (Accessed May 2019).

Kamens, D. and McNeely, C. 2010. Globalization and the Growth of International Educational Testing and National Assessment. *Comparative Education Review*, Vol. 54, No. 1, pp. 5-25.

Kellaghan, T. and Greaney, V. 2001. *Using assessment to improve the quality of education*. Paris: UNESCO-IIEP.

Kellaghan, T., Bethell, G. and Ross, J. 2011. *National and international assessments of student achievement. Guidance Note*. A DFID Practice Paper. London: Department for International Development (DFID).

King, K. 2017. Lost in translation? The challenge of translating the global education goal and targets into global indicators. *Compare: A Journal of Comparative and International Education*, Vol. 47, No. 6, pp. 801-817

Komatsu, H. and Rappleye, J. 2017. A new global policy regime founded on invalid statistics? Hanushek, Woessmann, PISA, and economic growth. *Comparative Education*, Vol. 53, No. 2, pp. 166-191.

Koretz, D. 2017. *The testing charade. Pretending to make schools better.* Chicago and London: The University of Chicago Press.

Larsen, M. A. 2005. A Critical Analysis of Teacher Evaluation Policy Trends. *Australian Journal of Education*, Vol. 49, No. 3, pp. 292-305.

Lingard, B. and Grek, S. 2007. T*he OECD, indicators and PISA: an exploration of events and theoretical perspectives*. ESRC/ESF Research Project on Fabricating Quality in Education Working Paper 2. University of Edinburgh.

Lingard, B., and Sellar, S. 2013. 'Catalyst data': Perverse systemic effects of audit and accountability in Australian schooling. *Journal of Education Policy*, Vol. 28, No. 5, pp. 634-656.

LMTF (Learning Metrics Task Force). 2013. *Toward Universal Learning: Recommendations from the Learning Metrics Task Force*. Montreal and Washington, D. C.: UNESCO Institute for Statistics and Center for Universal Education at the Brookings Institution.

Lockheed, M., 2015. *Why do countries participate in international large-scale assessments? The Case of PISA*. Policy Research Working Paper. World Bank Group.

Mbiti, I., Romero, M. and Schipper, Y. 2018. *The Challenge of Designing Effective Teacher Performance Pay Programs: Experimental Evidence from Tanzania*. Working Paper. http://barrett.dyson.cornell.edu/NEUDC/paper_613.pdf

Ministry of Education, Finland. 2006. *Education Strategy for Finland's Development Cooperation*. Finland: Ministry of Education.

Mitra, D., Mann, B. and Hlavacik, M. 2016. Opting Out: Parents Creating Contested Spaces to Challenge Standardized Tests. *Education Policy Analysis Archives*, Vol. 24, No. 31.

Montoya, S. 2016. *The Cost of Ignorance Revisited: A Reply.* https://norrag.wordpress.com/2016/02/08/measuring-learning-the-cost-of-ignorance/ (Accessed June 2017).

Niemann, D., Martens, K. and Teltemann, J. 2017. PISA and its consequences: Shaping education policies through international comparisons. *European Journal of Education*, Vol. 52, No. 2, pp. 175-183.

Nikel, J. and Lowe, L. 2010. Talking of fabric: a multi- dimensional model of quality in education. *Compare: A Journal of Comparative and International Education*, Vol. 40, No. 5, pp. 589-605.

Norwegian Agency for Development Cooperation (NORAD). 2018. Education Webpage. https://norad.no/en/front/thematic-areas/education/ (Accessed July 2018).

OECD. 2009. *PISA Technical Report*. Paris: OECD.

OECD. 2010. *The high cost of low educational performance: The long-run economic impact of improving PISA outcomes*. Paris: OECD.

OECD. 2011. *Improving access to PISA for students with disabilities and other special education needs*. Paris: OECD.

OECD. 2013. *Synergies for Better Learning. An international perspective on evaluation and assessment. OECD Reviews of Evaluation and Assessment in Education*. Paris: OECD.

OECD. 2014. *Education Policy Outlook: Denmark*. Paris, OECD.

OECD. 2015. Running in Hot Weather – Introduction. PISA 2015 test item. http://www.oecd.org/pisa/PISA2015Questions/platform/index.html?user=&domain=SCI&unit=S623-RunningInHotWeather&lang=eng-ZZZ (Accessed December 2018).

OECD. 2016a. *PISA Results in Focus*. Paris: OECD.

OECD. 2016b. PISA for Development Brochure. http://www.oecd.org/pisa/pisa-for-development/PISA-D_brochure_2016_ENG.pdf (Accessed August 2017).

OECD. 2017. *PISA 2015 Collaborative Problem-Solving Framework*. Paris: OECD.

OECD. 2018a. *The Future of Education and Skills: Education 2030*. Paris: OECD.

OECD. 2018b. Data Management Procedures. In *PISA for Development Technical Report*. Paris: OECD. http://www.oecd.org/pisa/pisa-for-development/pisafordevelopment2018technicalreport/PISA-D%20TR%20Chapter%2010%20-%20DataManagement%20-%20final.pdf (Accessed May 2019).

Parker, R. and Fraillon, J. 2016. *Southeast Asia Primary Learning Metrics (SEA-PLM) - Global Citizenship Domain Assessment Framework*. Melbourne: ACER.

PASEC. 2015. *PASEC 2014 Education System Performance in Francophone Sub-Saharan Africa: Competencies and learning factors in primary education*. Dakar: Programme d'Analyse des Systèmes Educatifs de la CONFEMEN.

People's Action for Learning (PAL) Network. 2019. Our growth. https://palnetwork.org/our-growth/ (Accessed February 2019).

Pizmony-Levy, O. 2013. *Back to the Future in International Assessments*. http://blogs.edweek.org/edweek/assessing_the_assessments/2014/04/back_to_the_future_how_international_large-scale_assessments_came_about.html (Accessed August 2017).

Pizmony-Levy, O., Harvey, J., Schmidt, W.H., Noonan, R., Engel, L., Feuer, M.J., Braun, H., Santorno, Rotberg, I.C., Ash, P., Chatterji, M. and Torney-Purta, J. 2014. On the merits of, and myths about, international assessments. *Quality Assurance in Education*, Vol. 22, No. 4, pp. 319-338.

Pizmony-Levy, O. and Green Saraisky, N. 2016. *Who opts out and why? Results from a national survey on opting out of standardized tests*. New York: Teachers College – Columbia University.

Ramirez, F. O., Schofer, E. and Meyer, J.W. 2018. International Tests, National Assessments, and Educational Development (1970–2012). *Comparative Education Review*, Vol. 62, No. 3, pp. 344-64.

Raudonyte, I. 2019. *Use of learning assessment data in education policy-making*. IIEP-UNESCO Working Paper. Paris: IIEP-UNESCO.

Ritzen J. 2013. International large-scale assessments as change agents. von Davier M, Gonzalez E, Kirsch I, et al. (eds). *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research*. Dordrecht, Netherlands: Springer, pp.13-24.

Rowe, E. 2017. Australia Country Case Study. Background paper for Global Education Monitoring Report 2017/18 on the practice of selective sampling in Australia.

Rutkowski, L. and Rutkowski, D. 2016. A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher*, Vol. 45, No. 4, pp. 252-257.

Ryan, S.V., von der Embse, N.P., Pendergast, L.L., Saeki, E., Segool, N. and Schwing, S. 2017. Leaving the teaching profession: The role of teacher stress and educational accountability policies on turnover intent. *Teaching and Teacher Education,* Vol. 66, pp. 1-11.

Sahlberg, P. 2007. Education policies for raising student learning: The Finnish approach. *Journal of education policy*, Vol. 22, No. 2, pp. 147-171.

Sayed, Y., Ahmed, R. and Mogliacci, R. 2018. The 2030 Global Education Agenda and the SDGs: Process, Policy and Prospects. In A. Verger, M. Novelli, & H. K. Altinyelken (eds.). *Global Education Policy and International Development (2nd edition)*. London/New York: Bloomsbury, pp. 185-208.

Shewbridge, C., Jang, E., Matthews, P. and Santiago, P. 2011. *OECD Reviews of Evaluation and Assessment in Education*. Paris: OECD.

Siddiq, F., Hatlevik, O. E., Olsen, R. V., Throndsen, I. and Scherer, R. 2016. Taking a future perspective by learning from the past – A systematic review of assessment instruments that aim to measure primary and secondary school students' ICT literacy. *Educational Research Review*, Vol. 19, pp. 58-84.

Singal and Sabates. 2016. *Access and learning are equally important for children with disabilities*. Global Partnership for Education website. https://www.globalpartnership.org/blog/access-and-learning-are-equally-important-children-disabilities (Accessed May 2019).

Smith, W.C. 2016. An Introduction to the Global Testing Culture. W. C. Smith (ed.). In *The Global Testing Culture; shaping educational policy, perceptions and practice.* Oxford Studies in Comparative Education Series: Symposium Books.

Smith, W.C. 2019. Evaluating the Global Testing Culture: A Response to Ramirez and Colleagues. *Comparative Education Review*, Vol. 63, No. 1, pp. 121-126.

Southeast Asia Primary Learning Metrics (SEA-PLM), SEAMEO, UNICEF and ACER. 2017. Assessment Framework. Revised draft document incorporating feedback from the Domain Technical Review Teams.

Steiner-Khamsi, G. 2003. The politics of league tables. *Journal of Social Science Education*. OJSSE 1-2003: Civic Education.

Steiner-Khamsi, G., Appleton, M. and Vellani, S. 2017. Understanding business interests in international large-scale student assessments: a media analysis of *The Economist*, *Financial Times*, and *Wall Street Journal*. *Oxford Review of Education*, Vol. 44, No. 2, pp. 190-203.

Swedish International Development Cooperation Agency (Sida). 2018. Sida education webpage. https://www.sida.se/English/how-we-work/our-fields-of-work/education/ (Accessed July 2018).

Tawil, S., Akkari, A. and Macedo, B. 2012. Beyond the Conceptual Maze: The notion of quality in education. *Education Research and Foresight (ERF) Working Papers 02.* Paris*:* UNESCO.

The Education Commission (The International Commission on Financing Global Education Opportunity). 2016. The Learning Generation - Investing in education for a changing world. *Comparative Education Review*, Vol. 61, No. 2, pp. 214-217.

Takayama, K. 2008. The politics of international league tables: PISA in Japan's achievement crisis debate. *Comparative Education,* Vol. 44, No. 4, pp. 387-407.

Tikly, I. and Barrett, A.M. 2007. *Education Quality; Research priorities and approaches in the global era*. Bristol: University of Bristol.

Tobin, M., Lietz, P., Nugroho, D., Vivekanandan, R. and Nyamkhuu, T. 2015. *Using large-scale assessments of students' learning to inform education policy Insights from the Asia-Pacific region*. ACER and UNESCO Bangkok: ACER.

Torney-Purta, J. and Amadeo, J.A. 2013. International large-scale assessments: Challenges in reporting and potentials for secondary analysis. *Research in Comparative and International Education*, Vol. 8, No. 3, pp. 248-258.

UN Secretary General's Global Initiative on Education. 2012 -2016. *Global Education First Initiative*.

UNESCO, UNICEF, Brookings Institution and the World Bank. 2017. MELQO. Overview. *Measuring Early Learning Quality and Outcomes*. Paris: UNESCO.

UNESCO Institute for Statistics. 2016a. *The cost of not assessing learning outcomes*. Information Paper No. 26. Montreal: UNESCO Institute for Statistics.

UNESCO Institute for Statistics. 2016b. *Understanding What Works in Oral Reading Assessments*. Montreal: UNESCO Institute for Statistics.

UNESCO Institute for Statistics. 2018a. The impact of large-scale learning assessments. Montreal: UNESCO Institute for Statistics.

UNESCO Institute for Statistics. 2018b. *Costs and Benefits of Different Approaches to Measuring the Learning Proficiency of Students (SDG Indicator 4.1.1)*. Information Paper No. 53. Montreal: UNESCO Institute for Statistics.

UNESCO Institute for Statistics. 2018c. *SDG4 Data Digest. Data to Nurture Learning*. Montreal: UNESCO Institute for Statistics.

UNESCO Institute for Statistics. 2019a. Database. http://uis.unesco.org/ (Accessed January 2019).

UNESCO Institute for Statistics. 2019b. Glossary. http://uis.unesco.org/en/glossary (Accessed January 2019).

UNESCO International Bureau of Education (UNESCO-IBE). 2013. *Learning in the post-2015 education and development agenda*.

UNESCO International Institute for Educational Planning (UNESCO-IIEP). 2017. *Planning education, building the future. Medium-term strategy 2018-2021*. Paris: UNESCO IIEP.

UNESCO. 2007. *Quality Education For All: A human rights issue*. Educational policies within the framework of the II Intergovernmental Meeting of the Regional Project in Education for Latin America and the Caribbean (EFA/PRELAC). Background document. Santiago de Chile: UNESCO Regional Bureau for Latin America and the Caribbean.

UNESCO. 1974. *Recommendation concerning Education for International Understanding, Co-operation and Peace and Education relation to Human Rights and Fundamental Freedoms*. In Records of the General Conference. Eighteenth Session. Paris, 17 October - 23 November 1974, Vol. 1. Resolutions. Paris: UNESCO.

UNESCO. 1996. *Learning: The Treasure within. Report of the International Commission on Education for the 21st Century (Highlights)*. Chair: J. Delors. Paris: UNESCO.

UNESCO. 2014. *EFA Global Monitoring Report 2013/4. Teaching and Learning: Achieving Quality for All*. Paris: UNESCO.

UNESCO. 2015a. *Rethinking Education: Towards a global common good?* Paris: UNESCO.

UNESCO. 2015b. *EFA Global Monitoring Report 2015: Education for all 2000-2015: Achievements and Challenges*. Paris: UNESCO.

UNESCO. 2016a. *Unpacking Sustainable Development Goal 4 - Education 2030*. Paris: UNESCO.

UNESCO. 2016b. *Global Education Monitoring Report 2016*: *Education for People and Planet.* Paris: UNESCO.

UNESCO. 2016c. *Incheon Declaration and Framework for Action: Towards Inclusive and Equitable Quality Education and Lifelong Learning for All*. Paris: UNESCO.

UNESCO. 2017a. *Analyzing and utilizing assessment data for better learning outcomes*. Paris and Bangkok: UNESCO.

UNESCO. 2017b. *A guide for ensuring inclusion and equity in education*. Paris: UNESCO.

UNESCO. 2017c. *Learning assessment at UNESCO: Ensuring effective and relevant learning for all*. Paris: UNESCO. http://unesdoc.unesco.org/images/0026/002603/260325e.pdf

UNESCO. 2017d. *Accountability in Education: Meeting our commitments*. Global Education Monitoring report. Paris: UNESCO.

UNESCO. 2017e. *Large-scale Assessment Data and Learning Outcomes: Linking Assessments to Evidence-based Policy Making and Improved Learning*. Bangkok: UNESCO.

UNESCO. 2017f. *SDG 4 Framework for Action: Symposium on Policy implications of learning assessments to enhance educational quality in the Gulf Cooperation Council States*. Doha: UNESCO.

UNESCO. 2018. *Building tomorrow's digital skills - what conclusions can we draw from international comparative indicators?* Working Papers of Education Policy No. 6. Paris: UNESCO.

Ungerleider, C. 2006. Reflections on the use of large-scale student assessment for improving student success. *Canadian Journal of Education*, Vol. 29, No. 3, pp. 873-883.

Unterhalter, E. 2019. The Many Meanings of Quality Education: Politics of Targets and Indicators in SDG 4. *Global Policy*, Vol. 10, pp. 39-51.

Unterhalter, E and Brighouse, H. 2007. Distribution of what for social justice in education? The case of Education for All by 2015. M. Walker and E. Unterhalter (eds.). *Amartya Sen's capability approach and social justice in education.* Palgrave Macmillan: New York, pp. 67-86.

USAID. 2011. *Education Opportunity Through Learning*. USAID Education Strategy. Washington D.C.: USAID.

Verger, A. and Parcerisa, L. 2017. Accountability and education in the post-2015 scenario: International trends, enactment dynamics and socio-educational effects. Background paper for the *Global Monitoring Report 2017/18*.

Verger, A., Parcerisa, L. and Fontdevila, C. 2018. The growth and spread of large-scale assessments and test-based accountabilities: a political sociology of global education reforms. *Educational Review*, Vol. 71, No. 1, pp. 5-30.

Verger, A., Fontdevila, C. and Parcerisa, L. 2019. Reforming governance through policy instruments: how and to what extent standards, tests and accountability in education spread worldwide. *Discourse: Studies in the Cultural Politics of Education*, pp. 1-23.

Wagner, D.A. 2011. *Smaller, Quicker, Cheaper Improving Learning Assessments for Developing Countries*. Paris: UNESCO-IIEP.

Wagner, D., Lockheed, M., Mullis, I., Martin, M., Kanjee, A., Gove, A. and Dowd, A. 2012. The debate on learning assessments in developing countries. *Compare: A Journal of Comparative and International Education*, Vol. 42, No. 3, pp. 509-545.

Wagner, D.A., Wolf, S. and Boruch, R.F. (eds.) 2018. *Learning at the bottom of the pyramid: Science, measurement, and policy in low-income countries*. Paris: UNESCO-IIEP.

Waldow, F., Takayama, K. and Sung, Y.K. 2014. Rethinking the Pattern of External Policy Referencing: Media Discourses Over the 'Asian Tigers'. PISA Success in Australia, Germany and South Korea. *Comparative Education,* Vol. 50, No. 3, pp. 302–321.

Walter, S. and Benson, C. 2012. Language policy and medium of instruction in formal education. B. Spolsky (ed.). *The Cambridge Handbook of Language Policy*. Cambridge: Cambridge University Press, pp. 278-300.

Weiss, I., Knapp, M., Hollweg, K. and Burrill, G. 2002. *Investigating the influence of Standards: A framework for research in mathematics, science and technology*. Washington D.C.: National Academy Press.

Winthrop, R. and Anderson, K. 2013. Can international large-scale assessments inform a global learning goal? Insights from the Learning Metrics Task Force. *Research in Comparative and International Education*, Vol. 8, No. 3, pp. 279-295.

World Bank. 2011. *Learning for All: Investing in People's Knowledge and Skills to Promote Development*. Washington D.C.: The World Bank.

World Bank. 2018. *Learning to Realize Education's Promise. The World Development Report 2018*. Washington, D.C.: The World Bank.

Yasukawa, K., Hamilton, M. and Evans, J. 2017. A comparative analysis of national media responses to the OECD Survey of Adult Skills: policy making from the global to the local? *Compare: a journal of comparative and international education*, Vol. 47, No. 2, pp. 271-285.

Yasunaga, M. 2014. *Non-formal education as a means to meet learning needs of out-of-school children and adolescent*s. Online publication: UNESCO/UNICEF.

Zadja, J. 2014. Understanding the quality debate in education. Conference paper. Faculty of Education and Arts, Australian Catholic University (Melbourne Campus).

Annex 1:

Cross-national learning assessments

INTERNATIONAL STUDIES						
Assessment	Test domains	Target groups	Implementing agency	Years conducted	Updated list of participating countries	Consolidated research and analysis – curated by implementing agencies
TIMSS TIMSS Numeracy TIMSS Advanced	Mathematics, Science	Grade 4 and 8 Grade 4 to 6 Students in the last year of upper secondary school	International Association for the Evaluation of Educational Achievement (IEA)	1995, 1999, 2003, 2008, 2007, 2011, 2015 Next study: 2019	http://timss2015.org/timss-2015/about-timss-2015/ http://timssandpirls.bc.edu/timss2015/international-results/advanced/timss-advanced-2015/about-timss-advanced/	https://timssandpirls.bc.edu/isc/publications.html https://iccs.iea.nl/resources/publications.html https://icils.acer.org/ http://pub.iea.nl/national_reports.html
PIRLS PIRLS Literacy	Reading Literacy	Grade 4 Grade 4 to 6		2001, 2006, 2011, 2016 Next study: 2021	http://timssandpirls.bc.edu/pirls2016/international-results/pirls/about-pirls-2016/pirls-2016-countries/	
ICCS	Civics and Citizenship	Grade 8		2009, 2016	https://iccs.iea.nl/cycles/2016/countries.html	
ICILS	Computer and Information Literacy	Grade 8		2013 Next study: 2018	https://www.iea.nl/icils	
LaNA	Reading and Numeracy	Grade 4 to 6		Piloting	*Aiming for developing countries, not ready for TIMSS or PIRLS*	

INTERNATIONAL STUDIES						
Assessment	Test domains	Target groups	Implementing agency	Years conducted	Updated list of participating countries	Consolidated research and analysis – curated by implementing agencies
PISA	Reading, Mathematics and Science Innovative domains Financial literacy (optional)	Pupils at age 15	Organisation for Economic Co-operation and Development (OECD)	2000, 2003, 2006, 2009, 2012, 2015, 2018	https://www.oecd.org/pisa/aboutpisa/pisa-2018-participants.htm	http://www.oecd.org/pisa/publications/ https://www.oecd.org/pisa/pisa-for-development/
PISA-D	Reading, Mathematics and Science	Pupils at age 15		2017	https://www.oecd.org/pisa/aboutpisa/pisa-for-development-participating-countries.htm	

REGIONAL STUDIES						
Assessment	Test domains	Target groups	Implementing agency	Years conducted	Updated list of participating countries	Consolidated research and analysis – curated by implementing agencies
PASEC[46]	Mathematics and Reading skills	Grades 2 and 6	Conférence des ministres de l'Education des Etats et gouvernements de la Francophonie (Conference of Ministers of Education of States and Governments of Francophonie) (CONFEMEN)	2014 Next study: 2019	http://www.pasec.confemen.org/wp-content/uploads/2015/12/Resume_Pasec2014_FR_BD1.pdf (see section Contexte, objectifs et méthodologie de l'enquête PASEC2014)	http://www.pasec.confemen.org/pasec2014-3/
PILNA	Literacy and Numeracy	Grades 4 and 6	Educational Quality and Assessment Programme (EQAP) of the Pacific Community (SPC)	2012 and 2015	https://www.eqap.org.fj/getattachment/Resources/Final-Regional-Report.pdf.aspx (see Executive summary)	https://www.eqap.org.fj/work/Assessment.aspx
PERCE	Mathematics, Science and Reading	Grades 3 and 6	Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (Latin American Laboratory for Assessment of the Quality of Education) (LLECE)	1997	https://unesdoc.unesco.org/ark:/48223/pf0000243532 (see section Introducción general)	http://www.unesco.org/new/es/santiago/education/education-assessment-llece/first-regional-comparative-and-explanatory-study/
SERCE				2006		http://www.unesco.org/new/es/santiago/education/education-assessment-llece/second-regional-comparative-and-explanatory-study-serce/
TERCE				2013		*http://www.unesco.org/new/es/santiago/education/education-assessment-llece/terce/documents/*

---

46   Before 2014, PASEC performed 35 national assessments in more than 20 francophone countries in Africa, Asia and the Middle East.

REGIONAL STUDIES						
Assessment	Test domains	Target groups	Implementing agency	Years conducted	Updated list of participating countries	Consolidated research and analysis – curated by implementing agencies
SACMEQ or SEACMEQ	Reading (1995-1999) Reading and Mathematics (rest of implementation cycles) HIV-AIDS knowledge in 2006-2010	Grade 6	Southern and Eastern African Consortium for Monitoring Educational Quality (SACMEQ/SEACMEQ)	1995-1999, 2000-2004, 2006-2010, 2012 - 2014	http://www.sacmeq.org/?q=sacmeq-projects/sacmeq-iv	http://www.sacmeq.org/?q=publications
SEA-PLM	Reading, Writing, Mathematics and Global citizenship	Grade 5	South East Asian Ministers of Education Organisation (SEAMEO), in collaboration with UNICEF and the Australian Council for Educational Research (ACER)	2018-2019	http://www.seaplm.org/seaplm/index.php/about/countries	http://www.seaplm.org/seaplm/index.php/capacity-building-2

Note: For the purposes of this research, EGRA and EGMA (see Annex 2) are *not* considered cross-national assessments because even though the assessment models administered in different countries rely on a common methodological framework, they are *not* intended to enable cross-country comparability.

Annex 2: National assessments with common methodological frameworks

Early grade reading assessments (EGRA) and Early grade mathematics assessments (EGMA). These assessments are oral assessments, administered face-to-face and generally in schools. Their results are often used to advocate for educational reform. They are designed to provide simple, low-cost measures of reading and numeracy. They were developed by the Research Triangle Institute (RTI International) – EGRA in 2006 and EGMA a few years later. The EGRA (Reading) assesses letter recognition, phonemic awareness, reading simple words, and listening comprehension. EGMA (Mathematics) assesses number recognition, comparisons, and ordering sets of objects. EGRA has been used by 30+ organizations, in 70+ countries and adapted for 120+ languages; while EGMA has been conducted in 22 countries and adapted for 24 languages (Beggs, 2016).

Citizen-led assessments. These assessments are generally led by citizens or civil-society organizations, rather than by governments. They are conducted in households rather than in schools and measure basic reading and numeracy skills. Citizen-led assessments can provide recurrent estimates of children's basic learning levels and (so far) tend to be similar in design and administration. Fourteen countries are implementing citizen-led assessments and convened through the People's Action for Learning Network. The movement was initiated in India (2005), and followed by Pakistan (2006), Kenya and Tanzania (2007), Uganda (2009), Mali (2010), Senegal (2010), Mexico (2013), Nigeria (2014), Bangladesh, Cameroon, Ghana and Mozambique (2016), and Nepal (2017). Combined, they have tested 7.5 M children in most vulnerable situations through 48 assessments in 30 languages and have enrolled the help of 600,000 volunteers (PAL Network, 2019).

The promise of large-scale learning assessments

Acknowledging limits to unlock opportunities

Education in the 2030 Agenda for Sustainable Development places new emphasis on equitable, effective and relevant learning for all with greater efforts being made to chart progress towards this goal. Central to this endeavor are large-scale learning assessments (LSLAs), which have gained prominence as a critical strategic component to monitor the results of educational processes. LSLAs have indeed spread worldwide in the past two decades and are now administered in more than half of the countries around the globe. The utility of LSLAs as drivers of educational progress is well-documented in an ever-growing body of specialized literature.

As the potential uses for this monitoring tool continue to grow, however, less is known about their limits. Concerns have been raised among a range of education stakeholders and researchers on the unintended consequences of LSLAs. The potential side effects stem from how they are designed and from the possible misuse of assessment data to drive education policy and practice. Ideally, an understanding of how they serve to further the education agenda should be balanced with an equal understanding of the potential pitfalls when used incorrectly. This publication contributes to the on-going global debate by reviewing what we do know about the positive aspects of LSLAs while clarifying the negative unintended consequences they may entail.

9 789231 003332