International Journal of Educational Development 96 (2023) 102684

Contents lists available at ScienceDirect

International Journal of Educational Development

journal homepage: www.elsevier.com/locate/ijedudev

The politics of good enough data. Developments, dilemmas and deadlocks in the production of global learning metrics

Clara Fontdevila [a,b]

[a] Department of Sociology - Universitat Autònoma de Barcelona, Avinguda Eix Central. Edifici B, 08193 Bellaterra (Cerdanyola del Vallès), Spain
[b] School of Education - University of Glasgow, St Andrew's Building, 11 Eldon Street,G3 6NHGlasgow, United Kingdom

ARTICLE INFO

Keywords:
Sustainable Development Goal 4
Learning metrics
UNESCO Institute for Statistics
Data production
International large-scale assessments
Global indicators

ABSTRACT

The indicator framework established for the monitoring of SDG4 is unambiguous on the need to advance in the production of global learning data. However, the production of SDG4 learning metrics has been riddled with technical and political difficulties. Drawing on a combination of documentary analysis, interviews and non-participant observation, this paper reconstructs the process of negotiation of the data suppliers, statistical routines and reporting standards necessary for the production of SDG4 learning metrics. The paper thus offers insight into the mechanics of global quantification, and on the transformative impact of such processes on the agendas and relationships of partaking organizations.

1. Introduction

Goal- and target-setting have a long history in the field of education, constituting today a key instrument of global governance and being both a source and a manifestation of increasing transnational interdependence (King, 2016; Mundy, 2010). The Sustainable Development Goal 4/Education 2030 agenda, adopted in 2015, adds to a long list of efforts in global coordination to support the universal right to education. Along with 16 other goals, SDG4 is integrated into the UN Sustainable Development Agenda adopted by the United Nations General Assembly and is largely conceived as the natural successor of both the Education for All agenda and the education-related Millennium Development Goals (Sachs-Israel, 2017).

One of the most significant shifts entailed by the SDG4 agenda is the focus on learning outcomes – a transformation sometimes referred to as *the quality turn* or *learning turn* and defined as an effort to transcend a focus on schooling and enrolment figures as key indicators of progress (Fontdevila, 2021; Sayed, Ahmed and Mogliacci, 2018). Accordingly, the indicator framework established for the monitoring of SDG4 is unambiguous on the need to advance in the production of global learning data – so that student achievement can be reported in an internationally comparable way. Up to five targets in SDG4 include one or more learning-related indicator, and 4 out of 12 global indicators require reporting on student learning, skills or knowledge.

Thus, with the advent of SDG4, the production of globally-comparable learning data has become an institutional priority for the UNESCO Institute for Statistics (UIS), designated as the custodian agency for most SDG4 global indicators. However, and despite broad agreement on the centrality of learning measurement, the production of SDG4 learning metrics has been far from straightforward and conflict-free. Difficulties encountered in the process have not been exclusively technical in nature, but also stem from the political obstacles inherent to the collective nature of indicator-making. Thus, the production of global learning metrics relies necessarily on data suppliers other than statistical offices and national governments – most notably, producers of cross-national assessments, whose relationship with the UIS is comparatively less institutionalized. Expectedly, reconciling the multiple expectations of these different data suppliers and international organizations has proven a considerable challenge for the UIS.

However, the process of indicator-production behind the making of SDG4 learning data remains an empirically under-researched process. To be sure, there has been some comment on the process by which

E-mail address: Clara.Fontdevila@uab.cat.

https://doi.org/10.1016/j.ijedudev.2022.102684
Received 1 November 2021; Received in revised form 12 September 2022; Accepted 4 October 2022
Available online 4 November 2022

learning-related targets were translated into quantitative indicators, and the effects that such transformation had on our understanding of education quality (Unterhalter, 2019; Wulff, 2020). Nevertheless, such accounts have tended to focus on the moment of indicator *selection*, while leaving unaddressed the very *production* of such data – that is, the negotiation of the data sources and methodological considerations driving data-collection and data-harmonization efforts. Indeed, the quantification labor performed by international organizations, and the mundane operations through which global datasets are constructed, remain an empirical gap.

In this light, this paper explores the process of production of the global indicator corresponding to Target 4.1, a commitment to '*ensure that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes*' (UNGA, 2017, p. 8). Known as Indicator 4.1.1, this metric refers to the '*Proportion of children and young people (a) in grades 2/3, (b) at the end of primary, and (c) at the end of lower secondary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex*' (UNGA, 2017, p. 8). This is in fact perceived by a range of education stakeholders as a crucial metric that, combined with completion rates, constitutes the basis for the SDG4 unofficial '*leading indicator*' – that is, the main benchmark against which the overall progress of the SDG4 agenda will be measured.

Drawing on a combination of documentary analysis, interviews and non-participant observation, the paper reconstructs the process through which this indicator has been attached to specific data suppliers, statistical routines and reporting standards. In so doing, the paper aims to gain insight into the mechanics of global quantification, and to understand the transformative impact of an inherently collaborative project on the agendas and relationships of the UIS and other partaking organizations. Overall, the paper finds that if the *learning turn* was a story of consensus-building, the quantification of learning targets is one of fragmentation – a process in which many of the ambiguities[1] strategically exploited in the negotiation of SDG4 have become readily apparent.

2. Trying and failing and trying again. Historical attempts at the harmonization of learning data

While cross-national learning data enjoys an unprecedented prominence in the SDG4/Education 2030 agenda, attempts at the production of globally comparable learning data have a long history within UN circles – in fact, it is possible to document a rich history of *unsuccessful* attempts to harmonize learning data at a global level. The magnitude of the challenge faced by UIS in relation to the SDG4 learning indicators needs to be understood in relation to UNESCO's erratic historical record in the collection and the harmonization of learning data.

It is important to bear in mind that, when it comes to the *production* of learning data, UNESCO's trajectory is a rather irregular one, with a number of one-off attempts with limited continuity over time. In fact, UNESCO's efforts in the production of learning-data date back to a twelve-country cross-national study conducted between 1959 and 1969 under the sponsorship of UNESCO Institute for Education in Hamburg (UIE, 1962). Later on, in 1992, UNESCO launched a project to monitor student achievement in selected countries – the Monitoring Learning Achievement initiative (MLA), implemented jointly with UNICEF. With a similar objective, the Southern African Consortium for Monitoring Education Quality (SACMEQ) was launched in 1995 with the support of IIEP-UNESCO, and tasked with monitoring student learning through the production of comparable reports (Cussó and d'Amico, 2005). In the case of UIS', one of its first inroads into the area of learning assessment was the creation of the Literacy Assessment and Monitoring Programme (LAMP), initiated in 2003 and oriented at measuring adult literacy. However, the project suffered from a number of political, technical and institutional setbacks, and 'never gained sufficient global prestige' (Addey, 2018, p. 400).

In the area of data *harmonization*, the UIS appears to be something of a late-comer.[2] Initiatives oriented at the harmonization (rather than the collection) of learning data include the Assessment of Learning Outcomes set in place in 2008 and the Observatory of Learning Outcomes established in 2011. However, such attempts at the production of global learning data found very limited success. This was very much a consequence of the reluctance to share certain data (e.g. test items) exhibited by the main global and regional organizations involved in learning assessments, along with the competition between the IEA and the OECD[3] – but also of the skepticism exhibited by some countries.

The establishment of the Learning Metrics Task Force (LMTF) in 2012 is probably the most visible and better-known example of UIS's early efforts to the harmonization of learning data. Envisaged as a multi-stakeholder partnership, it was co-convened by the Center for Universal Education (CUE) at the Brookings Institution and the UIS. The group organized three consultative processes and brought together up to 30 organizations that met on different occasions from July 2012 to September 2013, benefiting also from the contributions of nearly 200 experts. Importantly, the socialization effect brought about by LMTF meetings played an crucial role in constructing new and shared meanings and the legitimation of the assessment program, and in setting the parameters of the learning measurement debate (Fontdevila, 2020).

However, and despite the inclusive and pluralistic character of the LMTF, some have observed that the engagement of certain constituencies was frequently superficial or simply nominal and that, despite being presented as a collaboration between the UIS and CUE-Brookings, the latter was always fare more in control of the agenda than the former. Hence, and while the LMTF was central in revitalizing the debate around the harmonization of learning data and in creating a sense of common purpose among different agents involved in the measurement of learning, its legacy was somewhat tainted by the prominence of Brookings – which, to some extent, came at the expense of UIS's centrality in such debate.

Thus, the UIS transitioned into the post-2015 era without the necessary infrastructure or resources to produce globally-comparable learning data. At the same time, and given the unprecedented emphasis put on education quality and learning outcomes, with the adoption of the SDG4 agenda the production of such data became an inescapable responsibility. In particular, the production of the global indicator corresponding to SDG4 Target 4.1 became a *de facto*

[1] As I have argued elsewhere (Fontdevila, 2021), the process of consensus-building behind the negotiation of SDG4 often relied on word-tweaking or 'word-smithing' practices oriented at preventing the alienation of any constituency or country. This was the case of heavily contested issues such as the role of the private sector, but also of the use of learning outcomes as a measure of education quality.

[2] This cannot be disconnected from the fact that the production of globally comparable learning data only came to perceived as part of UIS' *core* monitoring mandate with the advent of SDG4. Hence, while one of the Education for All monitoring indicators approved in 2000 focused on the share of grade 4 students having mastered a set of learning competencies, such competencies were to be defined nationally (World Education Forum, 2000). Consequently, this EFA indicator did not entail a clear harmonization mandate for UIS.

[3] While IEA is widely recognized as a pioneer in the area of cross-national learning assessment, since the turn of the century its work has been overshadowed by the visibility of PISA – and, more in general, by the expansion of the OECD in the area of educational measurement. In consequence, and even if their testing programs differ in terms of target population and assessment approach, the OECD and IEA are perceived as rival organizations (cf. Morgan, 2011). It should be however noted that, according to different informants, such animosity or inter-organizational infighting has tended to fluctuate over time – eventually decreasing with the retirement of some key personalities (Fontdevila, 2021).

organizational priority for the Institute, and the production of Indicator 4.1.1. ended up acquiring a high-stakes quality for the UIS.

Importantly, what was at stake was not only the credibility of UIS within education circles, but also within the broader UN environment and the international statistical community. Thus, the UIS is accountable to the education agencies and Ministries of Education brought together by the SDG4 initiative, but also to the UN Statistical Commission and the Inter-Agency and Expert Group on SDG Indicators (IAEG-SDGs), responsible for the development of the global framework of indicators for the monitoring of SDGs (ECOSOC, 2017). Agencies responsible for SDG data, as is the case for the UIS, are expected to report periodically on the development of methodologies and data-collection tools. This process represents an important source of reputational pressure for the UIS and is perceived by the global education community and donor circles as a test of UIS' competence.

Taking into consideration these developments, this paper analyzes how the UIS has dealt with the challenges posed by SDG4 in terms of learning measurement – an area in which the UIS was not well positioned to affirm its leadership and expert authority. Specifically, the paper inquires first into the political and technical strategies mobilized by the UIS in order to produce the learning metrics required by the SDG4 monitoring framework, and then analyzes how this labor has impacted on the authority and credibility of the UIS within the global education field.

3. Opening the black box of indicator production: analytical considerations

In order to make sense of these dynamics, it is necessary to theoretically articulate the mechanics of global quantification as an object of study. This section introduces some conceptual remarks necessary to examine an oft-neglected aspect of the quantification labor performed by international organizations – namely, the *production* of indicators (i.e., the process through which specific data suppliers, statistical routines and reporting standards are selected and developed. The section also develops a methodological framework appropriate to the empirical challenges posed by the study of such processes, integrating the advances made by the so-called practice turn (Neumann, 2002) with the principles of thick description (Merriam and Tisdell, 2016).

3.1. The making of global indicators

The production of global metrics has significantly accelerated over the last two decades – accordingly, a profuse scholarship on global indicators has emerged. However, research examining global quantification practices has tended to focus on a particular subset of metrics – specifically, those reporting systems that rate, rank or categorize countries (see for instance Kelley and Simmons's (2014) or Cooley's (2015) work) – and has tended to focus on the impact of indicators over the monitored populations. As we will see in this section, such thematic focus has come at the expenses of an understanding of (*1*) the collective nature and the micro-social foundations of the processes of indicator-making; and (*2*) the effects of global indicators over the very organizations producing them. This section advances a series of concepts oriented at bringing to the fore such themes in a theoretically-informed way.

3.2. The production of global indicators as a collective endeavor

Literature on benchmarking practices has arguably tended to exaggerate the power and intentionality of the international organizations responsible for global indicators. It is thus implicitly assumed that international organizations operate as all-powerful agents deliberately instrumentalizing quantification practices as part of self-serving agenda. This totalizing understanding of the global quantification practices has, however, been called into question. Thus, Erkkilä and Piironen (2018) have recently noted that the rationales guiding the production of indicators are more complex than most accounts suggest – noting critically that the production of indicators is unlikely to be driven by a single, rational and well-defined motivation. Similarly, Dahler-Larsen (2014) has taken issue with the notion that indicators are guided by a consistent and coherent vision regarding their ultimate objectives. Such observations suggest that, far from being a neat process, the making of global indicators is a messy endeavor, driven by different and even contradictory rationalities.

It follows from this that it is necessary to unpack the making of global indicators. One strategy for doing so is to focus on the very *production* of global indicators. This notion was introduced by Davis, Kingsbury and Merry (2012) in their operationalization of the trajectory of global indicators and refers to the phase in which the conceptualization of the indicator is coupled to an already existing or newly created dataset. Importantly, the authors caution against placing excessive emphasis on the promulgators of a given indicator (that is, the organization in charge of its packaging and dissemination), since the production of an indicator involves a much wider and dispersed range of actors in charge of data-collection efforts (including international organizations, national statistical offices and NGOs etc.). The authors note that the collective character of indicator production is likely to translate into challenges associated with the difficulty of securing the cooperation and diligence of data suppliers, and with tensions and friction between different data sources.

A key point to bear in mind when examining processes of indicator production is the social labor behind the transformation of raw data into authoritative representations. This point has been raised by Espeland and Stevens (2008), who draw attention to the fact that indicators become authoritative as they 'move upwards'. Such views have more recently been echoed by Rocha de Siqueira (2017), who argues that producers of indicators are aware of the imperfection of their own datasets but do not see this approximate character as inherently problematic. The author has put forward the notion of *good enough data* in order to shed light on these dynamics. The concept captures the specificity of data recognized by both producers and consumers as imperfect (or even prone to error) but accepted as an authoritative source by virtue of their practicality or convenience.

3.3. Making sense of organizational effects

As noted above, much of the literature on global indicators has focused on the monitored populations or a loosely defined audience. Conversely, the effects of indicators on the organizations producing them have been less systematically examined. Freistein (2016) has recently put forward the notion of *organizational effects* – a concept that captures the impact of indicators on the operational logic of international organizations in charge of their creation or dissemination. The author thus notes that the creation of indicators is at least partially motivated by the need to signal and ascertain authority over certain issues.

Echoing Freistein's (2016) arguments, there is growing recognition that international organizations engage in the production of indicators as a strategy to position and (re)brand themselves. It follows from this literature that it is important to pay attention to both *external* and *internal* organizational effects. External effects refer to the fact that the production of global indicators affects the position of the promulgating organization vis-à-vis other international agencies (fostering relationships of competition or cooperation). These dynamics have for instance been captured by Cooley (2015), who argues that rankings are used by

both governmental and non-governmental international organizations as a means to assert their authority over specific issues.[4]

Internal effects capture the notion that the production of indicators also affects the relationships between different units operating within an organization. Such a phenomenon has for instance been documented by Arndt (2008), who draws attention to the impact of quantification labor on intra-organizational dynamics. In her discussion of the World Bank's governance indicators, the author notes that institutional reasons are one of the key drivers behind the rise of governance indicators and observes that the production of indicators constitutes a means to raise the status of particular units or divisions within a given organization (Arndt, 2008).

This study engages directly with the collective nature of indicator production, as well as with its organizational effects over institutions partaking in such a process. In this way, it contributes to a better understanding of such theoretical constructs by providing an informed illustration of the mechanics of global quantification – an area of study in which it is possible to identify a certain disconnect between theoretical and normative debates and empirical research.

3.4. Analytical strategy and data sources

This paper applies the conceptual and theoretical remarks outlined above to the study of the UIS' leading role in the production of SDG4 learning metrics. It focuses thus on the micro-social foundations of indicator-making processes, as well as on the effects of indicator-making on the interactions between and within organizations engaged in the production of data. This investigation is premised on the need to focus on tangible and observable instances of activity as key units of analysis – in line with the tenets of the so-called practice turn as applied to the study of international organizations. Introduced by the seminal works of Neumann (2002), and Adler and Pouliot (2011), the practice turn emphasizes the need to focus on the more mundane unfolding of global social action, and particularly favors greater empirical attention to the practices, routines and worldviews exhibited by international organizations.

In order to capture these micro-processes, the analytical strategy used in this investigation relies on the principles of thick description – that is, an idiographic, contextualized and detail-oriented approach, characterized by a sequence-focused reasoning style. Popularized by Geertz (1973), the notion of thick description was originally used to refer to interpretive work explicitly oriented towards capturing the insider views of the actors involved in a given process or phenomenon. However, and as noted by Merriam and Tisdell (2016), with the passage of time 'it has come to be used to refer to a highly descriptive, detailed presentation of the setting and in particular, the findings of a study' (p. 257). This approach is adequate given the preoccupation of the study with the chain of events through which the SDG4 learning metrics came into being, and the organizational dynamics and micro-sociological forces that shaped such processes.

In terms of data sources, the study relies on the triangulation of three main methods – namely, interview data, documentary analysis and non-participant observation. A corpus of 59 semi-structured interviews is the primary source of data. Interviewees were recruited through a purposive sampling strategy – a criterion-based approach oriented at capturing the perspective of all those individuals displaying a feature that the researcher judged to be of interest, rather than ensuring the representativeness of different groups (Ritchie, Lewis and Elam, 2003). The table below offers a breakdown of the pool of interviewees according to six basic categories.[5](Table 1).

Table 1
Pool of interviews – disaggregated by categories.

Interview group	Total
UN agencies: UNESCO (including IIEP and the UIS), UNICEF, UNHCR (UN)	15
Assessment producers (international organizations, NGOs, expert consortia) (LASS)	16
Multilateral and bilateral donors (DON)	12
Civil society organizations and non-governmental organizations (CSNGO)	5
Country representatives (CR)	3
Private sector and experts (PRI)	8
Total	*59*

Source: Author's elaboration.
Note: The bracketed acronyms correspond to the nomenclature used to quote interviewees.

Complementarily, the research relies on the analysis of a broad corpus of documentation collected over the course of the research in an iterative way, including research reports, position papers, public statements made by different stakeholders, blog posts published by negotiators, meeting agendas and evaluation reports. The collection of such documents had an iterative nature and did not follow a pre-determined protocol or search strategy. Hence, and on the basis of exploratory searches and pointers and recommendations given by interviewees, I conducted regular screening of a range of institutional portals such as the UIS web pages dedicated to the Global Alliance to Monitor Learning (GAML) and the Technical Cooperation Group (TCG), or the World Education Blog curated by the Global Education Monitoring Report (GEMR). A number of ad hoc searches were also conducted in response to new developments of the debate and to imprecisions in the accounts offered by the interviewees. These searches were also to test and substantiate the emerging analytical insight.

Finally, the study incorporated observation data in order to contextualize the insights gained on the basis of interview and documentary data. Specifically, non-participation observation was carried out on the occasion of three multi-day events directly related to the monitoring of the SDG4 agenda and the production of its associated metrics – namely (*1*) the 2nd Meeting of the Technical Cooperation Group on the Indicators for SDG4-Education 2030 (16–18 October 2016, Madrid); (*2*), the 3rd Meeting of the Global Alliance to Monitor Learning (11 and 12 May 2017, Mexico City); and (*3*), the 5th Meeting of the Global Alliance to Monitor Learning (17 and 18 October 2018, Hamburg).

4. The twists and turns of harmonizing learning data

This section is concerned with the first objective of the paper, namely, to gain insight into the political and technical strategies mobilized by the UIS to produce the learning metrics required by the SDG4 monitoring framework. The section shows that, in order to overcome the historical challenges associated to the creation of global learning data, the UIS relied on a transparency and inclusion strategy oriented at maximizing political legitimacy, in combination with a *bricolage* strategy oriented at combining pre-existing approaches to data harmonization. In order to document this process, the section reconstructs three main episodes corresponding to three of the major challenges faced by the UIS in its most recent attempt to harmonize

[4] Note that these observations apply primarily to organizations not having a formal mandate, and/or enjoying only limited legitimacy in the measurement of a given area. This is ultimately reflective of the emphasis of quantification literature on comparative and ranking -like format – which has partially come at the expense of the study on the development of the international statistical system, and UN statistical activity in particular (for an exception, see Ward, 2004). Hence, the quantification practices of UN agencies with a *formal mandate* in the measurement and monitoring of specific themes remain largely understudied.

[5] Note that, for the purposes of this table, representatives of IOs and NGOs selected on account of their role in the production of learning data are included in the *Assessment producers* category (LASS) rather than in the UN and CSNGO categories.

global learning data – namely, the creation of a legitimate infrastructure, the selection of data suppliers, and the design of an alignment strategy allowing to equate existing learning assessments.

4.1. The creation of a legitimate infrastructure

As the negotiation of the SDGs progressed, the UIS started to anticipate the need to venture into the production of globally comparable learning data. Thus, in April 2014, the UIS and the World Bank convened a meeting with the objective of defining criteria to monitor reading in primary education. This meeting eventually gave rise to the creation of the Learning Metrics Partnership (LMP), a joint initiative of the UIS and the Australian Council for Educational Research Centre for Global Education Monitoring (ACER-GEM)[6] supported by the Australian Department of Foreign Affairs and Trade, and which aimed to establish a common scale against which existing assessment could be plotted.

However, the LMP initiative was short-lived – a change in the leadership of the UIS in 2015 precipitated an early termination of the project. According to different interviewees, the incoming UIS Director decided to put an end to the partnership with ACER in order to reaffirm the centrality of the UIS and its visibility in the creation of globally comparable learning data. At the same time, the move was also prompted by reservations expressed by some representatives of regional assessments, who perceived the initiative as excessively top-down in nature. A former UIS analyst noted that the LMP had fallen apart after a particularly contentious meeting in which regional assessments perceived that they had been unduly sidelined (Interview UN7). Finally, the limited support for the LMP within the learning measurement community was also a consequence of the leading role played by ACER in this venture. Some LMP participants were under the impression that the LMP was being used by ACER as a 'shop front' to market its services, and that ACER was instrumentalizing global targets to create a niche of its own in the assessment market – especially after its central role as a PISA contractor came to an end after the 2012 survey (Interview LASS10).

It is precisely as a response to the failure of the LPM venture that the Global Alliance to Monitor Learning (GAML) was created under the auspices of the UIS. Announced in late 2015, it was devised as a group bringing together a wide range of stakeholders (most notably assessment agencies, education-related international organizations and NGOs, and national authorities) with the objective of reaching an agreement on the specific measures and tools necessary to monitor learning in the context of the SDGs (Montoya, 2015). Since its formalization in 2016, GAML has been meeting on a periodic basis.

The creation of GAML was largely welcomed by the education community on account of its inclusive character. The creation of a new space was portrayed by a number of interviewees as a positive development – an opportunity to start with a clean slate. This was especially true of those informants less well-versed in the decades-old infighting between the IEA and the OECD, and who were very critical of the paralyzing effect of such disputes. Additionally, the advent of GAML was perceived as creating more space for data producers other than large assessment consortia. As an NGO officer put it:

> It's sort of a breath of fresh air because the rest of us have been sitting at these tables for years now, literally! […] We've been sitting here for four years going around and around and around, with vested interests and new models, and not moving forward. (Interview CSNGO6).

In line with its self-professed open and participatory nature, GAML membership has, since the start, been open to any individual or organization interested in contributing to its work. However, the first GAML meetings were mainly attended by international agencies, development partners, research organizations and foundations with a global scope (UIS, 2016b). A wide range of individuals with different institutional affiliations were critical of the absence of country representatives at GAML meetings. In combination with such reliance on self-selection dynamics, the limited formalization of internal decision-making procedures soon came under the critical scrutiny of a number of members. Also, and as noted by Benavot and Smith (2020), donors have tended to 'dominate meetings and have played a significant role in directing the focus toward Target 4.1' (p. 254) – an observation that suggests that power asymmetries remain within GAML, despite GAML's commitment towards horizontality and inclusion.

These dynamics were particularly problematic for those attendees representing large international bureaucracies, and who were consequently less willing or capable of making snap decisions (Interviews DON6, CSNGO6, PRI3). In response to such concerns, and as a way to sensitize and familiarize countries with the outcomes of the GAML-led technical negotiation, since 2018 an explicit effort has been made to incorporate country representatives. Similarly, a governance structure has been progressively clarified and refined. According to the *Governance and organization* note issued in 2017 (UIS, 2017), the technical work would fall on thematic task forces, but the inputs and recommendation prepared by them should be discussed and endorsed by plenary meetings. This was paralleled by efforts to draw a clear separation between GAML's role (making recommendations) and the mandate of the Technical Cooperation Group (TCG)[7] (responsible for their discussion and adoption). Hence, it was made explicit that the Secretariat (served by UIS) was expected to prepare recommendations and report to the TCG – the group ultimately responsible for its discussion and approval (UIS, 2017).

Overall, an effort has been made to secure a more democratic environment – and despite some reservations, the move has proved successful in reinforcing both GAML's and UIS's authority. However, it is important to bear in mind that the democratization of GAML remains an unfinished process. Thus, in spite of the aforementioned governance innovations, some participants remain unclear on the real locus of decision-making. Particularly within civil society circles, it has been noted that despite the emphasis placed on deliberative forums and GAML's plenary meetings, the UIS frequently steers the debate so as to secure specific outcomes, and that much was being discussed behind closed doors (Interview CSNGO5). Some even perceived the GAML as operating as a rubber-stamping board in which the UIS was simply looking for a 'seal of approval' for previously negotiated agreements, and argued that the inclusion of country representatives essentially followed a tokenistic logic.

This episode suggests that the UIS has enhanced its legitimacy as a focal point for the negotiation of learning metrics by recourse to the maximization of inclusion and transparency as a means to overcome the limited confidence, misgivings and suspicion triggered by prior attempts at the production of learning data. At the same time, this democratization strategy pursued through the creation of GAML has not been without costs and might be preventing the UIS from making progress at the pace required by the international statistical community. This

[6] ACER is a not-for-profit, research-oriented organisation with a focus on education. Along with a handful of testing organizations and research centres (e.g. ETS), it represents one of the few contractors with the necessary expertise to cater to the needs of international organizations and consortia administering cross-national assessments. Historically, ACER has played a key role in the implementation and administration of large-scale assessments including OECD's PISA and IEA's TIMSS and PIRLS.

[7] A platform convened by the UIS (which serves as its secretariat), the TCG is tasked with the political mandate to develop and debate SDG4 thematic indicators. The TCG is composed of regionally-representative UNESCO Member States, and incorporates also representatives of civil society organization, and a range of multilateral partner agencies *i.a.* (UIS, 2022a).

explains the fact that, as noted by different interviewees and confirmed by non-participant observation, GAML meetings are increasingly scripted – with attendees being presented with a narrow set of options rather than invited to propose alternative routes. Ultimately, such dynamics are indicative of the fact that, while the democratic imperative and the inclusion expectations placed on GAML are key sources of legitimacy, if taken to extremes they can also operate as constraining elements that might hamper the UIS' ability to deliver its mandate – thus creating a perverse incentive for the UIS to preserve a formally democratic structure while moving the real *locus* of decision-making away from these democratic and participatory spaces.

4.2. The selection of data suppliers

Despite the legitimacy gains brought to the UIS by the GAML initiative, the production of globally comparable learning data continues to present a number of technical difficulties that have historically defied consensus. One of the most pressing challenges in this regard is related to the availability and quality of data sources. Such challenges include the limited coverage of existing learning data (i.e., the fact that for a large number of countries, data on learning outcomes remains absent), but also the limited comparability of existing metrics, and the existence of multiple datasets that require prioritization. It is important to bear in mind that, by the mid-2010 s, there were several cross-national assessments (CNAs) in place, but no consolidated methodology to equate and harmonize them. In fact, the CNA category encompass both regional and international assessments and, as captured by Table 2, is notoriously heterogeneous in terms of targeted domain and grades, design, sampling, methods for score estimation, etc.

The debate was further complicated by the discussion around the possibility of integrating data sources other than CNAs. While by the mid-2010 s a considerable number of countries had a National Assessment (NA) in place, this data is not particularly amenable to comparative purposes. At the same time, the UIS had important incentives to use NAs for global reporting purposes – including the fact that such data allows for greater coverage of Indicator 4.1.1; and that discarding NAs as a valid source of comparable data was a risky move since it runs against country-ownership principles. Ruling out NAs might hence create a perverse incentive for resource-constrained countries to invest exclusively in CNAs (which, despite their capacity-building potential, are unlikely to realize the full potential of learning metrics, especially when it comes to education planning and policy design) (UIS, 2016a).

Table 2
Overview of cross-national assessments.

Target potentially informed by the assessment	Grade or target population	Name of the assessment	Domain – Literacy (L) or Numeracy (N)
–	1	EGMA	N
	1	EGRA	L
		PASEC	**L, N**
	2	EGMA	N
		EGRA	L
		LLECE	**L, N**
	3	EGMA	N
4.1.1a		EGRA	L
		PIRLS/ePIRLS	*L*
		TIMSS	*N*
	4	*LaNA*	*L, N*
		PILNA	**L, N**
	5	**SEA-PLM**	**L, N**
		LLECE	**L, N**
		PASEC	**L, N**
4.1.1b	6	**SACMEQ**	**L, N**
		PILNA	**L, N**
		LaNA	*L, N*
	8	*TIMSS*	*N*
4.1.1c	15 y.o.	*PISA*	*L, N*
	14–16 y.o.	*PISA-D*	*L, N*
–	–	ASER, Uwezo (5–16 year-olds)	L, N

Source: Author's elaboration on the basis of Treviño and Órdenes (2017) and UIS (2016a).
Notes:

- **In bold**: regional assessments; *In Italics*: international assessments; Regular font: assessments of foundational skills and population-based assessments.
- Includes only those domains relevant for global reporting purposes.

Taking stock of these trade-offs, since 2016 the UIS has been toying with different options, and rather than favoring a particular course of action, has explored a variety of reporting and harmonization strategies, commissioning work to a wide range of parties. The profusion of reporting protocols, mapping exercises, prospective studies and concept notes produced over the last 5 years is testimony to the multi-pronged strategy pursued by the UIS.

One of the first options discussed in early 2016 was the establishment of a new cross-national assessment specific to a target population and learning domain, to be implemented in all countries – an idea originally pitched by a group of analysts affiliated with the Center for Global Development. Advocates of the single-test option portrayed the common assessment as a way of maximizing the robustness of a global dataset, equated to *ex-ante* comparability (Birdsall, Bruns and Madan, 2016). The proposal, however, found limited resonance within the GAML community. Part of the resistance faced by this proposal stemmed from the fact that it was not perceived as being sufficiently disinterested in nature. According to a range of interviewees, it was not entirely clear to what extent some of the original promoters of the idea were acting bona fide or driven by business interests.[8] Concerns about the commercial implications of a universal test were compounded by a sense of 'expert skepticism' regarding the feasibility of such an ambitious project. The idea of an *ex-novo,* single-test was largely perceived as not grounded in reality (inattentive to implementation costs and time challenges); and concerns were voiced on the risks on relying on a single tool.

Once the idea of a new universal test started to lose steam, the idea of selecting an existing CNA and extending it to new countries was vigorously pushed by some assessment agencies – the IEA and the OECD in particular. These organizations seized SDG4 as an opportunity to make inroads into the development realm and expand their portfolio of countries.[9] As summarised by a member of the GAML-SPC:

> *You know, particularly IEA was going* 'Well, we are already at the primary level. We're already all over the world. We should be the test'. *And then PISA, they were trying to jump in to say,* 'Oh well, we should be 4.1.c because we are at the end of secondary'. *So they were trying to occupy that space* […] *They were basically trying to argue,* 'Look, those are valid, reliable, long-standing international assessments. Why are we even debating it? These should be the tools' […] *So they were, I'd say, taking an arrogant approach. And they were trying to intimidate the others based upon their technical prowess.* (Interview DON6).

As the quote above suggests, the behavior exhibited by OECD and IEA representatives was largely perceived as opportunistic, particularly on the part of some civil society and non-profit organizations. This was compounded by the fact that, within the donor community, there was limited appetite for this approach. Such reluctance was partially driven

[8] While the veracity of these misgivings is beyond the object of this study, the reservations expressed by a variety of individuals as to the true intent behind the single-test idea, even if fuelled by mere hearsay, are revelatory in their own right and speak to the governance challenges that characterized the early days of GAML and that motivated the development of an increasingly refined governance structure.

[9] An endeavor already initiated through IEA's Literacy and Numeracy Assessment for Developing Countries (LaNA) and the OECD's PISA for Development (PISA-D).

by a perceived lack of alignment with country priorities, and especially by donors' image concerns and certain preoccupation on the reputational risks associated with being perceived as imposing a rigid testing program on recipient countries (Interviews DON7; PRI8).

Given the limited headway made by the CNA-extension option, in 2017 the UIS centered its efforts in the development of an ensemble-like strategy allowing for the combination of multiple data sources, including CNAs and NAs. According to different interviewees, the idea of a common scale, against which different learning datasets could by plotted gradually, became attractive within UIS quarters. Not only did the idea offer a way-out of the political divisiveness caused by those approaches privileging a limited set of CNAs, but it was also perceived as better aligned with the principle of country ownership and a reporting strategy at the service of countries' statistical needs – rather that the other way around (Interviews UN18, UN16, PRI4). The idea of a common scale not imposing a specific data source enjoyed the support of a variety of countries but also parties such as NGO and bilateral donors – to whom in-built comparability was far from a deal-breaker, and who perceived the strategy as more conducive to data actionability and usability.

Despite the broad support enjoyed by the common-scale proposal, the idea was initially received with skepticism on the part of international assessment producers. In particular, the IEA and, to a lesser extent, the OECD, insisted upon the value of in-built comparability and the technical superiority of CNAs. Representatives of these organizations argued vehemently that *ex-post* comparability was a problematic idea – and that the only way of producing reliable and accurate data was through the administration of a common test. As one OECD official put it 'To be absolutely blunt, the only way you could compare results internationally is if everybody takes the same test. There's no other way' (Interview LASS3).

Anticipating that such emphasis on technical robustness and rigor risked having a paralyzing effect, the UIS responded by emphasizing the need to avoid 'letting the perfect be the enemy of the good'. This way, the notion of *fit-for-purpose data* started to gain traction within GAML circles. In addition, and in order to appease the concerns of the leading data producers (and the IEA in particular), the UIS coined the so-called 'stepping stone' approach (Montoya, 2017). This strategy was supportive of a variety of courses of action and was crucial in securing a consensus among partners with different agendas and priorities. Hence, this approach supported three different options, each of them associated with a different temporal horizon:

- A short-term reporting strategy according to which countries are allowed to submit data of their own choice;
- A mid-term strategy oriented towards linking existing assessments to a common scale;
- Support for countries to join CNAs in order to create a critical mass of comparable data – an objective with an undetermined horizon.

Thus, by encouraging multiple streams of work rather than privileging a specific strategy, the UIS succeeded in breaking the deadlock – with CNAs, NGOs and donors alike green-lighting the UIS strategy (Crouch and Bernard, 2017; Montoya, 2017). In coherence with this all-encompassing approach, the UIS engaged in an effort to increase the availability of learning data *without privileging any source in particular*. Thus, the UIS launched a series of publications prepared by consultants and oriented towards supporting both the establishment of NAs and participation in CNAs. An example of this are the quick guides *Implementing a National Learning Assessment* and *Making the Case for a Learning Assessment* – conceived respectively as a hands-on guide aimed to develop a NA; and to initiate a policy dialogue around the need for large-scale learning assessments (UIS, 2017; UIS, 2018a).

Also in line with this flexible, pragmatic approach, the UIS has been actively advocating and working to create a global bank of test items – that is, a repository of test constructs crowdsourcing the items used in existing assessments. A concept note published in 2019 described the rationale and main features of the project, devised as a tool that would enable low- and middle-income countries to generate assessment data at a comparatively low-cost while allowing them to report on SDG 4.1.1. (UIS, 2019a). The project, currently in full swing as a collective endeavor benefiting of the input of multiple partners, is indicative of the UIS' willingness to reinforce its role as a provider of global goods and fortify its capacity-development function – by supporting countries in their efforts to strengthen national statistical systems.

Overall, these developments suggest that, as a means to overcome the political and technical impasses posed by the production of learning data, the UIS has relied on a *bricolage* strategy consisting of recombining a number of already available and legitimate models, recognizing the limitations of each approach and emphasizing the potential for complementarity. In other words, the UIS has accommodated the multiple (and contradictory) demands placed on the organization by resorting to a hybrid approach that maximizes data-source flexibility while recognizing the added-value of cross-national assessments.

4.3. Bypassing the linking debate

This *bricolage* strategy has, in fact, continued to drive UIS' more recent efforts in the learning measurement realm – with the so-called 'linking debate' being one of the most illustrative examples of such approach. Hence, by the end of 2017 a consensus had been attained regarding the convenience of using a common scale on which different assessments could be mapped, and during 2019, this common metric, known as Global Proficiency Framework (GPF) had been developed. GPF was oriented at articulating 'the minimum knowledge and skills that learners should be able to attain along their learning progressions at each of the targeted grade levels in the two subject areas [reading and mathematics]' (GAML, 2019, p. 3). Importantly, and as remarked by Smith and Benavot (2021), organizations such as USAID were over-represented in the group behind the development of GPF.

In any case, while agreement had been reached regarding the content and the minimum proficiency levels, there was still a need to decide on an alignment strategy (i.e., a procedure to equate existing assessments to the common scale). Hence, the consensus on the GPF was paralleled by a new debate relative to the linking strategy. This soon proved to be a challenging endeavor for the UIS since, once again, different organizations favored different options and insisted upon the technical and/or political superiority of their own alternative. Discussed over the 2018–2020 period, such strategies not only differed in terms of technical complexity and financial cost, but also in relation to more politically-sensitive issues, including their item- and data-sharing implications, their potential to effectively inform education planning and contribute to country capacity- building and, more importantly, the possibility to integrate NAs (cf. UIS, 2019b).

Leaving aside the technicalities of the debate, the crucial point here is that, once again, such trade-offs (and in particular the possibility of using NAs) have been strategically mobilized and exploited by the different participating organizations. Thus, ACER has repeatedly called for the need to devise a system able to incorporate NAs, highlighting the benefits of such an approach in terms of coverage, country ownership and capacity-building – and arguing in favor of a less orthodox approach to comparability. Conversely, more heterodox methods such as policy-linking and item-based linking approaches have been harshly criticized by IEA and OECD specialists, who cast doubt on the validity and reliability of this approach, and remain skeptical on the possibility of using NAs for global reporting purposes (Fontdevila, 2021). The following excerpts express the contempt with which an analyst of a major assessment organization perceive policy-linking and item-based linking:

> What is being proposed, or what was proposed early on and still being talked about is, 'Oh, we could use national assessments'. I'm

> sorry, that ain't going to work. The national assessment in Honduras, those results cannot really be compared to the results of national assessment in Australia, that's just technically not feasible. But there are lots of people who argue that 'No, you can draw up a common scale'. You can take these assessment items and you can equate them and link them. Okay, theoretically that could be done, but it is highly problematic. (Interview LASS3).

While the difficulty of reaching an expert consensus on the optimal linking strategy risked having a paralyzing effect, the UIS has nurtured again a hybrid or 'hedge-betting' approach, insisting upon the fact that the alternatives are not mutually exclusive and can even reinforce one another. Thus, the UIS has supported and commissioned technical work in relation to different options. Referred to as the portfolio approach, this strategy emphasizes the idea that the different options 'should be taken more as complementary routes than as alternative options in order to minimize risk if some of the approaches prove to be too costly, the margin of error is too high, politically-unfeasible or a combination of these issues' (UIS, 2018b, p. 19). This approach is also considered more respectful of countries' priorities and context-specificities – in that provides Member States with a menu of options. The rationale behind this pragmatist turn is captured by a GAML interviewee:

> Silvia [Montoya, UIS Director] shifted into full pragmatics mode, and just said, 'I don't have to pick a winner. So what we're going to have instead is a portfolio of options. So some countries want to do the Rosetta Stone, fine' – so she is giving something to the IEA. 'And if they want to do the reporting scales, [she said] fine'. So in effect, she pulled rank and said, 'I'm not picking any of your models because it's for the countries to decide'. (Interview DON6).

At the moment of writing, the reporting strategy for Indicator 4.1.1, relies primarily on a consensus reached among major CNAs regarding the alignment of their respective proficiency levels aligned with the global minimum proficiency level.[10] However, the UIS also supports the development of a much wider spectrum of strategies. These include the so-called policy-linking method, which allows for the use of NA data; as well as the psychometric test-based linking method – which builds on the Rosetta Stone proposal advanced by the IEA and aligns existing regional assessments to IEA's TIMSS and PIRLS achievement scales items (UNESCO, 2021; UIS, 2022b). More recently, and in the context of the COVID-19 crisis, the UIS has also launched the Monitoring Impacts on Learning Outcomes (MILO) project in Africa, with the financial support of the GPE and ACER's technical input. The MILO project makes it possible to link existing regional and national assessments with a global proficiency framework. Importantly, the project builds upon and continuing to expand the aforementioned global bank of items, thus maximizing country ownership and capacity development (UIS, 2022c).

As a final note of caution, and despite the significance of such developments and the growing consensus regarding data collection, reporting and harmonization standards, it is important to bear in mind that such agreements retain an unstable and fluctuating quality, and its real strength will only be seen over the long haul. However, the objective of this paper is not to analyze the robustness or adequacy of the indicator production efforts – let alone to speculate how the process will play out in the future. In this sense, the absence of a final agreement is not judged to be an impediment but as an opportunity to set the focus on the structural dynamics shaping the process – rather than on an end-product whose specifics might simply reflect contingent circumstances.

[10] In this sense, and as observed by Smith and Benavot (2021), a certain hierarchy of assessments seem to be emerging – presumably, as a temporary solution. Hence, international assessments are prioritized over regional and national assessments, and population-based assessments are seen as a last resort. Yet, as noted by the authors, a number of interrogation marks persist – for instance 'it is unclear to what extent national assessments will be reshaped to meet the robust requirements for participation' (p. 213).

5. The UIS at a crossroads

The following section is concerned with the second objective of the paper, namely, to analyze how the production of SDG4 learning data has impacted on the authority and credibility of the UIS within the global education field. To this end, the section delves first into the process of organizational evolution through which the UIS managed to relatively enhance its authority within the learning measurement field; and then turns to the circumstances that place the UIS in an ultimately fragile position and that appear to compromise its centrality.

5.1. The power of an honest broker

As discussed in the prior sections, the relative success of the UIS in the area of learning measurement owes much to its democratization strategy – as well as to its efforts to avoid a zero-sum approach and accommodate the use of different assessments and harmonization methods, and to support countries in navigating the 'assessment market' (cf. Montoya and Crouch, 2019). Generally speaking, such strategies have been largely welcomed within the GAML community, for they have allowed actors to bypass a number of competitive dynamics and misgivings that, in the past, had hindered global efforts towards the harmonization of learning data.

In this sense, the UIS has made an effort to posit itself as an honest broker driven by a public-service *ethos*, convening different parties and interests and building consensus in a fraught arena riddled with vested interests. While not all assessment partners are equally enthused by the pragmatic turn pursued by the UIS, most GAML participants were appreciative of the UIS's role as a mediator driven by the common interest, its efforts to create collaborative and inclusive spaces, and the emphasis placed on the principle of country ownership. Such views are captured in the words of an informant long-involved in the GAML space:

> You cannot make progress in this work without involving organizations with high capacity. But then the question is how do you make sure that then the outputs of that do not privilege a particular organization? It's a really delicate balancing exercise. and I think the UIS has given the credentials that they are not really favoring any organization. I think they are trying to move with some people. But they also need to satisfy certain standards in terms of how they collaborate, and what they make public, and what their agenda is. And [it] is not that easy. But from that point of view, I think the GAML is trying to accommodate as many players as possible. (Interview UN17).

Remarkably, the Institute has managed to turn a potential liability (namely, the imperfect character of global learning data and the politicized nature of the process) into an asset – an opportunity to affirm its authority in the education measurement realm. Rather than casting the reporting process as a purely technical challenge (or emphasizing expert knowledge as the most relevant attribute of the organization), the UIS has brought to the fore the political nature of the debate, and has tapped into its aura of publicness, neutrality and commitment to the common good as a means to bolster its credibility.

Similarly, rather than addressing technical rigor and *ex-ante* comparability as supreme values, the UIS has emphasized the need to combine such principles with considerations for country ownership, and to accommodate a diversity of data sources. The emphasis on the need for *fit-for-purpose data* (as opposed to 'perfectly accurate' data) has ultimately contributed to reasserting UIS' centrality. This is so as the fitness-for-purpose criterion entails an element of judgement that cannot rely exclusively on technical considerations – a role for which the UIS is ideally suited on account of its aura of neutrality.

To summarize, it appears that when it comes to learning data, the UIS derives its authority not from an appearance of scientific objectivity or expertise, but from an explicit recognition of the ever-perfectibility of data, the necessarily provisional character of figures, and the political

nature of the measurement debate. Such framing has proved functional in that it has enhanced the centrality of the UIS in its role as honest broker and standard-setter, rather than a mere data curator.

5.2. A fragile position

Although there is a certain agreement that the learning measurement mandate associated with SDG4 has contributed to ensuring a much more central role for the UIS, this position appears to be fragile. Likewise, UIS' authority gains are more clearly or more explicitly recognized within specific segments of the education-for-development field – such as measurement experts and psychometricians within the GAML and TCG circles. Thus, UIS' new-found position is far from being a secure or consolidated one. Indeed, as this section will discuss, a number of intertwined factors appear to endanger the authority and centrality of the UIS.

The first concerns the UIS' limited *in-house* expertise in the area of learning measurement –which some deem a consequence of the combination of financial difficulties and/or UNESCO's administrative rigidity. The UIS has thus tended to rely on external expertise to carry out some of the technical work required to harmonize different assessments. This explains UIS's heavy reliance on organizations such as Brookings or ACER in the early days of the SDG4 indicator debate – even when such collaborations came at a reputational cost. It also explains the fact that many UIS publications and initiatives launched over the last few years have been prepared in conjunction with consultancy firms, research organizations and assessment consortia (e.g., IEA, ACER) and independent researchers. While these collaborations ensure a certain degree of technical sophistication and allow the UIS to comply with the tight timeframe put in place by the global SDG reporting mechanisms, they are also likely to turn into a double-edged sword in the long term. The 'vicarious expertise' acquired by the UIS by means of partnering with others may ultimately pose significant risks in terms of sustainability, and even legitimacy.

Secondly, the UIS's rapport with UNESCO appears to be an uncertain one. On the one hand, the institutional distance between the UIS and UNESCO has, to some extent, proved helpful to the UIS. On the other hand, UNESCO's alignment and support to UIS's renewed vision and strategy in the learning measurement area remain an open question to many. There is certainly an overall lack of clarity about the extent to which UNESCO has been supportive of the UIS' efforts in the learning assessment domain, and some interviewees observed that the UIS's efforts to gain visibility have created some friction within UNESCO (UN18). This lack of alignment between the UIS and UNESCO's priorities appears to have ultimately had a detrimental effect on the activity of the Institute – especially since it has resulted in the UIS not receiving the political and financial support necessary for the challenges entailed by the production of SDG4 indicators.

This brings us to the third factor – namely, the UIS' economic situation. An evaluation recently conducted by UNESCO's Internal Oversight Service (UNESCO, 2018) finds that the Institute has long been in a situation of economic distress, which has led to a significant downsizing of its staff. Importantly, the bulk of UIS funding comes from bilateral organizations and private foundations. To be sure, UIS's work on learning outcomes stands out as one of the areas in which the Institute has been more successful in terms of resource mobilization, securing considerable support on the part of the Bill & Melinda Gates Foundation (Montoya, Beeharry and Woolf, 2019) and being one of the main beneficiaries of a DfID initiative oriented at improving education statistics[11] (DfID, 2018). However, the overall lack of budget stability perpetuates the Institute's dependence on consultants to fill knowledge gaps or even to perform core tasks or mandates (e.g. leading in GAML). In addition, the need for the UIS to comply with donors' priorities and timeframes makes it increasingly difficult for the Institute to rely on consensus-building and GAML as the main decision-making strategies. In fact, the UIS has increasingly tended to rely on technical meetings (i.e., meetings bringing together donors and data producers) as a means to overcome political impasses. The reliance on such agreements is, however, difficult to reconcile with the emphasis on participatory procedures from which the UIS derives its legitimacy.

Finally, some measurement projects recently launched on the part of the World Bank could also endanger UIS' new-found centrality in the learning measurement arena. This is for instance the case of the Human Capital Index (HCI), a new composite indicator launched in 2018 that combines metrics relative to different dimensions of human capital. The education component of the index aims precisely at capturing education quantity and quality and has motivated the development of a methodology to harmonize assessment scores (Altinok, Angrist and Patrinos, 2018). In addition, the launching of a Learning Target and a Learning Poverty Indicator in 2019 have also resulted in the development of a new indicator combining measurements relative to school access and learning into a single figure (World Bank, 2019). Overall, the production of global learning data has gained considerable prominence in the World Bank organizational agenda – a phenomenon perceived as an instance of 'mission creep' within UNESCO and UNESCO-adjacent circles (PRI6). As a UNESCO-affiliated interviewee noted:

> I'm quite concerned. So to me, this [HCI] didn't bode well for the Bank and its role. It really came across as a pretty half-baked idea. What's the sustainability here? Who's all this for really? […] It doesn't seem to be in much of a leadership role, and my worry is it's like other development agencies, they are more concerned about having a branded product than actually having an impact. (Interview UN9).

Interestingly, the World Bank is not oblivious to the reputational risks associated with organizational overlap and has instituted a preemptive strategy to dissipate such fears – especially as duplications have come under public scrutiny (GEMR, 2018). Thus, in 2019, encouraged by some key donors, the World Bank and the UIS signed a Memorandum of Understanding conceived as a means to give some stability to inter-organization collaboration. While the impact of this partnership will only be seen in the long-term, such dynamics are ultimately indicative that the learning assessment field is far from settled and retains an unstable quality.

6. Conclusions

The study of the negotiation of SDG4 learning data offers an opportunity to shed light on the intricacies and messy nature of global indicator-making, as well as the impact that such processes may have on the partaking organizations. The paper has first unpacked the data production stage. It has shown that, despite broad agreement on the centrality of learning measurement, the production of Indicator 4.1.1 has been far from straightforward and conflict-free. In line with the findings advanced by the literature on global indicator-making, the difficulties encountered in the process owe much to the collective nature of data-collection processes. Thus, obstacles have not been exclusively technical in nature, but also stem from the difficulty of reconciling the multiple (and sometimes conflicting) interests and priorities of the data suppliers engaged in such efforts. Specifically, the paper finds that the increasing centrality of learning data has been seized by producers of cross-national assessments as an opportunity to consolidate and expand their outreach and portfolio of activity– a process that inevitably creates issues of rivalry and overlap, sometimes exacerbating decades-long conflicts. In consequence, the orchestration labor performed by the UIS has proved a particularly challenging enterprise. The UIS has been tasked with keeping assessment producers on board (offering them reputational or material incentives powerful enough to get them to

[11] Namely, the *Better Education Statistics and global Action to improve learning* (BESTA).

engage) while 'taming' the assessment industry – for instance, preventing certain data producers from imposing their methodological preferences, or from using SDG4 as a 'product placement' scheme.

The study has also shown how the production of SDG4 learning data has had a number of organizational effects (Freistein, 2016) on the UIS. An immediate *external* effect has been the revitalization of the role of the UIS in the production of globally comparable data. The UIS has thus gained substantial visibility and centrality in an area in which it would have struggled to affirm its leadership. This reinvigoration appears to be largely the result of the UIS' ability to posit itself as an honest broker driven by a public-service *ethos* – one able to convene different parties and interests, and to build consensus in a fraught arena riddled with vested interests. Remarkably, the UIS has succeeded in this by bringing to the fore the necessarily imperfect nature of global datasets. While the construction of 'good enough data' typically occurs in the backroom, the UIS has succeeded in bolstering its own credibility by doing exactly the opposite – that is, by exposing the messiness, complexity and political difficulties behind the production of globally comparable learning data. In this sense, the paper corroborates Rocha de Siqueira's (2017) observations that data producers recognize global datasets as inherently imperfect but accept such inaccuracies out of pragmatism. However, the study also contends that the approximate and error-prone nature of global data is not simply accepted as a 'lesser evil' on the part of indicator promulgators – imperfections can become an opportunity for international organizations to reassert their own authority.

At the same time, the UIS' position is by no means a solid one, largely as a consequence of the emergence of parallel measurement projects such as the World Bank's Human Capital Index and Learning Poverty Indicator. This is compounded by the *internal* effects experienced by the UIS as a consequence of its reinvigorated role in the production of learning data. Thus, the limited institutional alignment between the UIS and UNESCO has translated into a lack of support that could ultimately jeopardize both the success of the global reporting effort and the leadership or pilotage capacity of the UIS.

Overall, the paper offers a number of empirically-informed insights into the production of global data – which remains a comparatively under-researched phase, in contrast with the conceptualization, use and impact of global metrics. However, the results might also have implications (or provide useful insight) for international development efforts. Thus, the paper shows that the production of SDG4 learning data has been very much shaped by a tension between the principles of in-built comparability and country ownership – the former being better served by large cross-national assessments, the latter more likely to be fulfilled by a hybrid approach that maximizes data-source flexibility. Such tensions are ultimately indicative of the fact that, in the absence of an honest broker, global reporting frameworks might end up trumping national statistical priorities. This is certainly the case when such frameworks encourage ad-hoc measurement exercises exclusively driven by comparability purposes, but lacking the granularity and frequency necessary to orient local and domestic policy-making. In this sense, the findings of the paper echo an emerging literature concerned with the risks of the disconnect between the supply and demand side of development data (Custer and Sethi, 2017; MacFeely and Barnat, 2017). The findings also suggest that, while the UIS has gone to great lengths to maximize data-source flexibility and ensure that SDG4 monitoring needs do not trump domestic measurement efforts, many remain scarcely informed of the opportunities and costs associated with different forms of learning assessment. The responsibility to help countries navigate the different assessment options cannot fall exclusively on the UIS or UNESCO; efforts are also required from donors and development partners who, through their technical and economic support, contribute directly to shaping domestic priorities in the area of learning measurement.

Conflict of interest

The author declares that there is no conflict of interest.

Acknowledgements

I am grateful to the three anonymous referees and to Oscar Valiente for their insightful comments on earlier versions of this paper, and to Antoni Verger for his advice on the dissertation from which this paper derives. All omissions and errors remain my own.

List of references

The references marked with an asterisk correspond to those documents gathered for the purposes of documentary analysis. Note however this is not an exhaustive list – the complete relation can be found in (Fontdevila, 2021).

References

Addey, C., 2018. The assessment culture of International Organisations: "From philosophical doubt to statistical certainty" through the appearance and growth of Internatioinal Large-Scale Assessments. In: Alarcón, C., Lawn, M. (Eds.), Assessment Cultures: Historical Perspectives. Peter Lang, Berlin, pp. 379–408.

Adler, E., Pouliot, V., 2011. International practices. Int. Theory 3 (1), 1–36.

Altinok, N., Angrist, N., Patrinos, H.A., 2018. Global Data Set on Education Quality (1965–2015) (Policy Research Working Paper No. 8314). The World Bank, Washington, D.C.

Arndt, C., 2008. The politics of governance ratings. Int. Public Manag. J. 11 (3), 275–297.

Benavot, A., Smith, W.C., 2020. Reshaping quality and equity: global learning metrics as a ready-made solution to a manufactured crisis. In: Wulff, A. (Ed.), Grading Goal Four: Tensions, Threats, and Opportunities in the Sustainable Development Goal on Quality Education. Brill Publishing, Leiden, pp. 238–261.

Birdsall, N., Bruns, B., Madan, J., 2016. Learning Data for Better Policy: A Global Agenda (CGD Policy Paper No. 96). Center for Global Development, Washington, D.C.

Cooley, A., 2015. The emerging politics of international rankings and ratings. A framework for analysis. In: Cooley, A., Snyder, J. (Eds.), Ranking the World. Grading States as a Tool of Global Governance. Cambridge University Press, Cambridge, pp. 1–38.

Fontdevila, C., 2021. *Global governance as promise-making. Negotiating and Monitoring Learning Goals in the Time of SDGs*. PhD thesis. Universitat Autònoma de Barcelona. Available at: https://www.tesisenred.net/handle/10803/672579#page1/41.

Fontdevila, C., 2020. Learning assessments in the time of SDGs. New actors and evolving alliances in the construction of a global field. In: Wulff (Ed.), *Grading Goal Four. Tensions, Threats, and Opportunities in the Sustainable Development Goal on Quality Education* Brill | Sense, Leiden, The Netherlands/Boston, MA, pp. 262–279.

Cussó, R., d'Amico, S., 2005. From development comparatism to globalization comparativism: Towards more normative international education statistics. Comp. Educ. 41 (2), 199–216.

Custer, S., Sethi, T., 2017. Avoiding data graveyards: insights from data producers and users. In: Custer, S., Sethi, T. (Eds.), Avoiding Data Graveyards: Insights from Data Producers & Users in Three Countries. AidData at the College of William & Mary, Williamsburg, VA, pp. 1–8.

Dahler-Larsen, P., 2014. Constitutive effects of performance indicators: getting beyond unintended consequences. Public Manag. Rev. 16 (7), 969–986.

Davis, K.E., Kingsbury, B., Merry, S.E., 2012. Introduction: Global governance by indicators. In: Davis, K.E., Fisher, A., Kingsbury, B., Merry, S.E. (Eds.), Governance by Indicators Global Power through Quantification and Rankings. Oxford University Press, Oxford, pp. 3–28.

Wulff, A., 2020. The twists and turns in negotiating a global education goal: a civil society perspective. In: Wulff, A. (Ed.), Grading Goal Four. Tensions, Threats and Opportunities in the Sustainable Development Goal on Quality Education. Brill | Sense, Leiden, The Netherlands/Boston, MA, pp. 28–64.

DfID. 2018, Better Education Statistics and global Action to improve learning (BESTA). Annual Review. DfID. Retrieved from: https://devtracker.fcdo.gov.uk/projects/GB-1–204695/documents.

ECOSOC, 2017. Report of the High-level Group for Partnership, Coordination and Capacity-Building for Statistics for the 2030 Agenda for Sustainable Development (E/CN.3/2017/3). ECOSOC, New York.

Erkkilä, T., Piironen, O., 2018. Rankings and Global Knowledge Governance. Higher Education, Innovation and Competitiveness. Palgrave Macmillan, Cham, Switzerland.

Espeland, W.N., Stevens, M.L., 2008. A sociology of quantification. Eur. J. Sociol. /Arch. Eur. De. Sociol. 49 (3), 401–436.

Freistein, K., 2016. Effects of indicator use: a comparison of poverty measuring instruments at the World Bank. J. Comp. Policy Anal.: Res. Pract. 18 (4), 366–381.

Geertz, C., 1973. The interpretation of cultures: Selected essays. Basic Books,, New York.
Kelley, J.G., & Simmons, B.A. (2014, August-September). The power of performance indicators: rankings, ratings and reactivity in International Relations. Paper presented at the Annual Meeting on the American Political Science Association, Washington D.C. Retrieved from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2451319.
King, K., 2016. The global targeting of education and skill: policy history and comparative perspectives. Comp.: A J. Comp. Int. Educ. 46 (6), 952–975.
MacFeely, S., Barnat, N., 2017. Statistical capacity building for sustainable development: developing the fundamental pillars necessary for modern national statistical systems. J. Int. Assoc. Off. Stat. 33 (4), 895–909.
Merriam, S.B., Tisdell, E.J., 2016. Qualitative Research. A Guide to Design and Implementation, 4th ed. Jossey-Bass, Wiley, San Francisco, CA.
GEMR, 2018. Is global education data heading toward fragmentation? Retrieved from: https://world-education-blog.org/2018/09/19/is-global-education-data-heading-toward-fragmentation/*.
Montoya, S., 2017, July 12. A pragmatic and unified approach to measure learning globally (Blog post – UIS). Retrieved from: http://uis.unesco.org/en/blog/pragmatic-and-unified-approach-measure-learning-globally?=undefined&wbdisable=true.
Montoya, S., 2015. *Monitoring Education for All Global and Thematic Indicators. ETAG Report to the EFA Drafting Group*. Presentation delivered by the UIS Director. *.
Morgan, C., 2011. Constructing the OECD Programme for International Student Assessment. In: Pereyra, M.A., Kotthoff, H.-G., Cowen, R. (Eds.), PISA Under Examination: Changing Knowledge, Changing Tests, and Changing Schools. Sense, Rotterdam, pp. 47–59.
Mundy, K., 2010. Education for All and the global governors. In: D'Avant, D., Finnemore, M., Sell, S.K. (Eds.), Who Governs the Globe? Cambridge University Press, Cambridge, pp. 333–355.
Neumann, I.B., 2002. Returning practice to the linguistic turn: the case of diplomacy. Millenn.: J. Int. Stud. 31 (3), 627–651.
Ritchie, J., Lewis, J., Elam, G., 2003. Designing and selecting samples. In: Ritchie, J., Lewis, J. (Eds.), Qualitative Research Practice. A Guide for Social Science Students and Researchers. SAGE, London, pp. 77–108.
Rocha de Siqueira, I., 2017. Development by trial and error: the authority of good enough numbers. Int. Political Sociol. 11 (2), 166–184.
Sachs-Israel, M., 2017. The SDG4-Education 2030 Agenda and its Framework For Action – The process of its development and first steps in taking it forward. Bild. und Erzieh. 69 (3), 269–290.
Sayed, Y., Ahmed, R., Mogliacci, R., 2018. The 2030 Global Education Agenda and the SDGs: process, policy and prospects. In: Verger, A., Novelli, M., Altinyelken, H.K. (Eds.), Global Education Policy and International Development, 2nd edition.,. Bloomsbury, London/New York, pp. 185–208.
Crouch, L., Bernard, J.-M., 2017, July 6. *Measure learning now to reduce inequality tomorrow* (Blog post – GPE). Retrieved from: https://www.globalpartnership.org/blog/measure-learning-now-reduce-inequality-tomorrow *.
UIS 2022a. TCG Composition. Retrieved from: https://tcg.uis.unesco.org/tcg-composition/*.
UIS, 2019a. Global Alliance to Monitor Learning: Update on progress. Presentation delivered by the UIS Director in the context of the 6th GAML meeting. Retrieved from: http://gaml.uis.unesco.org/wp-content/uploads/sites/2/2019/05/GAML6-Session1-Progress.pdf *.
UIS, 2016b. Global Alliance to Monitor Learning. List of participants. UIS, Montreal, Canada (Retrieved from). http://gaml.uis.unesco.org/wp-content/uploads/sites/2/2018/08/Confirmed-participants_GAML2-meeting_Website_20161101.pdf.
UIS, 2017. The Global Alliance to Monitor Learning (GAML): Governance and organization. UIS, Montreal, Canada. http://gaml.uis.unesco.org/wp-content/uploads/sites/2/2018/10/gaml-governance-organization-2017-en.pdf.
UIS, 2018a. Making the Case for a Learning Assessment. UIS, Montréal.
UIS, 2018b. Global Alliance to Monitor Learning (GAML): 2018 Progress report. UIS,, Montreal, Canada. http://gaml.uis.unesco.org/wp-content/uploads/sites/2/2018/10/GAML-5-Report.pdf.
UIS, 2019b. SDG 4 Data Digest How to Produce and Use the Global and Thematic Education Indicators. UIS, Montreal, Canada.
UIS, 2022b. Rosetta Stone Policy Brief. Establishing a concordance between regional (ERCE/PASEC) and international (TIMSS/PIRLS) assessments (UIS/UNESCO). UIS, Montréal.
UIS, 2022c. COVID-19 in Sub-Saharan Africa: Monitoring Impacts on Learning Outcomes. Main report. UIS, Montréal.
UNESCO, 2018. *Evaluation of the UNESCO Institute for Statistics* (IOS/EVS/PI/170 REV). UNESCO, Paris.
UNESCO, 2021. Global Education Monitoring Report. Non-state actors in education. Who chooses? Who loses? UNESCO, Paris.
UNGA, 2017. Resolution adopted by the General Assembly on 6 July 2017 - 71/313. Work of the Statistical Commission pertaining to the 2030 Agenda for Sustainable Development (A/RES/71/313). United Nations General Assembly. *, New York.
Unterhalter, E., 2019. The many meanings of quality education: politics of targets and indicators in SDG4. Global Policy 10 (S1), 39–51.
Ward, M., 2004. Quantifying the World. UN Ideas and Statistics. Indiana University Press, Bloomington, IN.
World Bank, 2019. Learning Poverty: What Will It Take? The World Bank, Washington, D.C. (*).
GAML, 2019. Global Proficiency Framework: Reading and mathematics. Grades 2 to 6. UIS. Retrieved from: https://gaml.uis.unesco.org/wp-content/uploads/sites/2/2019/05/GAML6-REF-16-GLOBAL-PROFICIENCY-FRAMEWORK.pdf.
Montoya, S., Crouch, L., 2019, April 26 and May, 20. The learning assessment market: pointers for countries (Part 1 and 2). (Blog post – GEMR). Retrieved from: https://gemreportunesco.wordpress.com/2019/04/26/the-learning-assessment-market-pointers-for-countries-part-1/.
Montoya, S., Beeharry, G. & Woolf, E., 2019, January 22. A partnership for a global public good: data to improve learning (Blog post – UIS). Retrieved from: https://sdg.uis.unesco.org/2019/01/22/a-partnership-for-a-global-public-good-data-to-improve-learning/.
Smith, W., Benavot, A., 2021. Quality education and global learning metrics. In: McCowan, T., Unterhalter, E. (Eds.), Education and International Development, 2nd ed. Bloomsbury, London, pp. 189–206.
Treviño, E., Órdenes, M., 2017. Exploring Commonalities and Differences in Regional and International Assessments (UIS Information Paper No. 48). UIS, Montreal, Canada.
UIE, 1962. Educational achievements of thirteen-year-olds in twelve countries. UNESCO Institute for Education, Hamburg (*).
UIS, 2016a. Proposal to develop global learning metrics. How to measure SDG 4.1. UIS, Montreal, Canada (Retrieved from). http://uis.unesco.org/sites/default/files/documents/proposal-to-develop-global-learning-metrics.pdf.
World Education Forum, 2000. The Dakar Framework for Action. UNESCO, Paris (*).