Commentaries on the discussion paper:

Options in achieving global comparability for reporting on SDG 4

By Silvia Montoya and Brenda Tay-Lim
UNESCO Institute for Statistics

Commentaries by

Kadriye Ercikan, University of British Columbia

Tünde Kovács Cerović, Belgrade University and Open Society Foundations

Radhika Gorur, Deakin University

William H. Schmidt, Michigan State University

Response to commentaries by

Luis A. Crouch, RTI International and Silvia Montoya, UNESCO Institute for Statistics

The research reported here was made possible (in part) by a grant from the Spencer Foundation Grant #201800045. The views expressed are those of the authors and do not necessarily reflect the views of the Spencer Foundation. The views also do not reflect the views of the Center for Advanced Studies in Global Education.

Will SDG4 achieve environmental sustainability?

http://dx.doi.org/10.14507/casge4.2018

CASGE working papers are circulated for discussion and comment purposes. They have not been peer-reviewed.

Commentary 1

Options in achieving global comparability for reporting on SDG 4: Some Thoughts and Suggestions

Kadriye Ercikan
Educational Testing Service/University of British Columbia

The objective of UNESCO's Institute for Statistics (UIS) is "to find ways to link different assessment results and to report them in a globally comparable way," and UIS is interested in building "a pragmatic system that could produce comparable results that allow for trend reporting."

Silvia Montoya and Brenda Tay-Lim lay out the challenges in achieving this goal and identify different statistical and non-statistical approaches to linking international assessments with national, regional, and local assessments. Linking these assessments has many benefits as identified by the authors of the paper, including cost savings, coherence in interpretability of scores to inform policy and planning decisions, and opportunity for closer alignment of national/localized assessments with local country curricula. However, the primary impetus for the linking explorations seems to be to meet the mandate from Sustainable Development Goals for Education (SD4) to report on the proportion of children and young people achieving a minimum proficiency level in reading and math in grades 2/3, at the end of primary, and at the end of lower secondary.

Let's take a look at what the targeted objectives imply. In order to meet the SD4 mandate, the assessments need to be linked to (1) be able to estimate/predict performance on local assessments based on international/regional assessments; and (2) be able to report performance results comparable across years (for trend analysis). These linking procedures should result in an estimation of proficiency in reading and math at the targeted grades.

In this commentary, I will identify factors that need to be taken into account when linking scores across assessments, comment on different linking approaches considered, and highlight issues that require further consideration.

Linking of performance results from different assessments are expected to be impacted by the similarity of *constructs*, *populations*, and *measurement characteristics* of the linked assessments (Kolen & Brennan, 2004).

- Constructs
    - Similarity of constructs refers to the degree to which the same construct is targeted by the assessment as would be identified by similarity of content frameworks.
    - For example, differential emphasis on different aspects of math and reading may impact the similarity of constructs assessed.

- Populations
  - Similarity of targeted populations (ages, grades etc.)
  - Sampling design and implementation (sample versus census, stratification, classroom level versus school level)
  - Inclusion/Exclusion criteria of special populations (vary across different international assessments as well as across countries)
- Measurement characteristics
  - Accessibility resources and tools (may vary for international assessments and national assessments)
  - Motivation (lower motivation on international/regional assessments)
  - Lengths of tests and types of items used
  - Measurement properties

Differences in assessments with respect to these factors are expected to impact the accuracy of estimates obtained from linking.

Four Linking Procedures

Given these factors, let's take a look at the four linking approaches considered.

Policy linking/ Social moderation – is in general considered as the weakest form of linking. It can be useful for initial understanding of similarities and differences across different scales (national and international). However, this approach to linking is quite limited in providing reasonable estimates of distribution of performance on one assessment based on another. Statistical moderation that focuses on aligning distributions of scores from different assessments can be added to the list of linking approaches that will allow empirical support/revisions to a linking established by a policy linking.

Statistical approach – Psychometric item based linking (2a). Important to note that even though the authors refer to this approach as item based equating, the method cannot be considered as equating that is expected to result in exchangeable scores. Equating is reserved "for a relationship between scores of different forms that are constructed according to the same content and statistical specifications" (Kolen & Brennan, 2004, p. 430). What the authors are describing is considered calibration. This approach involves inclusion of items with existing parameters in assessments for which the linking is needed. How many items, how the item parameters are obtained, what kinds of items they were combined with, and what kind of student sample was used to estimate the parameters are only some of the factors that will affect the validity of the item parameters. There was not sufficient information about where these anchor items would come from and where the associated parameters would come from. Added to this, is the comparability of construct, populations (age/grade), and measurement characteristics of the assessments that are targeted to be linked.

Statistical approach – Psychometric based linking (2b). This approach involves linking of assessment results from five regional assessments administered in mostly sixth grade, and one in fifth grade, and TIMSS and PIRLS 4th grade assessments. This is another calibration approach with intact TIMSS and PIRLS booklets. It is indeed stronger than policy linking and anchor item linking (2a) approaches described above. This linking will allow predicting performance on TIMSS and PIRLS based on performance on national assessments, as well as the other way round. A major concern in this case is appropriateness of grade 4 assessments to be administered at grades 5 and 6 in the regional assessments and representativeness of the anchor booklets of the total PIRLS and TIMSS scales.

Statistical approach – Alignment through Statistical Modeling. This approaches focuses on using data from several countries who took part in both international and regional assessments to establish linking between these sets of assessments that can be applied to other countries. Such an approach can be very valuable in evaluating different linking approaches.

Issues that Require Further Consideration

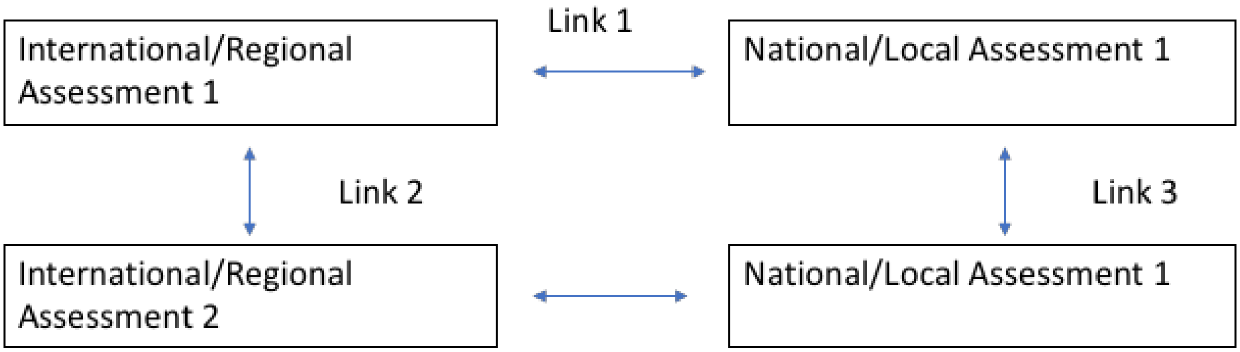
Linking needs to be aligned with the purposes it serves

Even though the emphasis has been on the SDG4 mandate and the impetus for the proposed linking plans, at least two of the approaches described above focus on predicting performance on international assessment based on regional assessments. It's not clear how this type of linking will help meet the SDG4 mandate. Are standards established by international assessment considered to be higher and more acceptable?

Trend over time requires consideration of alignment/linking of assessments across years

Performance trends over time are difficult to establish within countries due to changes in curricula, student populations, and test format/mode. Any linking established in one assessment cycle may not hold in the next one as multiple linking is implicated in such performance comparisons. Therefore, there needs to be consideration of performance results across years.

Figure 1. Multiple linking needed for comparing performance across assessments and across years


International/Regional Assessment 1
Link 1
National/Local Assessment 1
Link 2
Link 3
International/Regional Assessment 2
National/Local Assessment 1

Commentary 2
Zooming out: Possible threats to the process?

Tünde Kovács Cerović
Belgrade University

It is not easy to comment the Montoya and Tay-Lim paper. The condensed 10 pages take the reader on four exciting hypothetical roads of weaving together disparate assessment tools that measure disparate constructs, in different countries and under unknown circumstances, and asks us which of the four is the best or how could they be further combined. The paper, although highly technical, opens a new future global field of knowledge, data, and evidence-based policy making and addressing (hopefully also reducing) between-country and inside-country differences. It zooms in on the details of the steps (or possible steps) to link information from different kinds and types of National or International Assessments to outcomes operationalized through the SDG 4.1.1 indicators for the sake of precise and unbiased global comparison of learning.

I will use this opportunity to zoom out and highlight some contextual links that, if missed, can become reasons for concern, doubt, and even disengagement or misuse. I see these links important especially
- from the perspective of vulnerable groups, b) based on experiences in conducting pro-poor research, or c) with policymaking in times of educational reforms.

Concerns about the entry points

The described procedures start with the optimistic statement that 81% of states participates in an international learning assessment, and they all have administrative data and rely on some kind of EMIS.

Let me first problematize this end of the process, and raise a concern about the local level data entry, administrative data, sampling procedures and coverage, and formulate a qualified doubt about the quality of data analyses that build on data biased by discrimination, exclusion, integrity issues, or politicization. Too often have I seen in countries where Roma live (the most marginalized and discriminated against European minority of about 12 million peoples) how evidence-based policy making already staggers, stalls, and collapses at the very level of inclusive data collection, even in the midst of Europe, even in countries with a tradition of keeping records and evidence in social, health, and education sectors. The incredibly creative local solutions for establishing, re-establishing, and maintaining exclusion seem often to prevail over national or international policies of integration.[1] One can find how the municipality conveniently does not account for Roma settlements, just a couple of kilometers away from the city centers with children that were never in school.[2] Discrimination and marginalization means exactly that: being invisible in the best case. One can meet Roma school aged children on the streets of the *mahala* in Serbia, speaking perfect French or German since they grew up abroad, but not Serbian; hence they cannot go to school until they learn the language of instruction. Of course, without instruction they will not learn it. Or, one can find Roma children enrolled in special schools, even when inclusive education is a nation-wide policy[3] (the special school director will take care not to show you the classroom with Roma when you come to visit, and most probably will not keep that information in the school's administrative records[4]). Roma children can be enrolled in several neighboring schools as "ghost students," serving the need to maintain the municipal funding[5] and for the sake of the school, for teacher employment and not student learning. The chances of finding similar procedural disconnects in the mere administrative data about enrollment (or nonenrollment, that will be essential to report on, critically important for SDG 4.5.1. parity indicators) in other middle- or low-income countries are of course high, and they will inevitably influence the quality of data in further processing. I will never forget the school in the outskirts of a Central Asian capital where authentic and well-designed cooperative learning was taking place, although the school did not get the manuals and textbooks that were allegedly distributed, did not get the trainings for the new curriculum, has no water or sewage and only restricted electricity, and nobody has responded to their requests, claims, or suggestions for improvement. They will most probably not be included in national sampling either. Or, the complicated scenarios parents have to navigate to enroll their child in a prospective school, or for that matter in any school or preschool in some countries within deep transition. [6]

My intention is not to provide a full picture about the flip side of education - the flip side is much richer and diverse than I could fit into a couple of lines. My intention is to pose, building on these examples and selected without any serious systematization, two interrelated critical questions that would need to be incorporated in any further discussions about SDG 4 indicators.

- While highly appreciating the need to further the global agenda and refocus from enrollment data to learning assessment data in the new SDG framework, the question is: have we all learned the lessons from the MDG process, and are we sure administrative data refers to all categories of students, free of bias, based on discrimination, political affiliation, or corrosive

[1] Brown, P., Dwyer, P., Martin, P., Scullion, L., & Turley, H. (2015). *Rights, responsibilities and redress?. Research on policy and practice for Roma inclusion in ten member states.* Brussels: European Commission.

[2] Personal recollection from monitoring visit to Valjevo, Serbia.

[3] Bojadjijeva, A. (2015). *Roma Inclusion Index*. Budapest, Hungary: Decade of Roma Inclusion Secretariat Foundation.

[4] IPSOS. (2016). *Education in schools and classes for the education of children with disabilities in Serbia*. Belgrade: UNICEF & Ministry of Education, Science and Technological Development.

[5] Roma Education Fund. (2007). *Advancing Education of Roma in Bulgaria* (http://www.romaeducationfund.org/publications/country-assessments)

[6] OECD. (2017). *OECD Reviews of Integrity in Education: Ukraine, 2017*. OECD Publishing, Paris.

practices to the extent needed to accurately report on, as requested in the Montoya & Tay-Lim paper?

- Having in mind the procedural and data flaws described above, and possible contexts in which no measures are installed to counteract discrimination, political pressure, or corruption, are we sure that the learning assessments will be valid and reliable? Do we seriously hope that local politicizing will not exclude from the sampling and assessment exactly those that local and school politics ignore, hide, eliminate?

My answers based on my research and personal experiences from the field in diverse countries are unfortunately negative. Countries, even in Europe and even after six years of *Decade of Roma* integration, still struggled with their Roma-related statistics, with census usually showing a 3-4 times lower Roma population than the international and nongovernmental estimates.[7] Roma pedagogical assistants, in their narratives about education reform for social integration, have many doubts about how teachers address the Roma children and face discrimination themselves directly. The same processes that were identified in the 1960s by Rosenthal and Jacobson[8] are still working through teachers' low expectations towards the marginalized groups (such as Roma , or recently migrated students throughout Europe[9],[10]). Strong agency is needed to put the spotlight on concrete acts of discrimination and a strong voice needs to repeat to the teachers: "Don't put them in the back rows," "Do not neglect them at maths," "Somebody's future depends on your work," "Sometimes, give more than you can," "A little more effort will mean a lot," "Be the first role model for a Roma child throughout his/her life."[11] Without such voices the bias will be mediated by the data, but it will not disappear, and consequently "the results" will be inherently questionable.

Concern about the end points and use of global data

At the other end of the process I find further missing links, which lead to the question of how will the results be used and by whom. Here my questions and concerns derive from the perspective of a policy maker at the national level. There seem to be several critical steps that are not fully addressed in the paper, but that can and most probably will affect the propensity to use the results for policymaking, and

[7] European Commission. (2011). Communication from the Commission to the European parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: An EU Framework for National Roma Integration Strategies up to 2020. Brussels: European Commission. http://eurlex.europa.eu/legalcontent/EN/TXT/PDF/?uri=CELEX:52011DC0173 &from=en

[8] Rosenthal, R., & Jacobson, L. F. (1968). Teacher expectations for the disadvantaged. *Scientific American*, *218*(4), 19–23.

[9] Kovač Cerović, T., Grbić, S., & Vesić, D. (2018). How do schools integrate migrant students: Case studies from Serbia. In: F. Dovigo (Ed): *Challenges and opportunities in education for refugees in Europe*. Leiden; Boston: Brill Sense

[10] Kovač Cerović, T., Grbić, S., & Vesić, D. (2018). *Peer relationships between migrant and domicile students*. Paper presented at ECER, Bolzano, September 4-7, 2018.

[11] Daiute, C., & Kovács-Cerović, T. (2107). *Minority teachers - Roma in Serbia – narrate education reform*. Belgrade: Institute for Psychology & Kragujevac: APAS.

consequently also the care and diligence invested in data collection, adequate storage, cleaning, and all other procedures of creating evidence from raw data. We know from analysis that research evidence does not translate into policy directly, unmediated,[12] and that there are many disparities and gaps in the process of translation between evidence producers and evidence users[13]—indeed, that is why many countries use some kind of data broker services tasking them to help in the translation.[14] These services are less needed in cases when policy makers know what they are searching for, which kind of results will inform and feed into which of the unfolding policies, and in what ways the research evidence or information on global/national/local indicators will enter public debate and reinforce the new policies in public. After all, we should not forget that data collection, even on administrative issues, and increasingly so with learning assessments, has to be fully meaningful and purposeful at all the spots where data are generated, processed, stored, and analyzed—that is, from the perspective of every teacher, each school, each municipality, region, and so on. All these structures will have to counteract entropy, a wish to make their task easier, or to ensure that results will come out fine for their class, school, municipality, or nation. There are billions of facets of data manipulation, we increasingly know. And, of course, the less positive the scenario, the more work for broker services, for UIS, for all working on the global level. Nevertheless, that might not be enough—global comparisons can be used, not used, misused and abused, and I do not see any other mechanism to affect this but genuine participation in decision-making and the engagement of civil society.

And finally, it is also not surprising that even when national governments have a clear agenda that will need the 4.1.1 indicators, the meticulously organized comparative analyses will not be the major information source for them. They will want to compare, with their own progress, effects of interventions, and trends. They will want to know only broad comparative information, more for the sake of using it to mobilize the public or create interest for the reforms they are pursuing. Will they get the needed support in this respect? Or will we all face innovation fatigue and be content with the global comparative analysis?

Let me finish with a quick story about PISA in Serbia, illustrating use, non-use, misuse, and abuse in the realm of politics in the same country. Serbia joined in 2003, with a wish to get a clear understanding and diagnosis of where we are after the devastation (both financial and mental) of the education system during the 1990s, and a policy leaver for forthcoming changes. The government changed to a conservative line in 2004; the minister that received the first PISA results was shocked and treated the results as a way for the humiliation of the country by "the West," ignored them, and strengthened the rhetoric about the fantastic results of the Serbian students on the mathematic Olympiad. This Minister was quickly changed, and the next one allowed PISA due to pressure by international agencies. The

[12] OECD. (2007). Evidence in Education: Linking research and policy

[13] A world that counts: Mobilizing he data revolution for sustainable development (2014). www.undatarevolution.or

[14] European Commission/EACEA/Eurydice, 2017. *Support Mechanisms for Evidence-based PolicyMaking in Education. Eurydice Repor*t. Luxembourg: Publications Office of the European Union.

results were again bad but they were not even discussed. A new minister in 2009 promoted PISA as important, provided explanations about what PISA measures and how, and got much better results, just based on creating a positive atmosphere and basic understanding. But, the importance of comparison and ranking was already the dominant interpretative framework for PISA, and he was satisfied, and did not use the more detailed data for further policymaking. In 2012 the results were the same, in 2015, Serbia did not have the ambitions any more to participate in the next round.

I sincerely wish not to generate similar stories with the SG 4.1.1. Could we stop a bit at least at this Symposium and reflect on how to help governments to generate or appropriate authentic interest in the process, and thus ensure filling the missing link at the onset, at data accuracy level, and at the end the level of interpretation and use of results for policymaking by the government?

Commentary 3
Refocusing the project: What is the problem we are trying to solve?

Radhika Gorur,
Deakin University

As the custodian agency of SDG 4.1.1, UIS has embarked on a serious search for universally comparable metrics that are adequately rigorous and at the same time respectful of local cultures and contexts. Keeping in mind the financial, political, and technical challenges of this exercise, they take a pragmatic approach, accepting a trade-off between rigor on the one hand and feasibility, cost, and acceptability on the other. Moreover, rather than put all their eggs in one basket, they outline a plurality of approaches, hedging their bets, since all the outlined approaches are somewhat experimental, and not guaranteed to succeed. Already, it may be noted, Strategy 2a, the UIS-RS, has met with a roadblock, with regional assessments not sharing their items to help with calibration (p. 9 of the Position Paper). And contingency is also in place in the case of countries refusing to share their data (p. 1 of the Position Paper)

In responding to this paper, I draw upon my “sociology of numbers” and education policy research across several projects, which includes research on the establishment of large-scale comparisons; histories of the development of indicators; studies of national and international education policy assemblages; and research on the accountability practices that have attended the declaration of the SDGs. I will begin with a few comments on the methodologies for achieving global comparability outlined in the Position Paper by Sylvia Montoya and Brenda Tay-Lim, and then raise a few meta-questions to support my argument that, instead of finding solutions to the problem of achieving global comparability, the problem itself should be reconsidered, and the focus shifted from *improving global comparability* of learning to just *improving learning*.

Issues with the suggested methodologies

The Position Paper is already very thoughtful and raises several issues—technical, political and practical—in the project of developing globally comparable metrics. I submit some additional issues for consideration:

- All the suggested approaches involve the work of *defining standards* and *agreeing on proficiency levels and progression levels* etc. based on expert consensus and review processes that are seen as promoting rigor, fairness, and neutrality. These methodologies and processes are presented as a-political and technical. But defining and categorizing are ontological exercises that determine identities and bring different groups and individuals into focus as requiring particular interventions (Desrosières, 1998). They have implications for funding, for the very processes of teaching and learning, and for determining who and what comes to be valued in society, with serious consequences for individuals and societies (Gorur, 2016; Porter, 1995). These are deeply political exercises and should be judged not only for their technical rigor, but for the impacts they would unleash (Gorur, 2017).

- In assessing the relevance of PISA for middle- and low-income nations, it was determined that PISA was underpinned by a framework that was suited to OECD nations but provided little of relevance to middle- and low-income nations (Adams & Cresswell, 2016; Lockheed, Porkic-Bruer, & Shadrova, 2015). It is not clear how this issue is overcome in the case of the universal comparisons that developing universals scales for making comparisons—which is done in all the options suggested—discounts and erases whatever logics underpinned the national assessments. This is a deeply political exercise, and moreover, it may produce comparisons that have no meaning or relevance for the nations involved. Lockheed et al. (2015) contend that PISA produced little of value for middle- and low-income nations that spent enormous time and energy participating in that survey.

- All of the strategies outlined involve processes that reinforce existing assumptions that reading and mathematics are the most valued proficiencies or are good proxies for them. The defining of “proficiency levels” and “progressions” assumes that these domains of knowledge involve a linear progression, with one skill preceding the other; that this progression is similar in all cultures and languages; and moreover that this *ought to be* the progression. For example, Step 1 in Strategy 1 talks of building on “learning cognitive theory”—as if that theory was a singular, universally accepted and non-controversial theory. Other theories of learning exist—even in Western modes of thinking, such as constructivist theories, socio-cultural theories, etc. So the choices made in such exercises are neither self-evident nor apolitical. Even if one assumes that learning occurs in the same sequence in a particular domain of knowledge, would this same sequence hold good for all languages, all cultures, all mathematics? Research does not support this assumption (Verran, 2001).

- Strategy 2C, which attempts to harmonize data based on country participation across various assessments, suffers the same issues as research synthesis exercises across data from a range of methodologies, underpinned by a variety of assessments. The World Bank is using this as the basis for its Human Capital Index—and there are serious issues with these calculations, including calculations of learning in terms of months’ or years’ worth of learning (Lui, 2018).

- Legitimacy for the proficiency scales is sought through consensus—either cross-national consensus with representatives of as many nations as possible, or technical consensus, based on a select group of experts. Both of these processes are problematic and can reinforce existing inequities of expertise and monopolies over knowledge. There is a small group of international experts in the area of large-scale assessments. Most of the same experts have worked on developing most of the major international and regional assessments, and likely have advised on many of the national assessments as well. For example, ACER experts have worked on TIMSS and PIRLS, PISA, and SEA-PLM, and are working on the UIS-RS. Many experts serve on the advisory boards of each other’s organizations. Many of them have studied in the same universities. As a result, their views are already well articulated—and even the debates and differences well-worn. This leads to what Sally Merry (2017, p. 6) refers to as “expertise inertia,” where “insiders with skills and experience have a greater say in developing measurement systems than those without—a pattern that excludes the inexperienced and the powerless.” Even where a conscious effort is made to be more widely inclusive through inviting national representatives to consensus-building events, those without the cultural and epistemic capital of the established experts will find it difficult to challenge views or to contribute to the discussion. In fact, any process that seeks to universalize inevitably commits violence with the marginalized, the indigenous, the minorities, and the different.

- One of the justifications for introducing assessment exercises in low-income nations is that it would help “build capacity” in the receiving nations—but “building capacity” is a neo-colonial project that is almost invariably visualized as a one-way process (indeed, it begins with a “capacity needs analysis”—a deficit model). Established experts are seen as “building the capacity” of the “inexpert” rather than a more democratic inter-mingling of situated knowledges and exogenous expertise. Resultant misreading of local contexts, situations, and priorities by international agencies with a particular viewpoint and agenda can lead to costly mistakes (Rappleye & Un, 2018).
- Even if “local capacity” were “enhanced” with regard to statistical knowledge and knowledge of how to conduct surveys, there is the issue of “data inertia” (Merry, 2016), which refers to how, often, given the cost of gathering data, existing data must serve as proxy for what is desired to be measured. Thismeans “existing data determine what an indicator can measure” (Merry, 2016, p. 7). Moreover, training a small group of people (in some cases this is literally one or two experts) to conduct surveys in a particular way in a culture where nothing else in that society works in that way is not guaranteed to produce lasting benefits. Furthermore, there is no plan for training the next generation of experts when the first crop migrate overseas, join commercial organizations, or retire.
- Together, these two forms of inertia—“expertise inertia” and “data inertia”—“inhibit new approaches to measurement and tend to exclude inexperienced and resource-poor actors from having too much influence on what is measured” (Merry, 2016, p. 7). Indeed, here I would go beyond what Merry contends, by arguing that “inexperienced” and “resource-poor” are appellations that are relative to statistical expertise—often these actors have invaluable indigenous knowledges and situated expertise (Haraway, 1988) that underpin their own informal, local practices of learning, in which the role of assessment may not be the same as visualized in Western apparatuses (Scott, 1998; Silova, Komatsu, & Rappleye, 2018).
- The Position Paper anticipates the possibility that some countries may not want to part with their data, and a technical work-around is suggested to overcome this possibility. But such anticipation should give pause to this hegemonic exercise and raise not technical but moral questions.

Why do we need these data?

UIS's justification for developing these measures is that it has the global mandate to do so and is the custodian of data relating to 4.1.1. Herein lies the problem. 4.1.1 is not about promoting learning, but about *reporting on progress* in learning. In other words, the mandate is not to improve education but to monitor the progress towards this goal. Global comparisons are not designed to directly provide any input relevant to teachers and schools. The primary purpose of these assessments is to assist with accountability and regulation, which, in turn is expected to improve learning.

So if the main aim is to facilitate accountability, it is relevant to ask whether these metrics can serve the purposes of monitoring and accountability. This is not quite clear. What will be done with these data? How will they be used? At the national level, my research in India and Cambodia has shown that various datasets are called into play in policy debates. In countries like India, education is highly decentralized—and at the local level, most policy actors are not even aware they have access to the database and do not know how to download data from the EMIS (even when there are computers, Internet access, etc., which can by no means be taken for granted), let alone use it for governance.

At the international level, GPE has introduced some level of results-based funding, as have other organizations. If funding depends on showing results (and it may take years to produce statistically visible improvement), how might it affect populations ruled by corrupt or inefficient governments? If funding is withdrawn, would that not punish populations already oppressed by a corrupt government? What about nations where the average performance is great, but minorities score poorly? Given that UIS and UNESCO have no authority over sovereign nations, how exactly will this accountability work?

Moreover, as UIS has pointed out, over 81% of nations have already participated in some form of international assessment in the last five years. Why is this not adequate for the purposes of monitoring progress? Even if they have not participated in any international assessments, can nations not show "progress" from one cycle of a national assessment to the next, without requiring these assessments to be globally comparable? In other words, if India conducts the ASER survey or NAS regularly, can progress not be monitored on the basis of their performance on these assessments from one year to the next, without these being linked to some other (international) scale? It would appear that most nations have adequate data for the purposes of their own policymaking and governance, and that the universal comparisons are required mainly to satisfy the UIS requirement as the custodian of SDG 4.1.1.

What data are needed for improvement?

In an evocative and passionate analogy, Montoya and Crouch[1] have argued:

> In many countries, education ministers are like air traffic controllers, who see a storm on the horizon but find that 80% of their navigation instruments are either malfunctioning or non-existent. They simply don't have the data to steer their way out of a global learning crisis that affects more than one-half of all children of primary and lower secondary school age, according to estimates by the UNESCO Institute for Statistics (UIS). (Crouch & Montoya, 2018)

[1] https://www.norrag.org/sdg-4-data-investing-millions-today-will-save-billions-future-luis-crouch-silvia-montoya/

I would challenge this view. In my interviews with policy makers in India and Cambodia, I found that people in these positions were not ignorant at all, but very well cognizant of the complex issues that confronted their systems – the poverty in which many students lived, the lack of infrastructure, the poor quality of teacher education programs, the corruption and the political mafia that may be involved in diverting funding and hampering change, etc. At a school level, principals and teachers in small towns have spoken about how the mushrooming of low-cost private schooling has resulted in a migration of students, and how the poor quality of teaching in those schools has meant that children return to the public schools just before high-stakes examinations are held, because the private schools have dismissed poorly performing students from their school, to prevent their results from looking bad. I have met award-winning teachers who persuaded students to come to school on time by sourcing funds for providing breakfast. The problems identified and the solutions that are found are extremely complex and often these arise from logics that are very different to those that underpin the idea of monitoring and accountability by numbers. Many of the issues that arise might not even be visible, let alone be solved, by the kind of global comparisons that UIS is trying to develop.

The focus on globally comparable metrics has also diverted very large sums of money and attention towards this project. Again, Crouch and Montoya (2018) have argued that the cost of not having these metrics would be even greater, as education ministers would simply not know what the problem is and which problem to prioritize and address. I would argue that the reasons many of the issues have continued to dog education systems in low-income nations is not lack of global comparisons. Indeed, global comparisons have produced little of value in middle- and low-income nations (Lockheed et al., 2015).

To examine if having global comparisons is helpful, we need only look at whether education has improved in all the nations that have engaged in global comparisons. Even in nations that regularly participate in international comparisons and have rigorous national assessments, and where there is a great deal of policy focus on education, as in Australia, the results of international comparative data, even when well-analyzed, has not resulted in well-targeted reform or improvement in results. On the contrary, average national scores that are the basis of international comparisons obscure more than they reveal, and could point to policy changes that are ineffective at best and harmful at worst (Gorur & Wu, 2015). Most Anglophone nations have remained stagnant or have been sliding on the PISA rankings, despite having all the data their hearts could desire.

Refocusing the Project

What if we focused on *improving* student learning instead of *measuring* student learning? What if, instead of asking, “How can UIS create globally comparative assessments?” we were to ask, “What data does a school need to improve student learning?” or better still, “How can learning be improved in schools in this province?” And what if we involved teachers and principals in clusters of schools, and provided them with the resources and the support to improve learning? What if a small expert group were attached to these clusters to report on progress, to satisfy accountability requirements? I have seen considerable progress using locally developed programs, such as the Nali Kali program in Karnataka, India, which is quietly beginning to transform learning in public primary schools. Re-focusing the project from one of monitoring learning to one of learning itself might provide a way of overcoming our “expert inertia” and “data inertia” and doing something truly useful and innovative.

References

Adams, R., & Cresswell, J. (2016). *PISA for Development Technical Strand A: Enhancement of PISA Cognitive Instruments.* [Working Paper]. Paris: OECD.

Crouch, L., & Montoya, S. (2018, 24 February). SDG 4 Data: Investing in MIllions Today WIll Save Billions in the Future. [Blog]. https://sdg.uis.unesco.org/2018/01/26/sdg-4-data-investing-millions-today-will-save-billions-in-the-future/

Desrosières, A. (1998). *The politics of large numbers : A history of statistical reasoning*. Cambridge, Mass.: Harvard University Press.

Gorur, R. (2016). Seeing like PISA: A cautionary tale about the performativity of international assessments. *European Educational Research Journal, 15*(5), 598–616. http://dx.doi.org/10.1177/1474904116658299

Gorur, R. (2017). Towards Productive Critique of large-scale comparisons in education. *Critical Studies in Education*(Online). http://dx.doi.org/10.1080/17508487.2017.1327876

Gorur, R., & Wu, M. (2015). Leaning too far? PISA, policy and Australia's 'top five' ambitions. *Discourse-Studies in the Cultural Politics of Education, 36*(5), 647-664.

Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of artial perspective. *Feminist Studies, 14*(3), 575-599.

Lockheed, M., Porkic-Bruer, T., & Shadrova, A. (2015). E*xperience of middle-Iicome countries participating in PISA 2000-2015*. Washington, DC; Paris: World Bank, OECD Publishing.

Lui, J. (2018). Mind the learning gap: A methodological look into World Bank's New Human Capital Index. Retrieved from https://http://www.norrag.org/mind-the-learning-gap-a-methodological-look-into-world-banks-new-human-capital-index-by-ji-liu/

Merry, S. E. (2017). *The seducations of quantification: Measuring human rIghts, gender violence, and sex trafficking*. Chicago: Chicago University Press.

Porter, T. M. (1995). *Trust in numbers : The pursuit of objectivity in science and public life*. Princeton, NJ: Princeton University Press.

Rappleye, J., & Un, L. (2018). What drives failed poicy at the World Bank? An inside account of new aid modalities to higher education: context, blame and infallibility. *Comparative Education, 54*(1), 1-25.

Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. New Haven: Yale University Press.

Silova, I., Komatsu, H., & Rappleye, J. (2018). Facing the climate catastrophe: Education as solution or cause? Retrieved from https://http://www.norrag.org/facing-the-climate-change-catastrophe-education-as-solution-or-cause-by-iveta-silova-hikaru-komatsu-and-jeremy-rappleye/

Verran, H. (2001). *Science and the African logic*. Chicago: Chicago University Press.

Commentary 4

Options in achieving global comparability for reporting on SDG 4: Some Thoughts and Suggestions

William Schmidt
Michigan State University

Some sixty years ago a group of university professors from Sweden, England and the United States decided that to meaningfully compare educational systems across countries necessitated the collection of quantitative data including those derived from a common assessment. They decided first of all, to determine if such a common quantitative assessment of achievement across countries was feasible. Appropriately, the study was dubbed the "Pilot Study." They succeeded but encountered many problems in their attempt to represent country differences in achievement. Those identified difficulties are still salient in the world of international assessments to this day. They are also salient to the proposals made in the paper written by Montoya and Tay-Lim. The authors propose options for achieving global comparability for reporting on *The Sustainable Development Goals for Education (SDG 4)*. The goal in question is for all member states to report the "Proportion of children…achieving at least a minimum proficiency level" In this paper, I discuss the proposals but only with respect to mathematics.

Montoya and Tay-Lim's paper proposes four approaches; one conceptually driven while the other three are psychometric approaches. As the report suggests, each has its own set of difficulties and its own strengths. Later in the paper I will give my opinion as to which of the suggested approaches would work best for achieving the UIS goal.

I believe the inherent difficulty in achieving the UIS goal of providing common proficiency estimates for every country derived from different assessments – some international such as PISA and TIMSS, and others regional or even national in scope, is in defining the mathematics content to be used in the definition of proficiency. The founding fathers that created international assessments identified this as their number one concern. They phrased the problem slightly differently as they worried about the meaning of the differences among countries on the assessment because of the fact that the curricula of the participating countries were not the same at the grade level at which the assessment was administered.

The basic idea underlying my concern with the proposals made in the Montoya & Tay-Lim paper is that "a math test is not a math test." The content distribution associated with the test reflects a set of standards defined by the purpose of that assessment. Nations create their assessments to be given at specific grades to reflect the curriculum expectations in that country. Even at the international level, the mathematics employed in TIMSS is not the same as is employed in PISA. To ignore this type of fundamental difference around content is dangerous at best and fatal at worst. To use various psychometric scaling techniques to cover for this difficulty does not solve this most basic aspect of assessment. Issues around the content definition can be summarized by four points:

- Answering the question of “proficiency in what?” Mathematics includes multiple areas including number, algebra, geometry, and others depending upon the grade level. Typically countries might use the total score to arrive at an estimate of proficiency. This approach clouds the meaning of proficiency other than in reference to some general propensity in mathematics.
- At a given grade level, such as those suggested by the UIC, the content coverage across countries varies enormously. The international studies such as TIMSS have made that very clear. As a result, defining the content coverage in a common way so that the assessments have the same meaning would be enormously difficult if not politically impossible.
- This cross-country variation in content coverage at various grade levels raises the question of at what level of specificity the content should be defined. This is often called the “grain-size” issue. Agreeing for example that proficiency should be defined in terms of algebra leaves possible tremendous variation across countries as to what parts of algebra they include in their own assessments; this is an issue with an international and regional assessments as well.
- Another issue that needs to be carefully thought through is whether the proficiency is defined cumulatively over the grades leading up to the point at which the assessment is given.

All four of these issues would need to be carefully addressed under any of the proposals put forth in this paper. Clearly the policy linking approach would need to address these types of issues.

The fundamental core problem as captured by the four points can be simply stated as: the definition of the content determines the meaning and the variability of proficiency. This can be illustrated using TIMSS data. In 1995 the Third International Mathematics and Science Study (TIMSS) conducted the first and only largescale international curriculum analysis based on the document analyses of the country standards and textbooks. The study was accomplished in over 40 countries and produced multiple reports (Schmidt, McKnight, Houang, Wang, Wiley, Cogan, & Wolfe, 2001; Schmidt, McKnight, Valverde, Houang, & Wiley, 1997). The results of that study show the difficulty in accomplishing the goal of UIS, especially given the large variation in content coverage for the countries at different grade levels across countries. This was especially true at the end of lower secondary (8th grade for most countries).

The development of the content-coding framework from which these data developed took over a year to even gain agreement as to which topics should be included in the framework. That process which was so fundamental to the development of a common language enabling the data to communicate to all countries in a similar manner was ever so difficult. The paper by Montoya and Tay-Lim, indicates they have looked at over 100 mathematics national test frameworks but I cannot tell to what level of coding specificity that was done. This is because certain topics at a higher level of the mathematical framework can have multiple subtopics underneath it so examining the comparability across countries would depend on some extent on the level of specificity of the coding framework. In other words one topic may appear to be the same across two countries at one level of the framework but when examining it at a lower level in the framework it is different and hence saying they are comparable due to the common word at the higher level of the framework would produce the earlier conclusion.

Rarely was the same topic covered at exactly the same grade for all countries. This would of necessity be the first step in any of the four procedures suggested by Montoya and Tay-Lim. Perhaps this has been done, but the question remains at what level of content specificity, which I cannot tell from this paper. I cannot over emphasis the importance of the framework being specific enough so as to be able to say what mathematics content is truly being covered at the designated grade level.

Having the common definition and applying it to the official set of standards within each of the participating TIMSS 1995 countries enabled us to produce a document of the sort that I believe would be necessary to reach a definition of global competence. Figure 1 shows those results for each of grades 1-8, listing the set of common topics taught at that grade in the top achieving countries. The rule used for the inclusion of a topic was that two-thirds of the countries included the topic at that grade level. The agreement when including the rest of the countries in the table would produce the same topics being listed at virtually each grade level. An analysis of the textbooks for each of those countries confirmed the large variation at which mathematics topics were covered at which grade levels.

In defining proficiency in the TIMSS study, it was found that proficiency would vary across countries depending upon the content that was used to define proficiency. In TIMSS 1995, there were some 20 subtests each defined by different, specific mathematics content. The remarkable thing was that country rankings, in one sense indicating the relative proficiencies of these countries, changed depending upon the specific mathematics content that served as the basis of the test. Some countries ranked in the top five among countries in one area but fell lower in the distribution in another area which then shows how important it is to be specific in defining the proficiency of interest.

The reason I share the TIMSS 1995 results is to point to what I think is the central, fundamental yet almost impossible aspect of achieving the goal this paper identifies which is measuring the content in a comparable way that is mutually exclusive and exhaustive in topic coverage that may be obliterated in higher levels of aggregation since it is that mathematics content which is used to define proficiency this is extremely critical and very difficult to accomplish. For this reason, going back to the four approaches suggested by Montoya and Tay-Lim, I personally prefer an approach which would combine policy linking (strategy 1) and item-based linking (strategy 2a). The policy linking approach would help to define the common framework and the psychometric emphasis on developing items which could then be used across studies.

References

Schmidt, W. H., McKnight, C. C., Houang, R. T., Wang, H. A., Wiley, D. E., Cogan, L. S., & Wolfe, R. G. (2001). *Why schools matter: A cross-national comparison of curriculum and learning*. San Francisco: Jossey-Bass.

Schmidt, W. H., McKnight, C., Valverde, G. A., Houang, R. T., & Wiley, D. E. (1997). *Many visions, many aims, volume I: A cross-national investigation of curricular intentions in school mathematics*. Dordrecht/Boston/London: Kluwer.

Response to Commentaries

Learning from the Center for Advanced Studies in Global Education (CASGE) Symposium at Arizona State University

Luis Crouch, RTI International and Silvia Montoya, UNESCO Institute for Statistics

We are very happy to have been invited to a Symposium on Innovations in Global Learning Metrics, sponsored by CASGE, in November 2018. Silvia Montoya and Brenda Tay-Lim from the UNESCO Institute for Statistics (UIS) wrote a paper on "Options in achieving global comparability for reporting on SDG 4," and Luis Crouch presented for them due to their unavailability. We received excellent written pre-symposium commentary from Kadriye Ercikan (University of British Columbia), Tünde Kovács Cerović (Belgrade University and Open Society Foundations), Radhika Gorur (Deakin University), and William H. Schmidt (Michigan State University), as well as many live comments from the group. The paper and written comments are here.

In this blog we want to engage with the commentary—not just respond in a simple way to specific points. One of the broad discussion questions that arose was how researchers and academics can contribute more to policy directions and to policy critique. It'd be a bit off track to engage with that point now in great depth—maybe in some other blog or venue. But one simple, direct step is for us non-academics to simply engage in the discussion. That is why we are writing this blog.

First, the UN (and other) institutions who are custodians of the measurement at this point have a mandate. There is not much choice but to follow that mandate. The UN system is a membership organization and the member countries ultimately dictate. The measurement and tracking of performance, for a set of fixed indicators, and in a manner that is as standardized and comparable as reasonably feasible, are now a mandate, given to the custodian agencies. Within the Report of the Inter-Agency and Expert Group on Sustainable Development Goal Indicators, the language is very specific: "Global monitoring should be based, to the greatest possible extent, on comparable and standardized national data, obtained through well-established reporting mechanisms from countries to the international statistical system" (p. 8).

As was wisely noted at the meeting, though, it is true that while this is a policy or even political mandate now given to the professionals, professionals (and academics) do shape the agenda and provide policy makers and political leaders with a sense of what is possible. Professionals can't entirely "hide" behind a mandate. But we honestly think that, had the policy makers truly been responsive to a technocratic agenda instead of having opinions of their own, the indicators would not be nearly as demanding on us as they are. We are being forced to stretch, especially in areas such as adult learning, civic engagement and sustainability, and digital skills. We are not sure the public and NGO researchers and officials necessarily wished this difficult challenge on themselves.

But more importantly, and as was also wisely noted in the Symposium, professionals ought to have the moral courage to engage with their mandate, not just "obey."

To us, one of the most important reasons to have comparability and standardization has nothing to do with efficiency or cost savings and accountability and so on, and has a lot to do with equity and social justice, taking as a point of departure the contents and skills the kids/youths are entitled to. If we did not have the standardized and comparable measurements that we already have, which allow us to talk based on a common language and understanding, for instance, we would not know some of the things we increasingly know, in a comparable and multi-country (that is, pretty generalizable) manner, such as that:

- About half of the global cognitive inequality is between countries, and half is within countries, at least insofar as this can be measured using assessments. Knowing this should be helpful to both governments and development agencies in setting allocative priorities.
- We have a much clearer sense of what it takes to reduce that inequality within countries—less so between them.
- We increasingly know that factors such as wealth and ethnicity/ethno-linguistic discrimination or marginalization count for more, in driving cognitive inequality, than gender or (less clearly so) the "pure" urban-rural divide.
- We also increasingly know that, because much inequality (both between and within countries) cannot be explained by any clear "ascriptive" factors (gender, parental wealth, ethnicity), "simple" (but not really so simple!) lack of management capacity and quality assurance is a real problem. And data/evidence can help here, not just in setting policy but in managing and "moving the needle" on that policy.

You can't know how much inequality exists, or what drives it, unless you measure it—with a standardized measurement stick, otherwise it is literally difficult to judge that two things are not of equal length. But, we also note that the ideal might be "as much localization as possible, as much standardization as necessary." That is why UIS's emphasis has been on supporting the comparability of existing (and future national) assessments rather than on backing, adopting, "imposing," or even endorsing specific global assessments.

Second, it was noted that measurement isn't really the issue—action by teachers and systems is. This is true, and we would certainly back the idea that there be more funding of the "improvement" function than the "measurement" function. However, improvement can more easily gain traction if one knows what is going on. (There is, of course, already far more backing of the "regular business" aspects of education systems: assessment would be a tiny fraction of that cost. However, there is under-investment in how one actually uses assessments—the right combination of assessments—to improve.) But there is still a measurement mandate aimed at making the problem visible so resources for improvement are dedicated, and, since there are efficiencies in specialization, institutions such as UIS (and their equivalents at WHO, FAO, etc.) have to focus on measurement. But perhaps such specialized bodies ought to reach out more and support others, whose mission is to use the data to support teachers (or doctors and nurses, agricultural extension agents, etc.). Along those lines, though, we also suggested (with tongue only partially in cheek) that perhaps international assessments ought to be less, not more, relevant, or at least less determinant. That is, they ought to be only a reference point (albeit a useful one), and national assessments ought to have center stage. This is UIS's position.

A last major issue that was discussed, partly in reaction to the paper but partly also because it was "in the air," was whether (and how, and why) policy research and academic input influence policy. Some were skeptical or pessimistic. Others not as much. In our view, there is impact. Not, perhaps, immediately. And few if any policy makers make decisions solely based on evidence. Nor is the impact of research typically traceable to particular academics, books, papers, or conferences—it is a much more diffuse process, which can contribute to the sensation that one is not having impact. And, of course, political economy and just plain politics have a lot of influence. But JM Keynes got it about right: “Madmen in authority, who hear voices in the air, are distilling their frenzy from some academic scribbler of a few years back… Not, indeed, immediately, but after a certain interval… soon or late, it is ideas, not vested interests, which are dangerous for good or evil.” We can cite a few examples or suggest ways to think about this that are more optimistic about research impact:

- While politics and political economy play a role, no one likes to say so publicly. Policy makers seldom say, “Oh, that was a purely political decision.” They often pay lip-service to rationality, data, evidence, as well as, at least in democracies or semi-democracies, common sense round what is right and just. Academics and researchers can take advantage of this tendency to pay lip service and demand to be heard. In a similar manner, human rights get announced before they get enacted, and they get enacted partly because they were announced and someone then used that in order to push. As noted, it is not immediate, traceable to particular individuals, etc.
- A good example is the case for girls’ education and the progress that was made over the last 40 years or so. Researchers were instrumental in this. There was not necessarily a political case. Nor was there that much grassroots pressure from villagers or even urban dwellers. On the contrary, our experience suggests that, with regard to these issues, the grassroots were pretty feudal or patriarchal. Researchers and social activists, both global and local, eventually had an impact.
- It also helps if researchers are sensitive to issues, and if they gain the initial trust of policy makers by helping with smaller, relatively short-term, and relatively less weighty matters, as a way of gaining the space to have impact on the more serious issues. This can happen with individual researchers, and with think tanks, institutions, universities and centers such as CASGE. Admittedly, this is a long game, but social development does not happen overnight.
- “Situation rooms” that show data and modeling in visually-striking ways can be helpful, under certain circumstances. Generally, only as “just one more input.” And if one is indeed not naïve about things and over-estimates the impact one is likely to have. Policy makers often react against what they see as too much naiveté on the part of researchers, when they signal that they expect policy makers to act right away on the evidence presented. But in our experience, showing the impact of simulations in real time, in a policy discussion (e.g., projecting even a simple, Excel-based model on the wall) can be useful. This varies by bureaucratic culture, of course. And it is more useful if one can take the “situation room” (again, just a simple projection of a simulation model can be useful) to the policy makers rather than having the policy makers come to the “situation room”—unless they happen to be nearby.
- Finally, it is also important to take on board the fact that it is usually local intellectuals and activists who will carry the day. UN bodies, as was noted, can’t really “make” governments take action based on data/evidence. But the data can support local intellectuals and activists who can pressure governments, e.g., in eliminating school fees, in increasing investment in the younger children, etc.

There is no time, resources, and energies for questioning the commitments themselves. The 2030 Agenda is a call for everybody. Academia is not the exception, and initiatives such as the GPE's KIX are stressing the relevance of knowledge exchange and areas where academia can play a critical role if focused on building human capacities at all levels.

The Center for Advanced Studies in Global Education
P.O. Box 871611, Tempe, AZ 85287-1611
CASGE@asu.edu
(480) 727-5346
education.asu.edu/casge

@asucasge

ASU Mary Lou Fulton Teachers College
Arizona State University

Center for
Advanced Studies
in Global Education