United Nations
Educational, Scientific and
Cultural Organization

UNESCO INSTITUTE FOR STATISTICS

Cross-national Assessments technical meeting (September 2017, Hamburg): Summary

Global Alliance for Monitoring Learning
Fourth meeting
28-29 November 2017
Madrid, Spain

GAML4/9



Meeting Objectives

Silvia Montoya (UIS) opened the meeting by welcoming participants, including representatives from major international and regional assessment programs and members of the GAML. See Annex A for meeting agenda and Annex B for full participant list.

The primary objectives of the meeting were to:

- agree on a plan for UIS's interim reporting of SDG 4.1.1;
- recommend next steps towards a longer-term solution for reporting SDG 4.1.1 based on a common metric, to improve comparability and quality of the data reported.

There is a significant lack of coverage for the indicator. Many countries do not participate in cross-national assessments (international or regional) or have national assessments. Some countries prefer to use their national assessment data for reporting purposes even if they do participate in a cross-national assessment. Moreover, the quality and scope of national assessment data varies considerably. A potential solution to the challenge of learning outcomes data from different sources is, rather than insisting on a single assessment, to link the different assessments in some manner.

Jean-Marc Bernard (GPE) also welcomed participants and emphasized the importance of making progress towards reporting SDG 4.1.1. He noted that at the UN earlier in the week heads of state from countries including France, Norway, and Senegal had come out in support of making education funding a priority. While funding for education internationally has been decreasing, this marked a significant step forward. He emphasized the importance of having learning outcomes data for education to have credibility – it is essential to be able to show leaders where the gaps in learning are so that funding can be secured. He expressed appreciation to the international and regional assessment programs that are working together to find a common framework for reporting.

Interim Reporting of Indicator 4.1.1

Ms. Montoya presented a proposed approach to reporting on indicator 4.1.1 in the short term (2018 and likely 2019), recognizing that a universal scale and linked assessments is the ideal for the longer term. The interim plan would enable UIS, as custodian agency for SDG reporting, to meet its mandate using available data, while a longer-term strategy for increasing coverage is undertaken. The interim approach is based on the following principles:

- Be as pragmatic as possible while being as rigorous as possible; and
- Build on existing work and what is available.

In practice, this means that in the short term UIS must accept data that is not perfectly aligned with 4.1.1 or comparable with other countries, but that broadly meets the needs for reporting against 4.1.1.

UIS recommends starting with data from international and regional assessments, but allowing countries to submit national assessment data if they choose. If a country does not respond to UIS's request for data then UIS will decide which data source to use. It may also be necessary to accept other learning assessment data or even unofficial data in the short term in order to cover data gaps.

Considerations for Data Sources and Reporting Standards

UIS's proposal for an interim reporting approach includes the following recommendations:

- report the definition of reading and mathematics as proposed by each assessment;
- report on in-school students, with the exclusions taken by each assessment, as well as the target grade (with -1/+1 grade, if the target grade is not cleanly defined);
- identify any assessment used in 4.1.1 reporting that includes children or young people outside of school;
- preface Indicator 4.1.1 reporting with a clear explanation that assessment programs may measure varying levels of learning progress;
- report any major operational issues, in consultation with education systems;
- report the results of this analysis, which will be collected through Catalogue of Learning Assessments Module 2;
- work with education systems to ensure that technical documentation about scaling is available in the public domain, for any assessment programs used in 4.1.1 reporting (e.g., via the Catalogue of Learning Assessments);
- preface Indicator 4.1.1 reporting with a clear explanation that assessment programs may define minimum standards of proficiency in different ways; and
- report on the periodicity of each assessment, and if it is longitudinally equated.

The proposed interim approach to reporting includes recommendations for how the data on the *percentage of students meeting minimum proficiency standards* (for the relevant domain and measuring point) would be footnoted to denote: the data source and how it was selected, population covered, whether data is based on an assessment that is longitudinally equated, and whether out-of-school youth are included in the estimate.

Mapping Target Populations and ISCED Levels/SDG Reporting Levels

To further investigate "coverage," UIS has begun to map the target populations for the different international and regional assessments against the levels of education for which 4.1.1 will be reported (grades 2/3; end of primary; end of lower secondary) in different countries. The analysis indicates that while there is an "exact match" for many countries for end of primary, there are many cases where most countries could report data one or two or even more grade levels below end of primary and for end of lower secondary there are many countries that could report data one or two

grade levels above the end of lower secondary. For grades 2/3, many countries with available data have it for grades 2/3, but many countries do not have any at all.

Comparing Benchmarks

UIS has also begun to compare the benchmarks, definitions of associated proficiency levels, and minimum levels used in each international and regional assessment.

Discussion

There was broad support for being inclusive regarding the data that are used for interim reporting.

- Support alternative data sources to fill data gaps in the short term. For example, consider using data from MICS or examinations, but include caveats.
- Do not get too caught up with issues that are not critical to precision of data (e.g., a tiny population may have little impact on the overall estimate for a population so whether they are included or not may not change what is reported)
- Remember the purpose of the indicator is to report on the percent of students (at different levels in math and reading) meeting minimum proficiency and track progress over time.
- Allow countries to use a mix of international, regional and national assessment data for different levels if that is appropriate.

- Statistically linking international and regional assessments would increase coverage and strengthen comparability of data. This would be possible in 2019 when TIMSS and many regional assessments will be administered.
- Defining minimum proficiency
  - Need a general definition of "minimum proficiency," such as what is needed to succeed at the next level of education.
  - Be careful not to place the minimum too high or it will be difficult to show progress and will be frustrating for countries.
  - While only "minimum proficiency" is required for SDG reporting, it would be wise to consider more than one level in order to provide context for the minimum level and to prevent perverse incentives.

- Out-of-school youth
  - There was recognition that it is difficult and costly to assess this population, but also that it is important not to leave them out. Target 4.1 is about *children*, not students only.
  - Some out-of-school youth have attended school previously so it cannot be assumed that they will have the lowest levels of performance.
  - Also, some assessments, such as citizen-led assessments (CLAs), are already including out of school youth.

- While UIS is only responsible for reporting SDG indicators, it is important to do so in a way that encourages the use of the data by countries to improve learning.
- Important to decide *ahead of time* what is tolerable in terms of error and bias in the reported estimates. At the same time, this is not a dichotomous decision; there could be different categories of what is acceptable, similar to how TIMSS and PISA footnote countries or put them in different parts of a table or report based on how severe the problem is.
- Include a footnote about the denominator for the percentages reported (i.e., indicate coverage of the population)

Validation and Equating of the UIS Universal Reporting Scales (UIS-RS)

Maurice Walker (ACER) and Ursula Schwantner (ACER) presented on progress towards developing universal reporting scales for SDG reporting of Indicator 4.1.1 and, in particular, a proposed validation and equating process.

Background and Progress to Date

The purpose of the UIS-RS is to maximize the number of countries that can reliably report data for Indicator 4.1.1, even if the learning outcomes are based on different assessments, while at the same time developing capacity within countries to better utilize learning assessment data.

The reporting scales were developed to be fit-for-purpose for reporting against SDG 4.1 and cover learning from foundation to mid-secondary, representing a range of difficulties and content. The reporting scales are designed to provide an overarching framework that combines key concepts and skills found to be important in cross-national assessments and countries' curricula. The scales describe *learning progressions* in reading and mathematics and can locate the *distribution of learning* observed rather than simply whether a minimum standard has been met.

The GAML has articulated three options for reporting against SDG 4.1. Taken together, these options provide flexibility for countries in the data that are used as the basis of their reporting.

- Data based on an assessment that has been equated to the UIS-RS;
- Data based on an assessment that uses calibrated items from the UIS RS item pool; and
- Data based on a national assessment that has not been equated but that is qualitatively "aligned" with the UIS-RS.

Phase I of the development of the UIS-RS—development of draft reporting scales—has been completed and the draft scales are currently undergoing a broad review process. The conceptual frameworks for the scales are based on international and regional assessment frameworks and items—including PASEC, SACMEQ, LLECE, PILNA, TIMSS Numeracy, PIRLS Literacy and others—and assessments from a broad range of countries (e.g., ASER, UWEZO, Afghanistan's MTEG). Development included: 1. developing a conceptual framework based on assessment frameworks, curriculum documents, and the relevant learning domain literature; 2. an analysis of cognitive

demands of the items and comparisons of item difficult; 3., as well as qualitative validation by comparing with other existing reading and mathematics scales.

The review process currently involves gathering feedback on the domain and level descriptions, and example skill illustrations, and overall coverage of key concepts in reading and mathematics.

Equating and Validating

Phase II focuses on equating existing assessments (e.g., international or regional assessments) with the UIS-RS using item-based equating, establishing a pool of calibrated items that can be included in an assessment (e.g., a national assessment) such that that assessment can be linked to the UIS-RS, and validating the UIS-RS.

Equating existing assessments with the UIS-RS will allow any country using one of the equated assessments to report against the UIS-RS directly, and understand how that assessment and its standards align with the UIS-RS. Further, the equating will establish a pool of calibrated items that can be embedded into an assessment that is not already equated to the UIS-RS and to determine how that assessment aligns with the UIS-RS.

The validation process will involve multiple linking exercises across 10-15 countries. In each country, sets of items from the involved assessment program will be selected and administered to one or more samples of children. Each sample represents target population of interest. After all separate linking exercises, all items that were included will together form a pool of calibrated items – this will be a central tool in the future use of the UIS-RS.

The item-based equating also provides data to empirically validate the UIS RS since the draft scales were developed based on a conceptual empirical analysis of item difficulties and the equating would provide empirical validation at the country level. While less technically rigorous than test-based equating, item-based equating, which relies on non-equivalent groups and common items, has advantages in terms of reducing the burden on countries and children and it is more flexible. The resulting pool of items can be used to link other assessments to the UIS-RS in the future.

ACER is seeking the cooperation of international and regional assessment programs, as well as others such as EGRA and citizen-led assessments, for the validation exercise and creation of the item pools.

Discussion

There was broad recognition of the value of having universal reporting scales for SDG reporting and support for continuing to work towards universal scales to support a longer-term strategy for reporting. The UIS-RS approach would provide maximum flexibility to countries to report against the SDGs and could facilitate meaningful reporting of progress towards the SDGs. However, concerns and questions were raised about the construction of the scales and approach to validating the UIS-RS.

- The proposed approach for the UIS-RS is to have a single scale for mathematics and a single scale for reading. Each would span from early learning through end of lower secondary. A

number of participants raised concerns about this, asserting that the bridge across three levels is too long to be supported conceptually or empirically. There is also a concern with trying to equate early reading given differences in scripts and structures and how children learn to read.

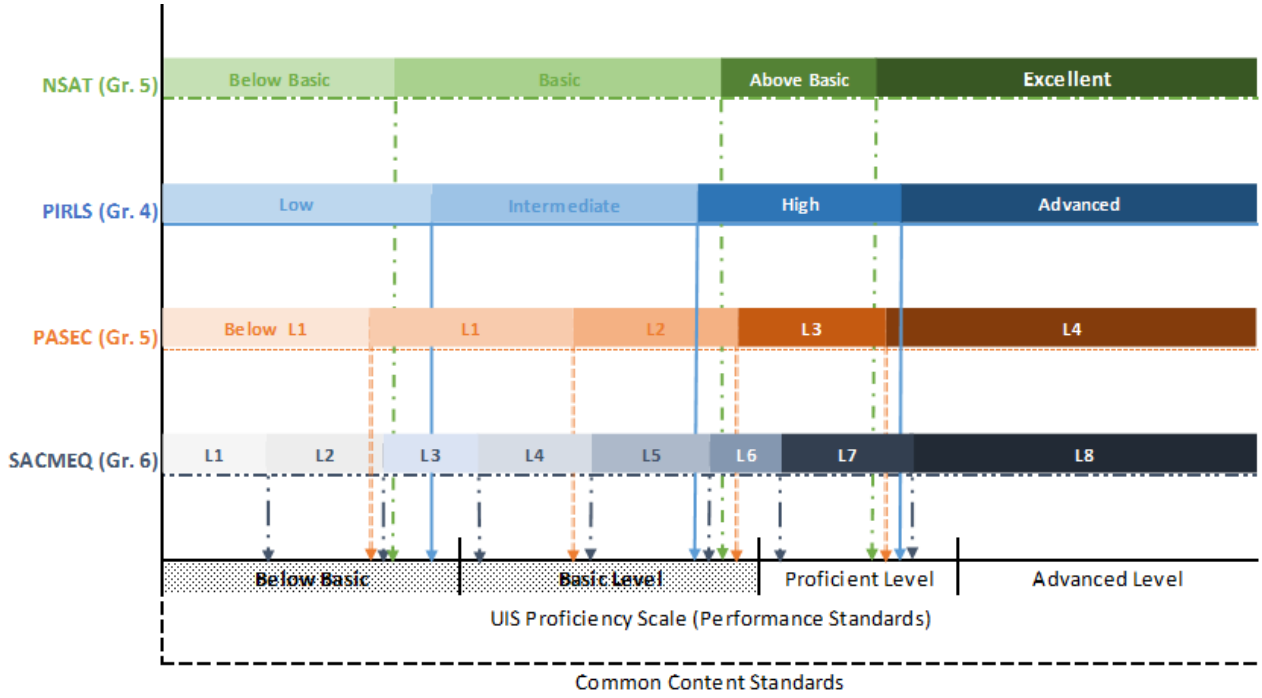
- while item based equating, and contextual matters like ordering of items within assessments may introduce uncertainty, a consideration of *acceptable* uncertainty should be made to accommodate the core agreed purpose of the UIS RS exercise: to *be as pragmatic as possible while being as rigorous as possible*.The equating and validation of the UIS-RS requires that assessment programs make items available for the equating exercises and then in perpetuity so that others can link their assessments to the UIS-RS. There is a need to find a way to propose alternatives to protect items.

- UIS reporting scales must be based on the *content* of countries' curricula and assessments based on the mapping done by IBE, so that the scales and reporting reflect what children in different countries are *learning*. UNESCO's International Bureau of Education (IBE)'s analysis of over 140 national assessments and cross-national assessments will be informative.

- There is a need to map the proficiency levels of different assessments. See Figure 1, below, for an illustration of how one can think about mapping proficiency levels used in different assessments against a UIS universal reporting scale. Again, UNESCO's International Bureau of Education (IBE)'s work to map proficiency levels used in different assessments will be informative, but a formal evaluation of alignment would need to be conducted.

- It is also necessary for UIS to define "minimally proficient" in terms of what students should know based on a Global Competencies framework for reference, and be able to do in reading and mathematics at the three levels of education. Moreover, UIS reporting should define not only "minimally proficient," which is required for SDG reporting, but also other levels to put minimally proficient in context and provide more useful information about student learning to countries.

Figure 1. Mapping performance standards on the UIS Proficiency Scale and National and Cross-National Assessments: An Example



Setting Benchmarks: Issues and Options

Dan Cloney (ACER) presented the key decisions that must be taken in order to establish benchmarks and define proficiency levels for SDG 4.1.1 reporting (in addition to defining reading and mathematics), with recommendations for the group's consideration.

Defining grades 2/3, the end of primary, and the end of lower secondary

SDG 4.1.1 implies grade-based definitions of levels of schooling by which to report learning outcomes. However, inherent in Indicator 4.1.1 is ambiguity about specific grade levels or how to account for out-of-school children. Moreover, countries vary considerably in the ages of children at different ISCED levels.

Recommendations:

- Recommendation 1 – Use ISCED as a cross-nationally standardized way of referring to the measurement points in Indicator 4.1.1.
- Recommendation 2 – Countries' specifications for the target grades that correspond to the measurement points in Indicator 4.1.1 will need to be adjudicated against an agreed set of criteria (e.g., +/-1 grade).
- Recommendation 3 – Adopt more precise interpretations (i.e., than the current "Grade 2/3") to support better cross-national comparability and to help appropriately consider the implications for an out-of-school equivalent.

Defining minimum proficiency in mathematics and reading

A fundamental question is whether minimum proficiency (or other proficiency levels) should be established against population norms or against substantive content related to curriculum outcomes.

- Recommendation 4 – Establish content-based standards that are informed by and mapped to local curricula and relevant national and international standards.
- Recommendation 5 – Conduct a content/curriculum audit across countries so that common expectations of minimum learning can be determined.

A further question is whether the standard(s) should be a point or a range on a scale.

- Recommendation 6 – Minimum proficiency levels should be established and located on the UIS RS using the descriptions of knowledge, skills and abilities in the strands for each scale.

Establishing Benchmarks

Standard setting relies on the use of expert judgment to define points (or levels) on a scale that represents a generic definition of minimum competency (or whatever levels of proficiency are desired). There are a number of procedures that can be used set standards, but all require broad expert input through a consultative process.

- Recommendation 7 – Establish panels of experts in reading and mathematics. Panel members should be selected from national nominees and have a high level of expertise in education in the relevant learning domain.
- Recommendation 8 – Benchmarks set by the expert panels should utilize existing mapping work and be submitted to a broader stakeholder consultation process before finalization.

Discussion

- Grade-level definitions of levels make sense but there are different grades spans used for ISCED levels and a lack of alignment with the language used in Indicator 4.1.1. For example, countries define “end of primary” anywhere from grade 5 to grade 7. One suggestion was to think about years of formal schooling and to think about the three levels as being after 3 years, 6 years, and 9 years of schooling, even if those don’t align perfectly with ISCED levels.

- Countries do not necessarily have data that align with the levels used in Indicator 4.1.1. For example, many countries have data for grade 4, but not grade 5 or 6 or 7. Countries will need to decide based on the available data what source to use for which levels and based on a clear reference to what are the capabilities students need to be successful in the next grade.

- There was agreement that standards should be content-based rather than norm-referenced.

- The purpose of SDG reporting is to look at progress over time and to inform policymakers about learning, not to compare countries against countries. As such, avoiding rankings and comparisons is advisable to the extent possible.

- It is preferable to have more than one benchmark at each grade point, even if only one is used for officially reporting against Indicator 4.1.1. Having more than one benchmarks helps put "minimally proficient" in context and is more informative to policymakers. The UIS RS provides a framework for doing this as it represents a continuum of abilities proximal to what will become the "minimally proficient" benchmark.

- There is a need for a technical background paper that outlines standard setting options, including the development of general (informative for policy) and detailed (content- and level-specific and informative for instruction) performance level descriptors for the UIS-RS.

Linking International and Regional Assessments

Dirk Hastedt (IEA) described a possible approach to statistically linking regional assessments with TIMSS in 2019 and PIRLS in 2021 and in doing so get more than 100 countries on the same metric. The approach would involve having TIMSS and the regional assessment (e.g., PASEC, PILNA) administered to the same students in two (although ideally more) countries in each region. This would produce a conversion table so that one could say what a particular score on the regional assessment corresponds to on the TIMSS scale (or PIRLS) and in that way "project" a country's score on TIMSS (or PIRLS). However, it would still be necessary to create a link between TIMSS (or PIRLS) and the UIS reporting metrics used for SDG 4.1.1 reporting.

The statistical linking exercise would be done once in 2019 for mathematics and 2021 for reading and repeated every 10 years or so. This could also be done for national assessments. The approach was likened to the method used to create the Purchasing Power Parity scale in which "rings" link up across regions.

Franceso Avvisati (OECD) said that this general approach could potentially be used for PISA too, in 2021. It would probably require oversampling to produce PISA grade-based estimates.

Discussion

- Participants were enthusiastic about statistically linking international and regional assessments. Representatives from PASEC and Laboratoria said they would be able to secure participation of countries. Other regional assessment representatives were positive and said they would take the idea back to their member countries.

- Statistically linking international and regional assessments would complement ACER's work to develop universal reporting scales.

- There is a need to move quickly to secure funds and get countries in place, given that the assessments will be administered in 2019.

- Dirk Hastedt will prepare a more detailed proposal including design options for linking TIMSS (and PIRLS) to regional assessments.

Conclusions and Recommendations

- Proceed with UIS' proposed interim reporting plan, taking into consideration comments and suggestions from participants.

- Continue validating and equating work in support of universal scales using item-based equating, with the intention to inform the approach of having three scales (one per "level") versus a single scale for each domain. Make technical report describing scale construction available.

- Finish and use content mapping of assessments and proficiency levels undertaken by IBE in the development of performance descriptors for UIS reporting metrics and evaluate alignment of performance descriptors used in different assessments with UIS reporting metrics.

- Produce a concept paper on social moderation as a potential method for setting performance standards for national and cross-national assessments to link with UIS reporting metrics.

- Prepare an investment case for statistically linking international and regional assessments in 2019 to support requests for funding.

- Prepare a proposal for statistically linking international and regional assessments in 2019, including possible designs and cost estimates, and that clearly lays out the logistical requirements, commitments and benefits of participants in regional studies who want to participate(IEA).

Annex A: Agenda

Day 1: Thursday, 21 September 2017	
08:30 – 09:00	**Registration**
09:00 – 09:30	**1. Opening session** a. Introduction of participants b. Objectives of the meeting ***Chair**: Silvia Montoya, UIS*
09:30 – 10:45	**2. Reporting** a. Presentation of reporting protocol – UIS b. Presentation of coverage issues – UIS i. Cross-national assessment mapping ii. National assessment mapping References: UIS concept note on reporting protocol UIS-IBE cross-national assessment mapping UIS-IBE national assessment coverage • PRESENT the immediate issues and solution on reporting in 2017 • REVIEW UIS reporting protocol • OUTLINE work of interim reporting DISCUSS and COMMENT ***Moderator**: Jeff Davies and Dana Kelly, MSI*
10:45 – 11:00	*Coffee Break*
11:00 – 12:30	**3. Reporting (cont.)** DISCUSSION ***Moderator**: Jeff Davies and Dana Kelly, MSI*
12:30 – 13:30	*Lunch*
13:30 – 15:00	**4. Linking regional and international assessments** a. Equating existing assessments and validating the UIS reporting scales– ACER b. Presentation of perspectives from cross-national assessments References: Equating existing assessments and validating the UIS reporting scales Monitoring the SDGs: an IEA's perspective • PRESENT the different options of linking cross-national assessments • IDENTIFY the technical requirements on cross-national assessments to enable linking • DECIDE on the best pragmatic option of linking • AGREE on way forward on equating cross-national assessments • DISCUSS item sourced from wide assessment programs (national and cross-national assessments) and item security • AGREE on the alignment to UIS Reporting Scale • AGREE on interim reporting DISCUSS and COMMENT ***Moderator**: Abdullah Ferdous and Jeff Davies, MSI*
15:00 – 15:15	*Coffee Break*

Day 1: Thursday, 21 September 2017	
15:15 – 16:45	**5. Options for linking regional and international assessments (cont.)** • APPLICATION from linking to reporting • DECIDE on way forward DISCUSSION ***Moderator**: Abdullah Ferdous and Jeff Davies, MSI*
16:45 -17:00	**6. Summary of first day discussion** ***Moderator:** Abdullah Ferdous and Jeff Davies, MSI*

Day 2: Friday, 22 September 2017	
09:00 – 09:15	**7. Outline for Day 2** • SUMMARY of Day 2 agenda ***Moderator:** Dana Kelly and Abdullah Ferdous, MSI*
09:15 – 10:45	**8. Extended discussion** a. Issues on reporting – establishing benchmark and defining proficiency level b. Remaining extended issues on linkages References: ACER Benchmark concept note • PRESENT different options of benchmarking to define proficiency levels • IDENTIFY the best pragmatic option • DISCUSS the way forward DISCUSSION ***Moderator**: Dana Kelly and Abdullah Ferdous, MSI*
10:45 – 11:00	*Coffee Break*
11:00 – 12:30	**9. Extended discussion (cont.)** DISCUSSION ***Moderator**: Dana Kelly and Abdullah Ferdous, MSI*
12:30 – 13:00	**10. Action items and next steps** a. Summary of the meeting b. Next action steps ***Moderator:** Silvia Montoya, UIS*
13:00	*Lunch*

Meeting documents are available at: *(google docs)*

Annex B: Participants

Australian Council for Educational Research (ACER)	Maurice Walker	Principal Research Fellow, International Surveys
	Ursula Schwantner	Senior Research Fellow
	Dan Cloney	Research Fellow
GPE	Jean-Marc Bernard	Deputy Chief Technical Officer
IEA	Dirk Hastedt	Executive Director
	Heiko Sibberns	Director IEA Hamburg
OECD	Francesco Avvisati	Analist, PISA
Pacific Community/ Education Quality and Assessment Programme	Torika Taoi	Educational Assessment Officer
PASEC	Hilaire Hounkpodote	PASEC Coordinator
	Bassile Zavier Tankeu	Technical advisor
RTI International	Luis Crouch	VP
	Simon King	Research Statistician / Analyst
SEAMEO Secretariat	Ethel Agnes P Valenzuela	Deputy Director (Programme & Development)
	Pattama Punthawangkul	Programme Officer I
The World Bank	Marguerite Clarke	Senior Education Specialist
UNESCO/Santiago	Adriana Viteri	LLECE Technical Assistant (Statistics and sampling)
UNESCO Institute for Statistics	Silvia Montoya	Director
UNESCO Institute for Lifelong Learning	Margarete Sachs-Israel	Chief programme coordinator
	Rakhat Zholdoshalieva	Programme coordinator
UNICEF	Manuel Cardoso	Education Specialist
UNICEF East Asia & Pacific	Camilla Woeldike	Project Manager – Education
USAID	Ben Sylla	Evidence Team Lead - Office of Education
Management Systems International (MSI)	Jeff Davis	Education Practice Area Lead, Technical Director
	Abdullah Ferdous	Technical Director
	Dana Kelly	Technical Director