

— Draft – do not quote —

8 May, 2018

*Developing a Monitoring Scheme
for Adult Numeracy as Part of
SDG Indicator 4.6.1:
Issues and Options for Discussion*

*Discussion Paper for the UNESCO Expert Meeting
on Adult Literacy and Numeracy Assessment Frameworks,
17 and 18 May 2018,
Paris*

Iddo Gal

Department of Human Services

University of Haifa, Israel

iddo@research.haifa.ac.il

Note: This draft has not undergone full language editing due to time constraints.

**The points of view, selection of facts and opinions expressed
are those of the author and do not necessarily coincide with official positions
of UNESCO or the UNESCO Institute for Lifelong Learning.**

Table of contents

Introduction 2

A brief overview of the PIAAC numeracy framework 4

The conceptualization of numeracy: Issues and questions from an assessment lens 10

Reporting issues and a reporting scheme for numeracy 25

Assessment methodology issues: Options for discussion 32

Bibliography 37

Appendices

Appendix 1: Definitions of reporting levels in numeracy surveys

Appendix 2: The PIAAC/ALL complexity scheme

Appendix 3: Bangladesh numeracy test

1 Introduction

1. This paper focuses on issues, dilemmas and options for assessment of the *numeracy* skills of adults, associated with Target 4.6.1¹ of the Sustainable Development Goals (SDGs). That target calls on countries to ‘ensure that all youth and a substantial proportion of adults, both men and women, achieve literacy and numeracy’ by 2030. The paper presents ideas for a monitoring and assessment strategy that can satisfy common criteria, i.e. policy relevance, credibility, comparability, feasibility in data collection, and so forth.² Methodologies developed for assessment of adult numeracy skills in high-income countries, e.g. as part of the Organisation for Economic Co-operation and Development’s Programme for International Assessment of Adult Competencies (PIAAC), are sometimes deemed inappropriate for lower-income countries, due to cost, necessary technical capacity, item content, and other factors. Hence, this paper assumes that countries interested in the numeracy monitoring process (hereafter, ‘target countries’) will be mainly those considered low- and middle-income, although some high-income countries may also be interested (according to the World Bank, at present there are 31 low-income, 53 lower-middle income and 56 upper-middle income countries, in addition to 78 high-income countries). Some of the same target countries may also opt to join PIAAC’s second cycle in 2021–2023
2. This paper takes into account recent work by UNESCO, in particular the UNESCO Institute for Statistics (UIS) and the UNESCO Institute for Lifelong Learning (UIL), and other stakeholders, and builds on the following key documents:
 - *Expert Meeting on SDG Indicator 4.6.1: Summary*, based on an expert meeting on measuring literacy and numeracy, convened by UNESCO partners (UIS, UIL) and held in Paris, November 2017 (UNESCO meeting, 2017).
 - Recommendations by UIS, (March 2018), written after the above expert meeting: Reducing financial, technical and operational burden of monitoring progress towards SDG 4.6: Options Paper (UIS options paper, 2018).
 - Earlier UIS recommendations (Dec 2017) regarding an alternative assessment tool for monitoring of SDG 4.6.1, summarized in the document, *Concept Paper on the Short Literacy and Numeracy Survey* (SLNS concept paper, 2017)
 - UIS (2017) report: *Implementation in diverse settings of the Literacy Assessment and Monitoring Programme (LAMP): Lessons for SDG 4* (LAMP report, 2017). Note: LAMP was administered in full in four middle-income countries (Jordan, Mongolia, Palestine, Paraguay), but has been piloted in others.

¹ See UN website, Goal 4 target 4.6: <https://sustainabledevelopment.un.org/sdg4>

- Review by Gal (2016), *Assessment of adult numeracy skills*, commissioned for UNESCO’s Global Education Monitoring Report (GEMR) (Gal, 2016).
3. As stated in the UNESCO meeting, 2017, UNESCO has decided to measure both literacy and numeracy for indicator 4.6.1, and to adopt for that purpose the PIAAC conceptualization of numeracy that was used in PIAAC Cycle 1 in 2013–2016. However, adaptations were needed in light of considerations described in the background documents listed above. This decision requires that UNESCO develop a new approach for defining, assessing and reporting numeracy levels, since, in the past, only literacy estimates were reported for countries.⁵ The UIS options paper, 2018, reviewed alternative assessment designs in terms of logistical demands, sampling, test instruments, costs, timeframes, and other parameters. It recommended that assessment of literacy and numeracy for indicator 4.6.1 be based on the LAMP methodology, which measured numeracy separately from literacy, but employed shortened cognitive tests, with necessary changes to the assessment design, and the possibility of attaching the data collection process to an existing national survey. The use of computer-based (tablet or laptop) administration was suggested to allow for an adaptive-testing approach that is time-efficient and cost-effective, yet still enables proficiency estimates at national level. In addition, the use of *paper*-based assessment drawing on LAMP materials was also posited as viable for some countries, given that LAMP procedures already exist for paper-based administration, whereas computer-based assessment will require extra steps.⁶ The UIS technical recommendations, and with the suggestion from the UNESCO meeting of 2017 that the conceptualization of numeracy for indicator 4.6.1 be based on the PIAAC framework, set the stage and provide a starting point for the present analysis. This paper focuses on a monitoring strategy for *numeracy*, and it reflects on issues associated with the three types of interrelated frameworks that are integral to any large-scale assessment:
- **Conceptual framework:** What to measure and why, organization of the content of the domain, its sub-constructs, and the knowledge and skills to be assessed.
 - **Reporting framework:** How will the results be reported in a way that responds to the information needs of the various stakeholders? What levels of proficiency will be used in reporting the results?
 - **Assessment framework:** The methodology that will be used, including the structure and length of the necessary tests and item pools, nature of specific items, administration mode (computer, printed, oral), scoring and other issues.
4. **Approach:** The remainder of this paper is organized in four parts. *Part 2* provides a brief review of the PIAAC numeracy framework, since it was chosen as a basis for conceptualizing numeracy for indicator 4.6.1.² *Part 3*

² PIAAC involved an optional ‘literacy components’ framework for assessing rudimentary reading skills (e.g. word meaning, comprehension of simple sentences) of adults with very low reading skills. OECD published

examines conceptual issues, while *Part 4* focuses on reporting issues and details a possible reporting scheme. *Part 5* examines two assessment designs for monitoring numeracy and describes two options: option 1 follows the approach proposed by UIS; option 2 proposes a simpler design. *Part 5* ends with a brief summary of the pros and cons of the two options, which can co-exist in the field. Some readers may wish to first skim *Parts 4* and *5*, and then work back through *Parts 2* and *3* in order to understand the issues that underlie the options discussed later on.

2 A brief overview of the PIAAC numeracy framework

- UNESCO has decided to adopt the conceptualization of numeracy used in PIAAC's first cycle (2013–2016; OECD, 2013a). That conceptualization involved a multi-faceted framework with three interlocking elements: a *definition* of the competency itself, a *model describing dimensions and specific facets of 'numerate behaviour'*, and the *numeracy complexity scheme*. Since these three elements served as the foundation for item production and interpretation of the assessment results in PIAAC, familiarization with their details is essential in order to understand the issues that may arise when applying them to the realities of monitoring indicator 4.6.1 on numeracy across a wide and diverse range of low- and middle-income countries (and possibly some high-income countries).

Table 1: Definitions of selected key constructs related to adult numeracy

<p>Numeracy and numerate behaviour – PIAAC (OECD Survey of Adult Skills, 2012)</p> <ul style="list-style-type: none">- <i>Numeracy</i> is the ability to access, use, interpret and communicate mathematical information and ideas, in order to engage in and manage the mathematical demands of a range of situations in adult life.- <i>Numerate behaviour</i> involves managing a situation or solving a problem in a real context, by responding to mathematical content/information/ideas represented in multiple ways. <p>Numeracy – LAMP (Literacy Assessment and Monitoring Programme, UIS, 2009)</p> <p><i>Numeracy</i> skills are measured using short tasks with mathematical content that are embedded in hypothetical contexts that simulate real-life situations. Successfully completing these tasks requires</p>

findings (Grotlüschen et al., 2016) regarding about 2.5 per cent of the total PIAAC sample that took the reading components. A comparable 'numeracy components' framework has not been developed so far but is discussed as an option for the next cycle of PIAAC in 2021–2023.

computing skills; estimating skills; an understanding of notions of shape, length, volume and monetary units; measuring skills; and understanding some statistical ideas or interpreting simple formulas.

Numeracy – KNALS (Kenya National Adult Literacy Survey, 2006)

Numeracy is the knowledge and skills required to effectively compute and respond to demands of diverse situations. This involves solving problems in daily life and work, and interpreting graphs, tables and diagrams. The numeracy items are divided into three main domains:

- *Number*: Operations and number line, square roots, rounding and place value, significant figures, fractions, percentages and ratios.

- *Measurement*: Related to distance, length, area, capacity, money and time. - *Space and data*: Geometric shapes, charts (bar, pie and line), tables of data.

Quantitative literacy and document literacy – IALS (International Adult Literacy Survey, 1996)

- *Quantitative literacy*: The knowledge and skills required to apply arithmetic operations, either alone or sequentially, to numbers embedded in printed materials (such as balancing a chequebook, figuring out a tip, completing an order form, interest on a loan).

- *Document literacy*: The knowledge and skills required to locate and use information contained in various formats (including job applications, schedules, tables, graphs, etc).

6. **Definition of numeracy.** The PIAAC definition is listed in *Table 1*, alongside the definition of ‘numerate behaviour’, a corollary construct discussed below, and additional definitions of numeracy and related constructs, also discussed below. Note that the PIAAC numeracy framework built on the conceptualization and assessment of numeracy in the Adult Literacy and Lifeskills survey (ALL, 2003–2006; see Gal et al., 2005), which itself built on earlier conceptualizations of quantitative literacy and document literacy in the International Adult Literacy Survey (IALS) and prior surveys. LAMP’s development drew on the IALS and ALL definitions, on which PIAAC is based. Thus, *Table 1* sketches a selected short history of the evolution of numeracy-related constructs in large-scale surveys.
7. **Numerate behaviour.** The definition of numerate behavior used in PIAAC is shown in *Table 2*. Numerate behavior is characterized using a model with four dimensions: *contexts*, *responses*, *content areas* and *representations*. For each one, key facets are listed to provide more details, i.e. about types of context, possible responses to numeracy tasks, four content areas (or types of mathematical information and quantitative ideas) for which knowledge and skills are expected, and types of representations of quantitative or statistical information. *Table 2* lists the percentages specifying the proportion of items used to cover each of the four content areas in PIAAC, attesting to their relative importance in that assessment. This model also

assumes that numerate behaviour is predicated on several enabling factors. While PIAAC created separate definitions for 'numeracy' and 'numerate behavior', the definitions of numeracy used in LAMP and KNALS merged the two constructs into a simpler structure that seemed more suitable for lower-income countries, as shown in *Table 1*.

8. **Complexity scheme.** This is an analytic tool developed for ALL and adopted by PIAAC to analyse task demands and estimate item difficulty before a piloting phase. Shown in full in *Appendix 2*, the PIAAC/ALL complexity scheme describes five factors that affect the difficulty level of numeracy tasks. *Table 3* lists the five complexity factors and (only) the description of the *lowest* (easiest) level of complexity (see Gal et al, 2005, for full background). Three of the factors were rated on a 1–3 range of difficulty, while two were on a 1–5 range, i.e. a task could be given a score ranging between 1 (low) and either 3 or 5. The simplest possible task would be given a score of 1 on all 5 factors. In internal analyses conducted during the ALL and PIAAC development phases, correlations around 0.70 were found between the complexity score given to each item based on the five factors, and the actual difficulty levels (IRT estimates) of numeracy tasks. Hence, while not perfect, the tool does provide useful approximations of expected item difficulty.
9. **The role of the PIAAC conceptual framework in assessment design.** The two schemes described in *Table 2* and *Table 3*, i.e. the dimensions of numerate behaviour and factors affecting item complexity/difficulty, are not only a part of the PIAAC conceptual framework, but also served as practical tools that guided item and scale development in PIAAC Cycle 1:
 - Numeracy items intended to cover all combinations of the facets and all four content areas in the scheme of numerate behaviour in *Table 2*. Items designed to span a range of difficulty levels, from low to high, in order to cover of the full spectrum of numeracy proficiency needed in the PIAAC countries. Given that PIAAC was designed for high-income countries, most PIAAC scales were designed to be at levels 2 to 4 (out of 5 main reporting levels); very few items were designed to be very easy and directly assess skills at Level 1, although the complexity scheme itself did describe that (low) level of numeracy.
10. **Questions?** At this stage, readers may have specific questions regarding the conceptualization of numeracy, such as about the importance of text reading as part of numerate behaviour, the relative importance of different content domains or contexts, or other concerns. Such key issues are addressed in *Part 3* later on.

Table 2: Numerate behaviour – Dimensions and facets, in PIAAC Cycle 1 (2013–2016)

Numerate behaviour involves managing a situation or solving a problem...

1. in a real context:

- everyday life
- work
- societal
- further learning

2. by responding:

- identify, locate or access act upon, use: order, count, estimate, compute, measure, model
- interpret/evaluate/analyse
- communicate

3. to mathematical content/information/ideas:³

- quantity and number (30%) dimension and shape (25%)
- pattern, relationships and change (20%)
- data and chance (25%)

4. represented in multiple ways:

- objects and pictures
- numbers and mathematical symbols
- formulae
- diagrams and maps, graphs, tables
- texts
- technology-based displays

³ The percentages next to the four content areas reflect the number of *items* that should examine each area. The percentages imply that PIAAC deemed all content areas as important, with a slight preference for quantity and number, given the ubiquity and centrality of this area in adults' lives. That said, it should be recognized that actual items sometimes require that respondents use or rely on knowledge from more than one content area at the same time. Hence, some items can be cross-classified as belonging to more than one content area. This means that the proportional coverage of each area in the item pool can be described by a range of values.

Numerate behavior is founded on the activation of several enabling factors and processes:

- mathematical knowledge and conceptual understanding
- adaptive reasoning and mathematical problem-solving skills
- literacy skills
- beliefs and attitudes
- numeracy-related practices and experience
- context/world knowledge

Table 3: PIAAC/ALL complexity scheme, with details of the lowest level of complexity
 (Note: see Appendix 2 for the full scheme of complexity factors and all their levels)

Complexity factor	What is the lowest score on this factor? (complexity score = 1) Note: Based on PIAAC Cycle 1 complexity scheme (and on ALL)
1. Type of match/problem transparency (Range: 1 to 3)	In the question/stimulus, the information, activity or operation required: <ul style="list-style-type: none"> - is clearly apparent and explicit – and all required information is provided - is specified in little or no text, using familiar objects and/or photographs or other clear, simple visualizations - is about locating obvious information or relationships only - closed question – not open-ended
2. Plausibility of distractors (Range: 1 to 3)	No other mathematical information is present, apart from that requested – no distractors
3. Complexity of mathematical information/answer required (Range: 1 to 5)	<p>Context: Very concrete, real-life activities, familiar to most in daily life.</p> <p>Quantity: Whole numbers: to 1,000</p> <p>Fractions, decimals, percentages: Benchmark fractions ($\frac{1}{2}$, $\frac{1}{4}$, $\frac{3}{4}$); decimal fractions for a half only (0.5) and equivalent as a percentage (50%).</p> <p>Pattern and relationship: Very simple whole number relations and patterns.</p> <p>Measures/dimension/space: Standard monetary values; common everyday measures for length (whole units); time (dates, hours, minutes); Simple, common 2D shapes; simple localized maps or plans (no scales).</p> <p>Chance/data: Simple graphs, tables, charts with few parameters and whole number values; simple whole number data or statistical information in text.</p>
4. Complexity of type of operation/skill (Range: 1 to 5)	<p>Communicate: No explanation is needed for a response, or a single simple response required (orally, or in writing)</p> <p>Compute: A simple arithmetical operation (+, -, x, ÷) with whole numbers or money</p> <p>Estimate: None at Level 1</p> <p>Use formula/model: None at Level 1</p> <p>Measure: Knowing common, straight-forward measures: naming, counting, comparing, or sorting values or shapes</p> <p>Interpret: Locating/identifying data in texts, graphs and tables: orientating oneself to maps and directions such as right, left, etc.</p>
5. Expected number of operations (Range: 1 to 3)	Only <u>one</u> operation, action or process.

3 The conceptualization of numeracy: Issues and questions from an assessment lens

11. This section raises issues, questions, and dilemmas regarding the conceptualization of numeracy that need to be discussed in order to provide an agreed basis for designing the required numeracy monitoring system. The previous section reviewed the conceptualization of numeracy used in PIAAC's first cycle, since it was proposed that UNESCO adopt it as a basis for the monitoring system. The PIAAC conceptualization of numeracy has so far been adopted by close to 40 countries that participated in PIAAC Cycle 1,⁴ and, hence could provide an accepted umbrella under which the assessment of numeracy for indicator 4.6.1 could be conducted. That said, as with other key constructs, the conceptualization of numeracy may also change over time,⁵ hence should be revisited by UNESCO. *Table 1* illustrated that the PIAAC definition may be consistent with the numeracy definition adopted by LAMP. Yet, the degree of this conceptual matching is an issue the UNESCO experts should discuss and confirm, if LAMP items are expected to be used as the basis for the monitoring system. Beyond the general need noted above for revisiting the PIAAC conceptualization of numeracy and its fit with the LAMP conceptualization, several more specific conceptual questions need to be answered, to enable the development of an assessment system that can yield meaningful and valid scores (Messick, 1995) for numeracy monitoring. These questions are organized below under four main topics:

- Topic 1: Skill levels to be monitored
- Topic 2: Literacy-numeracy dependencies
- Topic 3: Contextualization of assessment tasks
- Topic 4: Content areas and their relative importance

12. The answers to such and related questions can focus the planned assessment of numeracy and improve its potential to enable governments to identify skill gaps and plan corrective interventions. The four topics listed above are separated for ease of presentation, yet they are connected and influence each other. Hence, readers are asked to tolerate some overlap or fragmentation across discussions of these topics. After exploring these

4 The 40 PIAAC countries include one lower-middle income country (Indonesia/Jakarta) and four upper-middle income countries: Chile, which participated in PIAAC's second wave in 2016; and Ecuador, Kazakhstan and Peru, which are in the process of fielding PIAAC's third wave.

5 The Numeracy Expert group for PIAAC Cycle 2 (for 2021–2023) has started its work. Its review of past approaches and new needs may lead to (modest?) changes in one or more of the elements of the former PIAAC numeracy framework. UNESCO should follow-up on updates in this regard.

topics, tentative recommendations are presented regarding conceptual issues, before turning to *Part 4*, which examines reporting issues.

3.1 **Topic 1: Skill levels to be monitored**

13. **Challenges – What are the policy needs?** Deciding on skill levels to be monitored is the most important yet also complex conceptual challenge discussed in this paper, as it affects the value of the planned assessment for its many stakeholders. Indicator 4.6.1 calls on countries to ‘ensure that all youth and a substantial proportion of adults, both men and women, achieve literacy and numeracy’ by 2030. Hence, attention may be directed more to monitoring the proportion of the population at the lowest levels of numeracy, those whose skill level prevents them from effectively engaging with social and economic goals within their society. With this in mind, three key questions here are:

Q1. What skill levels (or proficiency groups) should the numeracy assessment cover? Can an ‘adequate’ skill level be defined a priori? What is it?

Q2. Should the planned assessment test some ‘component’ (very low-level) numeracy skills? Or include an oral component? Why?

Q3. If the PIAAC conceptualization of numeracy is adopted by UNESCO to inform item production, are any additions needed by UNESCO that go beyond the existing PIAAC complexity scheme, in order to cover component skills that are below the current descriptors of Level 1 tasks? Given the goals of indicator 4.6.1, there is a need to decide (a) what are the lowest and highest points of the numeracy continuum to be monitored, (b) is there a policy need, and, if so, whether it is conceptually possible to define what is an ‘adequate’ or ‘satisfactory’ level of numeracy on this continuum, via the nature of cognitive demands of the tasks to be used (not via cut-off points on a continuous distribution, as in LAMP, PIAAC). Setting such cut-off points is a thorny issue that has been debated in the comparative education literature. The need for discussing what is an ‘adequate’ level of numeracy (or literacy) is implied by the phrasing of indicator 4.6.1, which requires a country to measure the proportion of the population above a certain level. This need is also evident in the actual reporting labels chosen in national surveys, discussed in *Part 4*.

Since LAMP methodology and item pools (UIS, 2009) are recommended for use, the questions above have pragmatic implications for scale development, using existing items as calibration points to decide if the highest difficulty level of the planned tool will be as in LAMP (or lower, or higher), and if the lowest difficulty level should be as in LAMP, or if more resolution is needed at an even lower level than in LAMP (or PIAAC). Note that this issue is a conceptual one but overlaps with reporting issues, hence also addressed in *Part 4*.

14. **Comparing descriptors of numeracy levels.** To provide food for thought on target skill levels, this section first examines definitions of levels of numeracy proficiency in key surveys, shown in *Appendix 1* (which is based on an inventory provided by UIL). Later on, we examine some relevant statistics. Regarding the *lower end* of the numeracy continuum: *Appendix 1* shows that ‘Level 1’ in LAMP and PIAAC seem quite similar. However, PIAAC added a ‘below level 1’ describing more rudimentary capacities than LAMP Level 1. In contrast, the *national* descriptions from Kenya and Bangladesh vary, and describe more basic proficiencies compared to LAMP and PIAAC, while UK's Entry Level 1 describes a higher level. And, while seemingly outside the domain of adult numeracy, we should note the proposed assessment of mathematical literacy in *PISA for Development* (PISA-D), whose lowest levels⁶ cover skills that may reach below LAMP Level 1 or PIAAC Below Level 1.

Regarding *higher levels* of the numeracy continuum: LAMP Level 3 uses a rather concrete terminology compared to the more general descriptions included in PIAAC level descriptors. LAMP Level 3 describes knowledge and skills that seem mainly related to PIAAC Level 2 and at most to some of PIAAC Level 3, i.e. does not reach the more sophisticated numeracy-related knowledge and skills subsumed in PIAAC levels 3/4/5.

16. The comparison of level descriptors suggests two conclusions: (Note: This topic is split across this paper: see further discussion below, some *recommendations* regarding low-end *conceptualization* in paragraph 35; And a proposal for low-end *reporting* levels in *Table 7* on in *Part 4*).

A. What is considered ‘low numeracy’ varies across different assessments. LAMP and PIAAC levels are only partially comparable, at either the low end or high end. If LAMP items are to be used with the PIAAC conceptual framework, level descriptions and item assignments to reporting levels will need to be re-evaluated by an expert group, based on a combination of conceptual analysis and examination of psychometric data at the item and scale level. However, ultimately, the levels to be monitored should reflect the needs of policy-makers, i.e. tied to how the chosen levels can guide possible educational interventions.

B. The skills subsumed under LAMP Level 1 appear higher than the skills included in PIAAC ‘Below Level 1’. This means that the lowest-level LAMP items will need to be supplemented by additional simpler items in order to cover more rudimentary skills than what is covered by existing LAMP items. However, this issue is more

⁶ In PISA 2015, 32 per cent of all students from *low-income* countries tested in ‘Level 1a’ (the PISA ‘below level 1’ class). Hence PISA-D aims to split Level 1a into two new levels: Level 1b and Level 1c (lowest), as explained in the PISA-D 2017 draft Assessment and Analytic Framework. Level 1b includes ‘can understand questions involving everyday contexts where all relevant information is clearly given and defined in a short syntactically simple text... able to follow clearly prescribed instructions... can perform the first step of a two-step solution of a problem’. Level 1c (lowest) includes ‘can understand questions involving simple, everyday contexts where all relevant information is clearly given and defined in a very short syntactically simple text’.

complex than it seems, because there are different ways to conceptualize what are "simpler" numeracy skills (e.g. using simple print-based items or using an oral administration; each entails different cognitive demands).

17. Numeracy distributions. To explore the meaning of being at the lowest level of numeracy, *Table 4* assembles findings regarding the distribution of numeracy proficiency in the four LAMP countries, and in eight selected PIAAC countries (including three countries above the OECD average, and five below it, of which three are middle-income countries). In LAMP, proficiency was reported in three levels (1 to 3), whereas in PIAAC it was reported in five levels (1 to 5), but a 'Below Level 1' category was added. Note that all four LAMP countries were middle-income, while almost all PIAAC countries were high-income, and that LAMP was designed so that a portion of its items will test simpler skills and have reduced literacy load in order to provide more information at the low end.

LAMP results in *Table 4* show that the four middle-income countries in LAMP had, on average, between 17 per cent and 36 per cent of adults in Level 1. This figure is likely to rise higher in low-income countries with poorer educational systems and higher school drop-out rates. Note the column about LAMP *male-female differences*, which aims to disaggregate the national average. It shows that in two of the four countries, Jordan and Palestine, women perform substantially worse than men (e.g. in Palestine 36 per cent of all adults are in Level 1, but this figure includes 26 per cent of men and 46 per cent of women. This phenomenon could appear in other countries or regarding other variables, such as age, i.e. the national average masks the full picture about subgroups that may be vulnerable or in need of skill development.

A separate aspect of the LAMP results in *Table 4* involves the percentages of populations at Level 2 and Level 3. Close to *half* of the total population in each LAMP country is at level 2. Attention needs to be given to the implications of the level description, and to the number of levels. If level 2 is considered 'adequate' in terms of numeracy proficiency, this will affect the policy response that pertains to about *half of the total populations* to be measured with indicator 4.6.1. However, if Level 2 is considered as *lacking* some basic numeracy skills (as was also seen in the reporting levels used by Kenya and Bangladesh) then the population that requires a policy response is more than doubled!

LAMP results⁷ (adults 15+) (Note: Rounded percentages)

⁷ LAMP data taken from *Table 6* of the four relevant LAMP country reports (2009) at:
<http://www.uis.unesco.org/literacy/Pages/lamp-literacy-assessment.aspx>

Country	% of all adults at Level 1 numeracy	% of all adults at Level 2 numeracy	% of all adults at Level 3 numeracy	• % of males & females at Level 1 numeracy	% with primary education at Level 1 numeracy	% with secondary education at Level 1 numeracy
Jordan	25	45	30	M = 17F = 34	55	23
Mongolia	17	45	38	M = 17F = 17	65	17
Palestine	36	42	22	M = 26 F = 46	97	23
Paraguay	24	42	34	M = 21F = 27	51	13

Table 4: Numeracy rates at different levels in LAMP and PIAAC

PIAAC results⁸ (adults 16–65) (Note: rounded percentages)

	Below Level 1	Level 1	Level 2	Levels 3/4/5	Missing
Chile	31	31	26	12	0.3
Jakarta	26	34	29	11	(Indonesia) 0.01
Turkey	20	30	33	15	2.0
Spain	10	21	40	29	0.7
Israel	11	20	30	36	2.4
OECD avg	7	16	34	43	1.5
Australia	6	14	32	46	1.9

⁸ PIAAC data taken from OECD (2016), *Skills Matter: Further results from the survey of adult skills*, Figure 2.12, Numeracy proficiency among adults

Germany	5	14	32	49	1.5
Finland	3	10	29	58	0.0

15. The comparison of LAMP to PIAAC distributions in *Table 4* raises other questions that mix conceptual, reporting and assessment design issues. LAMP Level 1 captures between 17 per cent and 34 per cent of the distribution in *middle*-income countries. In PIAAC, which measured mainly *high*-income countries with better educational systems, the combined percentages at Level 1 and Below Level 1 categories are almost the same, but out of a five-level scheme. Further, the percentages of people in the PIAAC Level 1 and Below Level 1 categories rises sharply in the countries that are well below the OECD average, and reach 30 per cent to 60 per cent in Chile, Jakarta (Indonesia) and Turkey, which are all middle-income. This difference between what is captured by ‘Level 1’ in LAMP and PIAAC is important, and should lead to a discussion about test sensitivity and interpretation of what is intended by ‘Level 1’ (i.e. the lowest test level).
16. The comparison above is not a criticism of either LAMP or PIAAC, but simply suggests that LAMP Level 1 and PIAAC Level 1 do not capture exactly the same thing, when we compare the percentages in the middle-income countries. If the PIAAC ‘Level 1 + Below level 1’ categories together capture a much larger percentage of the population compared to Level 1 in LAMP, it is possible that what PIAAC considers ‘low numeracy’ is at a *higher* (more difficult) level than what LAMP understands by this term (this also explains why even in above-average PIAAC countries such as Australia or Finland, between 13 per cent and 20 per cent of the population is still placed in Level 1). In contrast, the LAMP ‘Level 1’ items seem to test lower-level skills, hence a lower percentage of the population has difficulty with them, and more persons are placed at Level 2. This could mean that the LAMP design for Level 1 items (and their psychometric parameters) has been successful, i.e. LAMP Level 1 items indeed reach lower difficulty levels than what is captured by Level 1 in PIAAC.
17. **How low is ‘low’? Can adults be ‘innumerate’?** We turn to a hidden conceptual problem with the meaning of the lowest performance level in a numeracy assessment. There is a fundamental difference between the nature of literacy and numeracy in this regard. Persons who are deemed ‘illiterate’, i.e. who cannot read and write at all, cannot automatically be considered ‘innumerate’”, because several mathematical practices are believed to be universal. Even in non-literate cultures, people engage in counting, measuring, designing, playing, or explaining with mathematical ‘objects’ using oral terms (Bishop, 1988). The literature suggests that people with little or no formal reading skills can still cope with selected mathematical tasks in everyday and work life, e.g. manage herds, conduct commercial transactions, engage in planning of construction, and perform other mental computation and other tasks that may require counting, estimation of time and distance, recognition of shapes, use of mathematical artefacts and tools, and so forth (Lave, Murtagh and de la Rocha, 1984; Greeno, 2003; Straesser, 2015). Functional mathematical skills of people with little or no

reading skills may be masked or remain undocumented either because of the *conceptualization* of numeracy (i.e. if it is defined as a skill that does not involve the ability to deal with written representations of numbers and quantities), due to *reporting* practices (i.e. combining literacy and numeracy in a single ‘non-literate’ category), or because of *assessment* practices (i.e. using a test that, for ease of administration, requires respondents to *read* written numeracy questions or *answer in writing*, or via a laptop). Whatever the explanation, these undocumented numeracy skills do have both social and economic values and, hence, conceptually are part of the construct of numeracy and should be measured. More important, from a policy perspective, information about the percentage of the population that has mental numeracy skills but cannot deal with written numeracy tasks can provide a base on which further adult education interventions can be planned, and hence may be valuable to cover as part of a monitoring tool for adult numeracy.

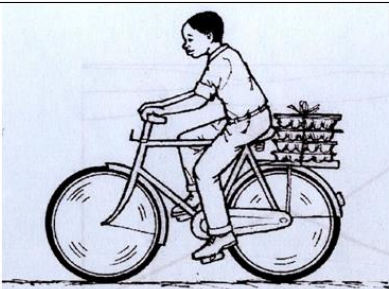

18. **Implications for necessary ‘multi-level’ discussions.** The review and analyses above suggest that there is a need for experts *and* policy-makers to discuss and attend both to defining the upper end of the numeracy continuum and to examining the need to extend the lower-end downwards. These issues are not just technical, but multi-faceted, since they combine questions about *conceptualization* (what are ‘very low-level’ numeracy tasks), *policy needs* and *reporting schemes* (do policy-makers need to distinguish ‘very low’ and ‘low’ numeracy? Are policy needs met with a three-level system?), and *assessment design* (should we add lower-level items that test more rudimentary skills to the LAMP item pool, in order to provide more resolution at the low end?). Some *recommendations* in this regard are presented at the end of this part, following the sub-sections examining numeracy-literacy dependencies, contextualization, and content issues. For sure, all this also suggests that careful *piloting* of new *items* and calibration of old and new *test scales* are needed, in order to examine how items and scales perform compared to existing LAMP items and scale with known difficulty profiles. Decisions about all of these issues have implications in terms of costs and development time.

3.2 **Topic 2: Literacy-numeracy dependencies**

19. **Challenges.** An underlying topic in assessments of quantitative (and other) basic skills is the *literacy* demands of the assessment, since the resulting test score may reflect one’s reading skill or text comprehension or writing ability, not the ‘true’ skill being measured. This issue has received much attention in the literature in mathematics education, but has a unique place in discussions about assessment of numeracy of adults. Concerns about the impact of literacy skills on numeracy performance are raised, for example, when adults have poor reading skills in the test language due to factors such as the quality of prior schooling (or lack of schooling), second-language or migration issues, age, reading practices, etc. Key questions in this regard are:
 - Q4. How or in what types of items can test design reduce the literacy demands on respondents, while overall remaining reasonably contextualized?

- Q5. Should the assessment involve an oral component which does not require the respondent to read questions or answer in writing (via a computer)?

Table 5: Sample items illustrating literacy-numeracy dependencies

<p>Item 1 (Source: Kenya's 2006 National Adult Literacy Survey)</p> <p>Question: Mr. Wafula cycles every day to the market with 4 trays of eggs. Each tray contains 30 eggs. How many eggs does he carry to the market every day?</p> <p>Answer: _____</p>	
<p>Item 2 (Source: Adult Literacy and Lifeskills survey (ALL) released item, OECD and Statistics Canada, 2005)</p> <p>Question: In total, how many bottles are there in the two full cases?</p> <p>Answer: _____</p>	

20. **Text is ubiquitous in life.** Quite often, text is an inherent part of real-life numeracy demands in work contexts and in personal and social life, as well as in learning. In a test, text can therefore be integral to what an assessment task aims to simulate, i.e. the text is an authentic element. An example could be a respondent having to read a store advertisement about a price deal, or a brief article in a newspaper regarding results from a recent opinion poll, and respond to questions about the meaning of the claims in the ad or the findings of the poll. In other cases, however, text is created by the assessment designers for the purpose of the assessment itself, and could be seen as a potential distracter or confounder that reduces the validity of the test score as an estimate of one's numeracy ability. *Table 5* illustrates these issues, using published items from a national and an international survey. The two items in *Table 5* were both used in assessments of adult numeracy. Both show a similar task: figure out the number of discrete elements in a rectangular package, and combine text and a visual display (stimulus) showing the target package. There are several differences, though, related to issues highlighted earlier: Item 1 is set in a seemingly real context, and includes a visual display, but the dimensions (physical layout) of the package are not visible, hence the question text provides both the needed context (to stimulate respondents' knowledge of what egg trays look like) and the necessary

mathematical information. Item 2 seems context-free, but its stimulus does show a real-world object that would be familiar to most adults, and the dimensions are visible, hence the stimulus itself provides the quantitative information and the question text is shorter than in Item 1. Item 1 provides written numbers and hence invites arithmetic operations using school-based procedures. In contrast, Item 2 provides quantitative information via visual means and thus allows the use of different mental operations or invented strategies. Item 1 may cause a higher cognitive load because the quantitative information has to be comprehended and memorized, while it is given directly in the Item 2 stimulus.

21. The examples in *Table 5* illustrate that the amount of text and reading demands can be adjusted by item designers, to some extent, in cases where the text is not already an integral part of the context. This issue requires trade-offs between competing design characteristics, but is seen in the literature as important in order to improve the assessment's validity. There are limits, of course, to the types of information that can be conveyed via stimulus (i.e. via a photo, drawing, chart, etc.) alone, without using text. Depending on the aims of the task, the test-taker may need more details about the context or the problem, and these must be provided within the question itself using text. This, of course, increases the cognitive demands and dependence of a numeracy score on literacy skills.
22. **Oral testing?** A question of principle is, thus, whether, for the purposes of monitoring indicator 4.6.1, there is interest in testing some facets of numeracy *without using printed texts*. This is possible via the use of picture-based images (whether on a tablet or in a paper booklet) and/or oral administration, whereby a question is read aloud by an interviewer (or digital tool) and/or the respondent answers in spoken words and the oral answer is captured by the interviewer (i.e. not written down by the respondent). A potent example for the implementation of an oral testing approach is the IVQ (*Information et Vie Quotidienne* or Information and Daily Life) survey, conducted in France in 2002–2004 to test the numeracy and literacy of 'low-level' adults. A unique feature of the test was its emphasis on oral administration (i.e. questions were read aloud by the interviewer, with content for some items shown on separate cards), to make the test available to people with low or no reading skills and avoid or minimize the impact of literacy skills on performance on numeracy tasks. Further, the numeracy test relied on a two-stage adaptive-testing (computer-based) design comprised of a screening (preliminary) test followed by one of two smaller testlets on different levels. A summary can be found in Gal (2016) and full details in Jeantheau (2005) and Murat (2008).
23. IVQ results for numeracy were reported using classical test theory approach, i.e. performance computed as percent of a total score. According to Jeantheau (2005), 32 per cent succeeded in at least 80 per cent of questions, while 15 per cent did not correctly answer 60 per cent of the questions in numeracy. Examination of connection between performance on literacy and numeracy tests showed that out of the 42 per cent who had difficulty with one or more of the three areas of literacy, 9 per cent had success in the numeracy test. In contrast, of those that showed strong performance in literacy, 8 per cent did poorly on the numeracy test.

24. **Implications for necessary discussions.** Issues around reducing literacy-numeracy dependencies, and the need for and benefits of oral testing, should receive attention when planning a numeracy monitoring strategy for indicator 4.6.1. The IVQ approach illustrates that reduction of literacy demands can improve the diagnostic value of a numeracy survey and help to identify sub-groups with unique skill profiles (see Lindberg and Silvennoinen, 2017, about unique literacy-numeracy profiles at the high end). However a task is presented, it needs to assess a [low-literacy] adult's ability to demonstrate his or her understanding of basic quantitative ideas and ability to perform selected operations 'in the head'. Examples are tasks related to quantity and number (e.g. conduct mental calculations, understand basic fractions such as $1/2$ or $1/4$), or dimension and shape (e.g. notions of length or weight).

3.3 **Topic 3: Contextualization of assessment tasks**

25. **Challenges.** The general issue of the degree of *contextualization* of tasks is a critical one as it affects item design, question wording and type of stimuli used, and creates issues in cultural adaptation of items. Contextualization issues have multiple dimensions that may be seen as having both positive and negative ramifications, as elaborated below. Relevant questions are:
- Q6. Should the numeracy test cover all four contexts listed in Table 2 (everyday life, work, societal, further learning)?
 - Q7. What proportion of the items should be used to cover each context?
 - Q8. Should the assessment use context-free items?
26. As background, the literature on 'everyday mathematics' suggests that differential performance can occur (though does not always have to) when assessment items seem 'school-like', i.e. have no context, or an artificial context, and lack a real purpose for a task given to adults, compared to items set in a realistic or familiar setting (Lave et al., 1984; Resnick, 1987). Thus, a general axiom of surveys of adult skills is that items (a task with a question and stimuli) should aim for a high degree of realism or authenticity, and be embedded in contexts that are meaningful, or reasonably familiar.
27. The rationale for the emphasis on contextualization and meaningfulness of tasks is that it can enable respondents to show their full ability to apply skills and use informal knowledge, or use 'invented' procedures that they may have gained via life experiences or numeracy/literacy practices, rather than attempt to cope with mathematical exercises that force them to draw on (faulty or faded) memories from past schooling. The use of realistic or contextualized tasks is also seen as motivating for respondents taking a 'low-stakes' survey, who otherwise may not care much for doing well on the assessment; in other words, the contextualization is supposed to improve the chances for investment of mental effort.

28. Indeed, PIAAC and LAMP present items that are contextualized, i.e. embedded in one of the four contexts listed above. Further, participating countries have been permitted (in fact, expected) to adapt contexts and stimuli to their own local conditions, so that they keep item intent (i.e. the item will test the same cognitive skills intended by the designers) but within a context that is familiar to respondents. Statistical analyses conducted both by PIAAC and LAMP show that for the most part, after such cultural adaptations, almost all items ‘behave’ psychometrically in the same way and do not show marked differential item functioning (DIF) across cultures and subgroups. That said, some recent studies have shown that sometimes the contextualization is ineffective (or perhaps tasks or stimuli were not contextualized properly) and this affects respondents' reasoning about some tasks in ways that were not foreseen by designers (Maddox, Zumbo, Tay-Lim and Qu, 2015).
29. **Examples.** While the use of tasks that lack a realistic context may seem undesirable for the many reasons given above, we should discuss to what extent the use of partially or fully decontextualized tasks may be ‘tolerable’ in some circumstances. For illustration, consider the items in *Table 6*. Item A is taken from the Kenyan KNALS survey (2011), and items B and C from the Bangladesh survey of adult skills (BBS, 2013). The three sample items are context-free, yet their use could be justified on the grounds that they aim to examine whether adults know fundamental properties of numbers, e.g. can decode written digits can identify missing numbers in a sequence (either single-digit or multi-digit numbers), or know basic symbols such as for addition or division (which may change from country to country but, often, are quite uniform within a country). On the other hand, it is more challenging to contextualize questions regarding such components skills. The use of context-free tasks can also help to standardize task demands in a consistent way regardless of the cultural context, aiding comparability of assessments.

<p>A. What is the missing number in the series below? 1, 2, 3, 4, 5, 6, 7, 8, __, 10 Answer:</p> <p>B. Arrange the following figures in ascending order: 125, 280, 70, 300, 50 Answer:</p> <p>C. Divide: 5) 25 (Answer:</p>
--

Table 6: Examples for decontextualized items in national numeracy assessments

30. **More theoretical concerns with applied implications.** Many issues are associated with the notion of contextualization, viewed broadly. Here are four separate but related concerns:
- The social setting of the overall assessment session, and its cultural relevance frame respondents' thinking and reactions ('Why do I have to do this?'). If the assessment is seen as 'low stakes' (Wise and DeMars, 2005),

this can reduce respondents' motivation, and limit their investment of effort during some assessment tasks (Maddox, Zumbo, Tay-Lim and Qu, 2015). This issue is not specific to numeracy assessment; it should inform the general assessment design and instructions to interviewers (enumerators) and their initial explanations.

- We need to worry whether the word 'mathematics' is used when introducing the assessment or its purpose, due to negative connotations that some adults associate with school mathematics or their (low) self-efficacy in this regard.
- Contextualization is important as it affects skill transfer and reasoning. How well tasks are presented in ways that fit (or do not fit) a respondent's life circumstances, numeracy practices (Coben and Alkema, 2017), or numerate environment (Evans, Yasukawa, Mallows and Creese, 2017), may affect what prior knowledge or habits of mind respondents activate or bring to bear on the given assessment problem. Testing of adult numeracy via the use of context-free school-based mathematical symbols (such as the division sign in Table 6) presents another challenge related in part to recall and memorization issues. It can be assumed that large numbers of adults (compared to pupils in school) will not remember formal notations. In countries where a sizable proportion of the adult population has not gone to school or dropped out early, or are immigrants, recall problems may be more serious due to the diversity of prior schooling contexts (and their diverse mathematical conventions), and multiple linguistic backgrounds which may affect performance.

31. **Implications for necessary discussions and decisions.** The analysis above of contextualization issues suggests that decisions need to be made regarding the nature of the contexts to be used (see the four contexts presents in the opening: *everyday life, work, societal, further learning*), and overall desired levels of contextualizing of items.

With regard to the use of *decontextualized* tasks, two specific questions arise:

- Can the use of partially or fully decontextualized tasks be 'tolerable' when assessing rudimentary mathematical knowledge or 'numeracy components'? A key example (see *Table 6* above) is if one of the goals of the assessment is expressly to determine (due to policy needs) if a person is familiar with basic *formal* notations and mathematical symbols (e.g. can decode and solve ' $2+2 =$ '), which, by definition, is context-free. Contextualization is harder to achieve (though, of course, not impossible) when designing items that test rudimentary or component skills. *Caveat*: Note that even when items are expressly designed as *written* items for the reasons given above, the underlying mathematical reasoning ability can often be addressed by an *oral* question (see paragraph 24 above) or at least by using a text-free stimulus. After all, the fact a person could not solve $2+2$ when presented in a written form, does not mean this person cannot add numbers or conduct similar basic operations in their head.
- Can the use of partially or fully decontextualized tasks be 'tolerable' for the purpose of assessing the context of 'further learning'? This context requires that learners cope with more formal types of representations, e.g.

formulas, more complex mathematical notations and symbols, which may be infrequent in the other three contexts.

32. *Adaptation guidelines*: The conceptual framework should clarify permissible changes and contextualizations by national teams during the translation and adaptation stages, either in the question, the stimulus or the overall social setting of a task. Contextualization issues surely will be central in an assessment that needs to fit the diverse contexts found in dozens of different target countries, because the contexts of questions and stimuli that seem suitable in one cultural context may not appear sufficiently familiar or valid in other cultures. Hence, specific guidelines need to be provided in this regard.
33. *Adding questions about numeracy practices in the background questionnaire*: Contextualization issues and their impact cannot be fully understood without understanding one's broader circumstances in life. Hence, there is a need to include questions in the *background questionnaire* about numeracy-related practices, as part of the numeracy conceptual framework for indicator 4.6.1.

3.4 **Topic 4: Content areas and their relative importance**

34. PIAAC has focused on the four content areas shown in *Table 2* (proportion of items in parentheses): Quantity and number (30%), Dimension and shape (25%), Pattern, relationships and change (20%), Data and chance (25%). These four content areas reflect 'big ideas' which are a cornerstone of mathematics and statistics education worldwide, and hence appear in international and national assessments, such as PISA and TIMSS, though sometimes under different names or groupings. When planning assessment of numeracy for indicator 4.6.1, it is still worthwhile reflecting on the relative importance of knowledge and skills in all of these areas, for adults in the target countries. Relevant questions are: Q9. *Should the numeracy indicator 4.6.1 cover all the four key content areas subsumed under the PIAAC scheme of numerate behavior?* Q10. *What proportion of the items should be used for each content area?*
35. Note that LAMP and KNALS covered all four content areas, as *Table 1* shows. Another point of comparison can be found in the numeracy sub-test that was included in the Bangladesh adult literacy survey (BBS, 2013), the composition of which is given in *Appendix 3*. This sub-test included 12 items, most of which focused on 'quantity and number', the rest on basic money-related issues that subsumed as part of 'dimension and shape'.
36. Further attention is needed to the coverage of the content area 'data and chance', which is made up of two separate sub-domains involving 'chance' (or probability and related topics), apart from big ideas and procedures related to 'data-analysis' or 'statistics'. First, perceptions of chance or probability of events, or the extent to which certain things are 'predictable' or 'unpredictable', may be influenced by belief systems that

are culture-bound or sensitive to the wording used. For example, questions about ‘What is the chance/probability/likelihood/risk ...’ and related terms may have different connotations or cultural nuances in different languages or subgroups, even within the same country. Hence, items related to chance or probability, such as questions about the results of rolling dice or financial risks, might produce unreliable information and have to be carefully piloted. Second, the UIS 2018 recommendation to use the LAMP item pools as a source for numeracy items raises a separate conceptual issue, related to the known overlap between the construct of *numeracy* and that of *document literacy* (defined in *Table 1*). Document literacy was part of the literacy assessment in LAMP, given its reliance on IALS. The problem is that both constructs involve some overlapping quantitative content and some statistical ideas. For example, both justify the use of tasks examining reading and interpreting graphs and numerical tables.

37. Given the concerns above, the level of coverage of data and chance has to be revisited and possibly reduced somewhat, as explained below. This, of course, requires a painful trade-off that is open to debate, given that knowledge related to data and chance, including the sub-areas of statistical literacy (Gal, 2002), probability literacy and risk literacy (Gal, 2004), are essential facets within the broad construct of numeracy, and increasingly important in adults’ lives, informing civic engagement with data-based evidence and with statistics in the social sphere.

3.5 **Recommendations related to conceptual issues: For discussion**

38. This paper proposes that the general definition of ‘numeracy’ and the model of ‘numerate behaviour’ based on PIAAC, which were reviewed in *Part 2*, will be used as a basis for monitoring numeracy for indicator 4.6.1 (note: there is a need to follow up on any updates from the new PIAAC NEG). Beyond that, more specific recommendations are listed below, to cover key aspects of the four conceptual topics examined above, and the 10 discussion questions and many of the concerns raised earlier. Note that a few other conceptual points are mentioned later when discussing reporting issues (*Part 4*) and assessment options (*Part 5*), where relevant.
39. **Tentative recommendations Topic 1: Skill levels to be monitored Regarding the lower end of the numeracy continuum:** There is a need to provide more resolution at the low end of the numeracy scale, and seek better coordination between LAMP and PIAAC item designs, in light of the different distributions in Level 1 among LAMP and PIAAC countries (see *Table 4*)a. Assuming that LAMP items will be the basis for an item pool for indicator 4.6.1, there is a need to revisit the conceptualization of LAMP Level 1 since it is higher (in terms of the wording used) than PIAAC Level 1 and Below Level 1, and to expand the item pools for Level 1 with lower-level items. LAMP level 2 conceptualization may also need checking since it included close to 45–50 per cent of all respondents.

b. The PIAAC complexity scheme already covers low-level skills in the lowest complexity level depicted in *Table 3*. It is possible to extend it downward and add a complexity level below the current *Level 1* described in *Appendix 2*: in the content area of *quantity and number* and possibly in *dimension and shape*. Adding low-level items is possible in different ways, such as the following:

- b1. Use very simple *printed*, context-free tasks requiring identification of single numbers, number line, etc. Items from Kenya and Bangladesh are examples of field-tested items.

- b2. Use tasks with *photographs or drawings of familiar objects or situations*, which offer a somewhat higher level of contextualization or at least realism, and require simple operations such as counting of objects shown, recognition of shapes, etc. This brings up issues regarding the localization of objects and scenes shown.

- b3. Use *oral* tasks involving mental operations with no dependence on print, involving simple addition, subtraction, etc.

Regarding the higher end of the numeracy continuum: The data in *Table 4* suggest that there is no merit in extending the numeracy monitoring system above the skills covered by the current LAMP Level 3 description, as the percentages in Level 3 of LAMP are already quite small (i.e. around one third), they are quite coordinated with PIAAC levels 2 and 3, and reach a difficulty level considered by OECD as ‘adequate’; hence, they can satisfy the target indicator 4.6.1.

40. Tentative recommendations Topic 2: Numeracy-literacy dependency

Items should vary in the degree to which the task is embedded in text. A few items should simulate everyday situations where mathematical ideas are embedded in text, while others should be designed to use little or no text.

- b. Offer an *oral* assessments module that involves a range of 5–6 items. Some will test component skills (e.g. number decoding, number line, shown on printed cards). Some will be fully oral (i.e. no printed text) and tasks will test via mental math higher content, mainly from quantity and number, that is equal to what is included in some Level 1 or low Level 2 printed items (e.g. addition of two single-digit numbers, division, what is a half of a given quantity). This module aims to reduce overlap with literacy scales, and enable diagnosis of selected mental operations of low-literate or illiterate persons. Note: ‘*Oral*’ is not a synonym with ‘*low*’. Mental operations, such as estimation, approximation or rounding, and the ability to conduct some mental operations in one’s head or understand relations of quantities, are also part of the skill set of any numerate person, hence testing them via oral means, as opposed to written means, can contribute to the coverage of the overall construct of numeracy, and increase the validity of the numeracy assessment.

41. Tentative recommendations Topic 3: Contextualization of tasks

- a. Contextualize as many items as possible.

b. Permit countries to adapt stimuli and details to their national circumstances.c. Allow for a limited number of context-free items (or for items embedded in ‘thin’ contexts) for an assessment of component skills.

42. **Tentative recommendations Topic 4: Content areas.** Assess all four content areas as in PIAAC (and LAMP), but consider reducing by 5 per cent the proportion of items in *data and chance*, and instead increase by 5 per cent the coverage of *quantity and numbers*, as follows: Quantity and number (35 per cent), dimension and shape (25 per cent), pattern, relationships and change (20 per cent), data and chance (20 per cent). This recommendation stems from (1) the overlap between numeracy and document literacy noted earlier, and (2) the concerns about problems with cultural comparability of questions related to chance, probability, and risk, which are otherwise important

b. In *data and chance*: avoid questions about probability or random-generating devices. Consider items that relate to risk and use percents

4 Reporting issues and a reporting scheme for numeracy

43. **Background:** In general, reporting frameworks of large-scale surveys of social or educational topics should be designed so as to respond to policy needs and provide useful information to decision-makers and stakeholders about the proportions and characteristics of key subgroups in the target populations. The reporting levels should identify groups or social categories that can benefit from interventions or policy attention – and if the data collection is repeated, enable detection of trends over time. Reporting (for a monitoring purpose) comes first; only after we agree on a reporting scheme, such as the one proposed in *Table 7*, the supporting assessment framework can be discussed (see *Part 5*).The envisioned system of indicators for Target 4.6.1 has to serve policy-makers across dozens of target countries with diverse characteristics, as well as other international agencies and stakeholders. The determination of reporting levels for such a system is a complex undertaking since many factors and influences have to be taken into account, mainly:

- reporting needs (explained above), and national or educational ‘politics’, which affect the labels or words assigned to different performance levels;
- conceptual aspects and theoretical models (about the phenomena being modeled by the indicator);assessment design, test characteristics and psychometric considerations;
- prior reporting practices;
- international standards or working norms;technical constraints.

44. As argued in an earlier analysis of assessment challenges in adult numeracy in the USA (Condelli, Safford-Ramus, Sherman, Coben, Gal and Hector-Mason, 2006; pp. 34–47), it is difficult for one assessment scheme that pertains to a complex construct such as numeracy to accommodate the information needs of all

stakeholders, even in a single high-income country. In a review of assessments for use in the UK for research or programmatic purposes regarding adult literacy and numeracy, Brooks, Heath and Pollard (2005) have demonstrated that each assessment system has unique design features that shape the types of inferences that can be made based on the data provided by the assessment. The features can pertain to available administration time, possible number of items and their scaling or difficulty levels, the designers' 'philosophy' about permissible responses (e.g. can we use multiple choice questions, or only constructed/open response?), and many other aspects.

45. Finding an agreeable path between the factors and pressures described above involves many trade-offs and difficult choices. Whatever reporting scheme is eventually adopted will never be seen as perfect by all stakeholders. For instance, in PISA, which now covers more than 70 countries, the reporting scheme involves reporting of continuous scores in the range 200 to 800, with an average around 500 and a standard deviation of 100. The publication of average scores for each country as the basic reporting mode has raised many concerns, e.g. regarding the development of 'league tables' and a race to improve averages without attention to the underlying distribution and to vulnerable groups. Further, PISA, TIMSS, PIAAC and the surveys that preceded them, such as IALS, have added seemingly value-free reporting schemes using numerical designators for segments of the distribution (i.e. Level 1, Level 2, etc) based on the standard deviation. These too have been critiqued for various reasons.
46. In the area of literacy statistics, UNESCO and most countries it serves have for decades used a dichotomous system (literate-illiterate). The advent of LAMP over a decade ago enabled UNESCO to move to a dual system similar to the one pioneered with IALS and former surveys in the USA and Canada, involving the use of a continuous literacy or numeracy score, which is divided into discrete levels for reporting. LAMP adopted a three-level system. However, as argued earlier in interpreting the data in *Table 4*, three levels do not provide enough resolution either at the low end (LAMP Level 1) or at middle of the distribution (LAMP Level 2). Based on these and other considerations explained below, this paper proposes a five-level reporting scheme, as sketched in *Table 7*.
47. **The lowest level:** We should pay special attention to the lowest reporting level, as this is where key social or educational interventions could focus and where much policy-setting may occur. The practice in LAMP, PIAAC and most other adult surveys has been to lump together three different subgroups in the lowest reporting level in numeracy (i.e. in Level 1):
 - a. Adults who have *low literacy skills* and thus had difficulty with the numeracy test which uses written questions, ending up with a low numeracy score,
 - b. Adults who have *low numeracy skills* (they may also have low literacy skills, or higher literacy skills, a phenomenon highlighted by the IVQ survey in France)

c. Adults with *no numeracy skills*, i.e. truly ‘innumerate’. With the exception of persons with severe intellectual or learning deficits, true innumeracy should be rare, both on theoretical grounds explained earlier, as well as based on results from studies of mathematical skills and practices of ‘indigenous’ or special populations who lack formal schooling and/or written scripts.

48. **Case studies for reporting levels:** In addition to the numbered level systems used by LAMP, PIAAC, RAAMA, etc., some countries have developed their own systems as part of large-scale national adult surveys on their own initiative (*Appendix 1*). For example:

Kenya used a six-level system (0 to 5) for separate scores in numeracy and literacy, but converted the levels into three partially overlapping verbal labels: *illiterate/below minimum mastery levels* (0 to 2), *minimum mastery level* (levels 3 to 5), *desired mastery level* (levels 4 to 5). In actuality, few adults in Kenya achieved Level 3, hence the two categories of ‘minimum’ and ‘desired’ were almost identical in terms of the actual percentages.

Bangladesh reported a unified literacy measure but 25 per cent of the score was based on a numeracy assessment. It used a four-level system of *non-literate*, *semi-literate*, and two levels classed as ‘literate’, i.e. *literate at initial level* and *literate at advanced level*. Those labels were not based on an analysis of the actual cognitive skills underlying each level, but on dividing the total possible score into four quarters and designating each one of them with a different label.

UK uses five levels as well, which mix verbal labels (three Entry Level categories) and two higher numerical categories (levels 1 and 2). This approach seems value-free but gently still implies what is the threshold for an ‘adequate’ level of performance. The UK system is also noteworthy because the same labels are also used by adult education programmes geared to each level, i.e. the assessment reporting levels are designed to inform interventions and synchronized with other tests and field activities.

49. **What can be concluded from analyzing the various approaches to reporting?** The diversity in reporting schemes set out in *Appendix 1* clarifies that it is not possible to fully compare numeracy rates or levels across reporting systems, given the plethora of numbers of levels and their wording. In the *national* examples above, the search for policy significance in the system of levels and labels is evident, as is the attempt to use of wording that implies what counts as an ‘adequate’ proficiency level. The number-based systems used in all *international* schemes are different and appear value-free given the much larger number of users and their diversity, but, of course, over time, countries and stakeholders do learn to use the numerical designators and apply values to them as well.
50. **Pressures:** The choice of the number of reporting levels, their actual meanings (i.e. what knowledge and skills they represent), the separation between them (i.e. how close they are) and the degree to which the labels attached to them are neutral or carry a potential social stigma, depend not only on the needs and choices of policy makers, but also on the assessment design: when more reporting levels are desired, more items are

usually needed to cover each level reliably. That said, PIAAC 'Below Level 1' was scarcely based on actual items; it covered those who could *not* answer correctly the items in the level above it, i.e. Level 1.

51. **Looking ahead:** UIS (2013) data show that compared to middle-income countries (such as the three reviewed in *Table 4*), low-income countries have lower adult literacy rates and a lower proportion of the adult population completing primary or secondary education, with a sizable percentage not attending school at all or dropping out early. Hence, if the LAMP assessment (which was finalized only in middle-income countries) was applied in *low-income* countries, compared to the statistics in *Table 4*, a higher percentage of adults in those low-income countries (possibly up to 50 per cent or more) may be classified at the lowest performance level conceptualized by LAMP. This projection strengthens the need for better resolution at the lowest level of numeracy performance for indicator 4.6.1.
52. **A tentative proposal:** Based on the analysis in the previous sections, for purposes of indicator 4.6.1 in numeracy, a possible scheme of reporting levels is sketched in *Table 7*. The scheme is comprised of five levels, inspired in part by ideas first raised in Wagner, Sabatini and Gal (1999) who proposed a four-level system, in part by the conceptual frameworks for numeracy from PIAAC and LAMP.

Readers are advised to read the list of seven 'reflection points' below, and use them to further evaluate the scheme in *Table 7*.

LEVEL	DESCRIPTION/DIFFICULTY LEVEL	COMMENTS
E	Skills related to PIAAC <i>lower</i> 'Level 3	'Adequate level
D	Skills related to PIAAC Level 2 / LAMP Level 2	
C	Skills related to LAMP Level 1 / PIAAC Level 1	
B	Can engage in some (possibly even advanced) mental calculations using indigenous number systems or measurement devices/techniques only. Knows few print-based or formal numeracy symbols and systems, though may be able to do very simply written math problems.	Based on use of minimal or no text, i.e. in part on an <i>oral</i> assessment, in part on items with text-free stimuli
A	HAS RELATIVELY FEW MENTAL CALCULATION SKILLS BEYOND COUNTING OR ADDING OF SIMPLE QUANTITIES. CANNOT RECOGNIZE THE MEANING OF WRITTEN DIGITS OR POSITIONS ON A NUMBER LINE.	

Table 7: A possible scheme of reporting levels for indicator 4.6.1 in numeracy
 Note: Proposed as a basis for discussion. Needs coordination with PIAAC and LAMP levels

Seven reflection points about the reporting scheme proposed in Table 7

53. *Table 7* shows a *conceptual* proposal for a reporting scheme that will have a good chance of being adopted by most of the target countries, especially low- and middle-income countries that may be interested in measuring and reporting on indicator 4.6.1. The shaping of a reporting scheme is a complex process that requires dealing with design conflicts and constraints. Hence, the scheme in *Table 7* has been shaped to provoke discussion about the policy needs, trade-offs, and other issues involved. Below are seven explanations and caveats that have to be taken into account when reflecting on the scheme in *Table 7* and considering what adaptations may be needed in it, and why.
54. **1 Monitoring needs – the upper end of the scheme – Level E.** The scheme in *Table 7* is meant to serve *monitoring* of Target 4.6: ‘...ensure that all youth and a substantial proportion of adults, both men and women, achieve literacy and numeracy’ by 2030. Thus, it does not aim to cover the full spectrum of possible numeracy levels, as in PIAAC, but to enable reporting of percentages of a population that reached the ‘adequate’ or ‘desired’ skill level, and support policy setting. There is no escape from having to make a value judgment and decide what proficiency level is ‘adequate’ enough, however that idea is worded. The scheme assumes that for reporting indicator 4.6.1, Level E will describe the percentage who reached an ‘adequate’ level of numeracy. The skills subsumed in Level E (which for now are not described) are to be based on the skills at PIAAC lower Level 3 (or the boundary between levels 2 and 3) because the analysis earlier, and many prior publications, suggest that they are sufficiently high to be considered ‘adequate’. Also, realistically, the percentage (in *Table 4*) reaching LAMP Level 3 or PIAAC Level 3 are sufficiently *low* in middle-income countries, so there is no need for a higher level.
55. **2 Policy needs – lower level skills – levels C to D.** Beyond setting the upper end for reporting at Level E, the scheme in *Table 7* aims to provide enough resolution at lower levels, where a sizable percentage of the population may be concentrated in many low- and middle-income countries (see Paragraph 46, and *Table 4*). This is the arena where policy-setting is likely to focus various interventions in order to raise the percentages in Level E. The scheme assumes that Level D has basic skills that are more than rudimentary, yet can enable persons to function well in some contexts. Also, they can engage in adult education programmes, using written materials since they can cope with written numeracy tasks. Countries which wish to, may combine levels D and E for further national-level discussions, if they feel the skills in Level D are ‘adequate’ for them.
56. **3 Lowest levels A to C, and literacy-numeracy dependencies.** Below Level D, the proposed scheme aims to enable a better separation between *low* formal skills in numeracy (Level C and some of B) and *no* formal skills involving written numeracy (Level B and below; i.e. ability to conduct mental operations but not using written text). This is based on (a) a theoretical necessity regarding the nature of numeracy and the ability of adults to activate mathematical reasoning, a key ‘enabling process’ numeracy performance, even without any literacy

skills, and (b) on the realization that persons to be designated in Level B cannot really engage in a numeracy programme, without first acquiring working literacy. In all, the descriptions of levels A to C also aim to enable better disaggregation of literacy and numeracy or reduce literacy-numeracy dependencies at the reporting phase.

57. **4 Wording and names of levels.** Note the use of neutral level designators with letters, from A (lowest) to E (highest) in order to avoid any confusion with the numerical level designations used by LAMP, PIAAC, PISA for Development, and other sources. For now, no word labels are proposed. However, for internal purposes, countries can develop their own verbal descriptors alongside the neutral system of levels A–E, comparable to the use of ‘mastery level’ or ‘advanced’ or ‘prerequisite’ as seen in *Appendix 1*, or following the UK usage of value-free low-end categories of ‘Entry level’.
58. **5 What is the reporting (and difficulty) level of mental arithmetic skills?** An issue that complicates the setting of a hierarchy of mathematical skills, and one which has challenged even the creation of the PIAAC complexity scheme, is the relative difficulty of simple written tasks, versus ‘high’ mental arithmetic or other ‘in the head’ processes. The literature on ‘street math’ and everyday cognition, and other types of functional mathematics, shows that some people can develop rather high computational or visual skills without the use of written, text-based, or much more formal mathematics instruction. As a result, some *mental* tasks may be more difficult than (simple) *written* problems. However, the scheme in *Table 7* includes advanced mental computation skills in Level B, even though some of them may in fact fit (higher) in Level C in terms of the conceptual complexity and the item-level psychometrics. In this regard, the current scheme is a compromise aiming to keep levels C to D linked to the PIAAC level descriptions for Level 1 to D, while acknowledging the role and importance of mental computations without the use of text or written information. However, this is a slippery issue which warrants further discussion.
59. **6 Can the assessment framework support the desired reporting framework?** The actual number of items that can be fielded for numeracy (or literacy) assessment is not large. UIS proposed to use cognitive tests that would be shorter than those used in LAMP, which already used by design relative short tests. This means that the number of levels have to be examined from a psychometric perspective to ensure that each of the levels B to E can be supported by enough assessment items (on assumption Level A does not require actual items). Rather, similar to the Below Level 1 used in PIAAC or PISA, respondents will be assigned to Level A if they fail tasks at the next higher level). If there will be severe constraints in terms of the number of items per level, the reporting scheme may have to be converged into four levels.
60. **7 The need for consensus, calibration and piloting.** The proposed scheme has to be ‘calibrated’ from two perspectives – political or conceptual, and statistical: *Political or conceptual* level: It is important to reach a sensible consensus that the skills proposed for Level E (i.e. what is an ‘adequate’ level of numeracy) are

agreeable to target countries, to ensure they are comfortable with the indicator and will invest the necessary energy and resources in data collection and reporting. After all, the diverse characteristics of different countries imply that what is ‘low numeracy’ for some may be ‘medium’ for others, depending on levels of industrialization and urbanization, formal skill demands of education programmes for out-of-school persons, or other country-specific factors (see Wagner et al., 1999). *Statistical or psychometric level:* Since LAMP items are expected to serve as the basis (though not the only source) of the item pool for monitoring numeracy, a serious *calibration and a pilot-test process* must be planned, for example to examine how LAMP Level 3 items map onto PIAAC Level 3, to ensure that the difficulty levels of the planned items, especially for levels D and E, are set properly, and to examine cultural bias, so that items perform well across a range of countries and societies.

5 Assessment methodology issues: Options for discussion

61. **Aims:** This part discusses two options for the assessment⁹ of numeracy of adults for Target 4.6.1 of the Sustainable Development Goals. Option 1 follows the approach proposed by UIS; Option 2 proposes a simpler design. The intention is not to present fully fleshed-out details, as this is premature and not possible before decisions about reporting and conceptual issues are made, but rather to provide food for thought for further discussion and debate.
62. As stated earlier, Target 4.6.1 calls on countries to ensure that by 2030 ‘all youth and a substantial proportion of adults, both men and women, achieve literacy and numeracy’. The UIS (2018) recommended to base the overall assessment on the LAMP methodology which already has an operational basis, using a computer-based (tablet or laptop) adaptive-testing approach that employs shortened versions of the LAMP cognitive tests for literacy and numeracy, with needed extensions and changes based on a revised conceptual framework. Sampling and data collection could be done via a dedicated survey, or by attaching the data collection process

⁹ In general, the assessment framework for a test of cognitive skills should be (a) coordinated with the desired reporting framework (i.e. be able to provide data needed for the number and nature of reporting levels envisioned as useful for policy-makers and data users), and (b) based on a solid conceptual framework that describes the target construct and its internal organization in terms of dimensions, facets, envisioned response types, and other elements. On the technical level, the assessment design framework should reflect the above and other parameters deemed important by the designers, and specify many details such as the desired number, types and range of necessary items in terms of content areas, difficulty levels, text loading, degree of authenticity, response mode (e.g. constructed-response vs. multiple choice), scoring logic, routing of respondents based on response patterns, and so forth (Brooks, Heath and Pollard, 2005; Gal and Tout, 2014).

as a module to an existing national survey, and using a paper-based testing as an option for countries who desire it. This approach was seen as a viable way to achieve a cost-effective assessment that still produces stable proficiency estimates at the national level.

63. The UIS proposal has clear advantages, since it enables countries to collect comparative data using accepted international benchmarks. Yet, we should consider various potential *threats* (as in a SWOT analysis): It is hard to imagine any single *survey-based* assessment scheme that can equally serve all target countries, which, according to the latest World Bank list, could include 31 low-income; 53 lower-middle income, and 56 upper-middle income countries, as well as some high-income countries that will not participate in PIAAC second wave. These target countries have diverse characteristics, including wide linguistic diversity and rural dispersions, which are much greater in comparison to many OECD countries that may take part in PIAAC. We need to take into account the sheer number of the target countries. The UIS proposal means that a single international agency (with the help of a technical consortium) has to design a multilingual computer-based assessment platform *and* a paper-based path for a huge number of scripts and cultural contexts, train personnel and conduct quality assurance regarding complex data collection operations, conduct centralized data reporting and analysis processes, and manage reporting in a timely manner. If most countries join, this has to be done for a couple of dozen of countries *each year, in parallel*, a feat not tried before. We must assume that not all countries will want or be able to implement Option 1 *in a timely* fashion. Some may not want to be locked into a single model with rather heavy resources demands that requires continuous work with an international agency.
64. **How can we promote adoption of a (numeracy) monitoring scheme?** There is a need to think how to increase the potential for widespread and early embracing and actual implementation of some type of a monitoring process, even if it is less than ideal. The rate-of-adoption issue may seem outside the mandate of the present paper, but must be flagged, as we are dealing with the ‘politics’ of comparative assessment, not just with technical issues. Note that STEP (World Bank) and MICS (UNICEF) run in waves, a few countries at a time, partly because the coordinating agencies face slow adoption rates by target countries, in part due to logistical bottlenecks. For LAMP, more than 10 countries piloted the assessment, but only four completed it, back in 2006. None have repeated it. France and the UK, high-income countries, likewise have not repeated their Skills for Life or IVQ surveys, even after a decade. Thus, the uptake of a new assessment model is a gradual process at best. Regarding adoption, we also need to consider the *timeline* for creating and fielding a numeracy assessment for indicator 4.6.1. The working assumption in this paper is that for both literacy and numeracy, the first measurement point will not start before 2020 at the earliest, given the need for preparations such as further conceptual development, item design, piloting and calibration studies. However, assessments should start in many countries on a rapid pace afterwards, since countries will need or should want a minimum of two data points (assessment cycles) between 2020 and 2030, but hopefully more, in order to evaluate progress towards indicator 4.6.1.

65. The concerns sketched above have to be borne in mind when designing the numeracy (and literacy) assessments for indicator 4.6.1. It is proposed here that we think along two pathways, to maximize the chance that as many countries as possible move early on to measure indicator 4.6.1 in some way. Thus, alongside the approach recommended by UIS, i.e. Option 1, we need to develop another approach, i.e. Option 2, in the spirit of the ‘smaller, faster, cheaper’ perspective advocated by Wagner (2011), which, by design, requires (possibly painful) trade-offs. Option 2 should be simple enough so that, notwithstanding any shortcomings it might (and, inevitably, will) have, countries will feel comfortable joining the new indicator system. This is not a simple matter since countries are burdened already by many other new SDG indicators. Thinking of a simpler option could also help to fulfil the spirit of the SDG system in general, which calls for a *sustainable* monitoring strategy, i.e. a country should be able to proceed mostly on its own, once a model for monitoring and some tools and training are made available.
66. **Assessment approach:** With the above in mind, and in light of discussions in earlier parts regarding conceptual and reporting issues, we sketch below general characteristics of two options for monitoring numeracy for indicator 4.6.1. The working assumption is that these two options will run in the field in parallel, and countries will report for indicator 4.6.1 regardless of which option they chose, using the reporting scheme with five levels described in *Part 4*. Thus, there is a need to create a common item pool and some guidelines for developing it and using it with a long-range view.
67. **A common/central item pool:** It is assumed that a general item pool for numeracy will be created by an expert group, based on existing items from LAMP, PIAAC (if an agreement is reached with OECD) and other international sources, with donated items from interested countries, and with the support of an international technical consortium and feedback from country experts on emerging items. The finalized item pool should include 40–50 items, whose mathematical content and cognitive demands are pre-determined (i.e. it is clear what each item intends to measure). The items should be created in advance to fit the four reporting levels B to E described in *Part 4* (e.g. the item pool has 8–12 items under Level B, 8–12 items for Level C, and so forth). More criteria can be established for item characteristics, in order to reduce literacy-numeracy dependencies and increase item authenticity, with items at Level B designed for oral administration or mostly using text-free stimuli, and items at levels C and D with limited text-reading demands. It may be possible to draw on selected quantitative or numeracy items that have already been adapted for use in different countries in relevant areas, such as in international surveys or tests of financial literacy, science literacy, consumer empowerment, and health numeracy. In each of these domains, there are surveys with specific items that test quantitative knowledge in sub-areas of much importance to policy-makers, and their use could both speed assessment design as well as provide for useful comparisons between ongoing assessment programmes.

It is assumed that, as with PIAAC and LAMP, countries are permitted to translate and adapt items from the common pool to local conditions and cultural preferences, within specified parameters, (e.g. in terms of

numbers, weights, measurement and time units, currency). They can also change stimuli (e.g. photos, drawings) to fit the item content to their national needs and circumstances, but otherwise have to retain the underlying *mathematical intent* of the items in terms of their mathematical and statistical demands.

68. **Option 1 – adaptive testing, mainly computer-based:** This follows the UIS options paper recommendation, summarized in paragraph 50. As Gal’s review (2016) describes, studies such as LAMP, PIAAC and STEP, or national work in Brazil, UK and France, use variants of an adaptive design for numeracy assessment. Such a design can often double the number of items in the main item pool, and the assessment as a whole covers a broader range of skills and difficulty levels and, hence, can cover the numeracy construct quite well. However, there is a need to consider how an *oral* testing component in the area of numeracy can be added into Option 1, in a way that can fit the assessment flow and routing algorithm, given the importance of identifying adults who may have limited or no literacy skills but non-trivial numeracy skills. A possible process could involve two subgroups:

a. **People who do not pass the screening** (and are routed to a paper-based test): Such respondents will be given a ‘numeracy components’ test comprised of two testlets that aim to enable determination as to whether the respondent will be in Level A or Level B (and compute a numeracy score within Level B). *Testlet 1* will be a diagnostic test with items about, e.g. recognition of numbers, using a number line, and possibly other mathematical elements to be determined by the expert group. *Testlet 2* will include *higher* level ‘mental-math’ items requiring more advanced mathematical operations and reasoning, using picture-based stimuli and oral administration. Some illustrative examples appear in *Table 5* and *Table 6*.

b. **People who passed the screening** (usually proceed with a computer-based assessment): These respondents could be given *Testlet 2*, as described above, to cover higher skills. The responses could be scored and added to the IRT-based proficiency estimation.

69. **Option 2 – paper-based test with a simple ‘semi-adaptive’ process:** This option aims for a simpler assessment model that does not use a computer-based administration, and can be implemented in the field without the ongoing involvement of an international agency (unless one is desired or mandated for quality assurance purposes). As a background, the Gal (2016) review of operational aspects of national and international assessments, and the UIS options paper, 2018, suggest that low- and middle-income countries can field on their own a typical large-scale household-based survey that affords 60–80 minutes of testing time per respondent. Within this timeframe and situation, *it is possible to employ 15 to 20 relatively short numeracy items*, as Kenya and Bangladesh demonstrated, each under very different operational conditions. It is proposed that under these operational assumptions, countries which do not wish to use Option 1 will be offered Option 2: administer a simple paper-based test and an assessment flow (routing), using a screening process similar to the one used in Option 1 for respondents who cannot use or refuse to use a computer. Based on this screening,

a decision can be made as to which of the two testlets respondents will be given: *Testlet 3* will be a print-based numeracy test (questions in a booklet) with 15–18 items covering levels C, D and E (each with 5–6 items, all taken from the common pool which is also used in Option 1). Those who fail the screening will be given *testlets 1 and 2* already described above, and could also be given a *subset of testlet 3* with only the items covering Level C, for full coverage of levels A–C. Calculation of scores can be based on a classical test approach, i.e., percent of correct responses. The mode of score computation will have to be further discussed, to clarify how scores are then converted into the reporting scheme, i.e. onto a Level A to Level E scale.

70. **Summary of assessment options:** This part of the paper has shown that at least two options exist for an assessment design for indicator 4.6.1 in numeracy, each with some important advantages and disadvantages or costs. Option 1 appears superior to Option 2 on all technical grounds, i.e. it can create scores that are more reliable and valid since it uses a longer test, covers the construct better, and has access to more sophisticated statistical procedures for proficiency estimates. However, it was argued that there may be adoption issues as well as logistical bottlenecks with Option 1. Nonetheless, several preliminary ideas were sketched to show that, within Option 1, the numeracy assessment should and can be expanded in light of concerns raised in this paper. The use of two testlets for the low-end of the numeracy continuum can enable a country to report the percentage of people in all five reporting levels A–E, by collating information from separate testing routes, just as it is done already in LAMP and PIAAC. Thus, Option 1 can fit the suggested reporting scheme for numeracy, as described in *Part 4*.
71. The new Option 2 process proposed here enables a country to produce a quick estimate of a person’s numeracy level, within a short testing time, using common items from a centralized international item pool. In this way, all countries cover the same four content areas, and address selected aspects of the model of numerate behaviour. Option 2 is likely to be inferior to Option 1 in terms of psychometric quality and the assessment information value because it is based on a smaller number of items and cannot use IRT estimates (which estimate proficiency for the whole population). Nonetheless, the test to be used in Option 2 will be consistent across countries (using items selected from the same item pool used for Option 1), hence Option 2 can yield a continuous numeracy score that enables global comparability, and its results can fit into the same five reporting levels proposed.

Yet, Option 2 has additional several advantages: Testing for Option 2 can be fielded within a household survey format that is already common in all countries, and the cost aspects should not be much different than current costs for such surveys, unlike Option 1, which is more costly and time-consuming. Hence, Option 2 can enjoy a fast adoption rate and a higher chance that the assessment will be repeated in frequent intervals, thus providing countries a better way to sustain their monitoring process in the long run, and generate monitoring statistics on a faster pace compared to Option 1.60. Overall, Option 2 is proposed here for its potential to offer an alternative approach that may be sufficient for the purposes of some countries in terms of monitoring

indicator 4.6.1. For countries that will not be able or willing to adopt Option 1 as proposed by UIS, Option 2 will enable such countries, for the first time ever, to report the percentage of their citizens at multiple levels of the numeracy proficiency continuum, using a common item pool and common reporting scheme and terminology. UNESCO can choose whether to report two systems of scores (given the psychometric differences between the two options), or report a single score on the five proposed levels (with the score based on either one of the two measurement options). Either way, UNESCO will have a more widespread system that covers more countries than would be the case if only Option 1 were offered.

6 Bibliography

- Bangladesh Bureau of Statistics (2013). *Literacy Assessment Survey (LAS) 2011*. Online: Coben, D., & Alkema, A. (2017). The case for measuring adults' numeracy practices. *Journal of Research and Practice for Adult Literacy, Secondary, and Basic Education*, 6(1), 20–32.
- Bishop, Alan J. (1988). The interactions of mathematics education with culture. *Cultural Dynamics*, 1(2), 145-157.
- Brooks, G., Heath, K., & Pollard, A. (2005). *Assessing adult literacy and numeracy: a review of assessment instruments*. London, UK: National Research and Development Centre for Adult Literacy and Numeracy. [online: www.nrdc.org.uk]
- Condelli, L., Safford-Ramus, K., Sherman, R., Coben, D., Gal, I., & Hector-Mason, A. (2006). *A review of the literature in adult numeracy: research and conceptual issues*. (Prepared by American Institutes for Research, for the Adult Numeracy Initiative of the US Department of Education's Office of Vocational and Adult Education). Online: <http://eric.ed.gov/?id=ED495456>
- Evans, J., Yasukawa, K., Mallows, D., & Creese, B. (2017). Numeracy skills and the numerate environment: Affordances and demands. *Adults Learning Mathematics: An International Journal*, 12(1), 17-26.
- Gal, I. (2002). Adults' Statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, 70(1), 1-25.
- Gal, I. (2005). Towards 'probability literacy' for all citizens. In G. Jones (ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 43-71). London: Kluwer Academic Publishers.
- Gal, I. & Tout, D. (2014). *Comparison of PIAAC and PISA Frameworks for Numeracy and Mathematical Literacy*. OECD Education Working Papers, No. 102, OECD Publishing. Online: <http://dx.doi.org/10.1787/5jz3wl63cs6f-en>
- Gal, I. (2016). Assessment of adult numeracy skills. Paper commissioned for the *UNESCO Global Education Monitoring Report 2016: Education for people and planet*. Paris, UNESCO. Online: <http://unesdoc.unesco.org/images/0024/002455/245573E.pdf>
- Gal, I., van Groenestijn, M., Manly, M., Schmitt, M. J., & Tout, D. (2005). Adult numeracy and its assessment in the ALL survey: A conceptual framework and pilot results. In Murray, S. T., Clermont, Y., & Binkley, M. (Eds),

Measuring adult literacy and life skills: New frameworks for assessment (pp. 137-191). Ottawa, Canada: Statistics Canada.

- Greeno, J. G. (2003). Situative research relevant to standards for school mathematics. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.). *A research companion to Principles and Standards for school mathematics* (pp. 304-332). Reston, VA: National Council of Teachers of Mathematics.
- Grotlüschen, A., Mallows, D., Reder, S., & Sabatini, J. (2016). Adults with low proficiency in literacy or numeracy. Grotlüschen, A., et al. (2016), "Adults with Low Proficiency in Literacy or Numeracy", *OECD Education Working Papers*, No. 131, OECD Publishing, Paris, <http://dx.doi.org/10.1787/5jm0v44bnmnx-en>.
- Jeantheau, J. P. (2005). Assessing low levels of literacy: IVQ Survey 2004-2005-Focus on the ANLCI module. *Literacy & Numeracy Studies*, 14(2), 75.
- Kenya Bureau of Statistics (2007a). *National Adult Literacy Survey: Literacy assessment*. Retrieved Jan 15, 2016 from: <http://statistics.knbs.or.ke/nada/index.php/catalog/58>.
- Kirsch, I. S., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). *Adult literacy in America*. Princeton, NJ: Educational Testing Service, National Center for Education Statistics.
- Lave, J., Murtagh, M., & de la Rocha, O. (1984). In B. Rogoff & J. Lave (Eds.), *Everyday cognition: Its development in social context*. Boston: Harvard University Press.
- Lindberg, M., & Silvennoinen, H. (2017). Assessing the basic skills of the highly educated in 21 OECD countries: An international benchmark study of graduates' proficiency in literacy and numeracy using the PIAAC 2012 data. *Comparative Education*, DOI: [10.1080/03050068.2017.1403676](https://doi.org/10.1080/03050068.2017.1403676)
- Maddox, B. (2015). The neglected situation: Assessment performance and interaction in context. *Assessment in Education: Principles, Policy & Practice*, 22(4), 427-443.
- Maddox, B., Zumbo, B. D., Tay-Lim, B., & Qu, D. (2015). An anthropologist among the psychometricians: Assessment events, ethnography, and differential item functioning in the Mongolian Gobi. *International Journal of Testing*, 15(4), 291-309.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9), 741.
- Murat, F. (2008). Adult skill assessment: Emerging methods. *Education Formations*, 78, 81-92. Retrieved Jan 5, 2015: http://media.education.gouv.fr/file/revue_78_Anglais/01/3/EDUC&FORM_EN_41013.pdf#page=80
- OECD (2013a). *OECD skills outlook 2013: First results from the survey of adult skills*. Paris: OECD Publishing. doi: 10.1787/9789264204256-en.
- OECD (2013b). *The Survey of Adult Skills: Reader's companion*. Paris: OECD Publishing. <http://dx.doi.org/10.1787/9789264204027>
- Straesser, R. (2015). Numeracy at work: a discussion of terms and results from empirical studies. *ZDM*, 47(4), 665-674.

- Tannenbaum, R. J., & Katz, I. R. (2013). Standard setting. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol. 3. Testing and assessment in school psychology and education* (pp. 455–477). Washington, DC: American Psychological Association
- Tout, D. and Gal, I. (2015). Perspectives on numeracy: Reflections from international assessments. *ZDM–International Journal of Mathematics Education*, 47(4), 691-706.
- UNESCO (2004). *The plurality of literacy and its implications for policies and programmes*. Paris: Author. Online: <http://unesdoc.unesco.org/images/0013/001362/136246e.pdf>
- UIS (2009). The next generation of literacy statistics: Implementing the Literacy Assessment and Monitoring Programme (LAMP). Online: <http://www.uis.unesco.org/Library/Documents/Tech1-eng.pdf>
- UIS (2017). Implementation in diverse settings of the Literacy Assessment and Monitoring Programme (LAMP): Lessons for SDG 4. Unpublished working paper.
- UIS (2018). Concept Paper on the Short Literacy and Numeracy Survey (SLNS). Unpublished working paper.
- UIS (March 2018). Reducing financial, technical and operational burden of monitoring progress towards SDG 4.6: Options Paper. Unpublished working paper.
- UNESCO (2017). Expert Meeting on SDG Indicator 4.6.1: Summary. Unpublished working paper.
- Wagner, D. A. (2011). *Smaller, Quicker, Cheaper: Improving Learning Assessments for Developing Countries*. Education for All Fast Track Initiative. Paris: International Institute for Educational Planning. Wagner, D. A., Sabatini, J., & Gal, I. (1999). *Assessing basic learning competencies among youth and young adults in developing countries: Analytic survey framework and implementation guidelines*. Philadelphia: International Literacy Institute/UNESCO. (Online: <http://literacyonline.org/products/ili/pdf/op9901.pdf>)
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10(1), 1-17.