GAML5/REF/4.6.1-02

— Draft – do not quote —

3 May 2018

Form follows function: A global framework for assessing and reporting literacy

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

T. Scott Murray

DataAngel Policy Research

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

Executive Summary

The following paper documents:

- How literacy has been defined, measured and reported across a range of national and international assessments.
- What uses are served by data on the level and distribution of literacy by proficiency level.
- What the listed uses imply for the statistical properties of the required estimates.

The paper also sets outs the design choices that must be made to guide the design of a global measurement framework, with particular reference to the lower levels of literacy, and what approaches yield valid, reliable, comparable and interpretable estimates.

All assessment programmes share a common objective of informing a broad range of public policies. To serve the purposes for which they are designed, assessment programmes must be based on a solid conceptual framework and report results in ways that are accessible and meet the needs of key groups of users.

At their most basic, assessment programmes need to identify that there is a literacy 'problem' that merits public attention, to decide who needs what help and what level of investment is implied.

Fundamentally, however, measures that shed light on the type of instruction needed to remedy the 'problem' of different groups of people needing help are significantly more useful.

To serve the full range of policy, needs assessment results must be valid, reliable, comparable and interpretable.

For developing countries, assessment programmes also need to be manageable, in the sense that they do not impose too high a technical and operational burden, and are affordable, in the sense that, at a minimum, measurement does not divert too much of the available resource from instruction.

The paper discusses the following topics and proposes questions for the experts to consider in reflecting on the development of recommendations: data uses; criteria for assessment, including validity, reliability, comparability and interpretability; definitions and measurement of literacy; improvement of measurement at the lower end of the scale; face validity/familiarity; and definition of proficiency levels. It also discusses additional conceptual issues which impact the ability to monitor literacy globally, including the definitions of functional literacy. Lastly, it offers some reflections on the technical and operational demands of assessments.

1 Introduction

This paper proposes a framework that will guide and inform the assessment of adult literacy skills globally. The content is meant to inform a discussion among assessment experts who will attempt to reach a consensus regarding key design choices.

Valid, reliable, comparable and interpretable estimates of average literacy scores, and the distribution of literacy skill by proficiency level, are needed for the overall adult population, and for key population subgroups, in order to support a range of purposes, including monitoring Sustainable Development Goal (SDG) Target 4.6.

Experience suggests that meeting these needs is technically, operationally, financially and politically demanding. Assessment programmes are only successful if they manage the inherent problems in a coherent way that reduces the level of risk of catastrophic and irreversible error to acceptable levels.

In an effort to achieve the requisite level of coherence, this report is organized in chapters, each of which addresses a fundamental aspect of design:

- This chapter, *Chapter 1*, introduces the report and outlines its objectives
- *Chapter 2* documents what data on literacy users need.
- *Chapter 3* identifies criteria that data must meet.
- *Chapter 4* provides a brief overview of definitions of literacy, what a framework is and how it informs the approach to measurement and, ultimately, the ability to support intended uses.
- *Chapter 5* provides an overview of the definition of proficiency levels and compares and contrasts a number of selected studies.
- *Chapter 6* compares approaches to measurement devoted to the lower regions of the literacy proficiency scale.
- *Chapter 7* discusses fixed proficiency levels.
- *Chapter 8* discusses a series of additional issues for design for literacy assessment.
- *Chapter 9* discusses options for assessment methodologies.

2 Uses of data

Theory provides a framework for classifying the uses to which official data is put.

A clear statement of intended uses is important because the fitness for use of any data produced by an assessment system can only be judged against the purposes to which the data will be put.

As documented in the following table adapted from the World Bank's volume on using assessment results, comparative data on the level and distribution of adult literacy skill is needed to serve five distinct purposes, each of which imposes a unique set of statistical demands to be fit for use (*Table 1*; see also UIS options paper 2018):

Table 1. The uses of data on literacy

Application type	General purpose	Related policy questions	Implication for data collection strategy
Knowledge generation	Identification of the causal mechanisms that link skill to outcomes. These data provide reasonable expectations on how rapidly skill distributions will respond to policy initiatives.	How do individuals acquire skill? How do they lose skill? How are skills linked to outcomes? What is the average skill level and distribution of skill for different age groups? What are the levels of social and economic demand for skill? Are they sufficient to meet national goals? How efficient are the markets that match skill supply and demand?	Needs longitudinal or repeated cross-sectional data with comparable measures of skill.
Policy and programme planning	Planning government response to identified needs to meet social and economic goals.	Which groups need skill upgrading? How many people are in need? Where is need concentrated? What measures are needed to improve market efficiency?	Needs profile of skill for key population subgroups.

Application type	General purpose	Related policy questions	Implication for data collection strategy
		Are measures needed to increase skill demand?	
	Determination of funding levels.	How much budget is needed to raise skills at the rate needed to achieve social and economic goals?	Need numbers of adults with different learning needs.
Monitoring	Adjustment of policies, programmes and funding levels.	Are skill levels rising at the expected rate? If not, what additional policy measures and programme investments are needed?	Needs repeated cross-sectional skill measures
		Are skill-based inequalities in outcomes shrinking?	Needs repeated cross-sectional skill measures for key population subgroups.
Evaluation	Formal process to determine if programmes are performing as expected.	Are government programmes effective? Are they efficient?	Needs data on skill gain/loss and costs for programmeparticipants.
		Are government programmess meeting their objectives?	
Administration	Making decisions about specific units: individuals, regions, programmes.	What criteria are applied to determine programme eligibility? To allocate funding to programmes?	Needs results that are reliable enough to keep type I and type II individual classification errors to acceptable levels and that can be aggregated to the programme level.

Most importantly, national governments need comparative data to set policy and programme priorities, to make the case for international support, to establish national funding allocations, to monitor progress towards stated targets, to evaluate the efficacy of public investments in skill generation and to administer programmes, as well as progress towards SDG 4.6.

Multilateral and bilateral donors also require comparative data to guide their policies and programmes and to monitor progress towards international and national targets, including SDG 4.6. SDG 4.6 has its own set of requirements that need to be met by the proposed assessment system. Specifically, the target states that 'By 2030, ensure that all youth and a substantial proportion of adults, both men and women, achieve literacy and numeracy'. The global indicator for SDG 4.6, the only indicator for this target directly related to the measurement of learning outcomes, is indicator 4.6.1: the percentage of the population in a given age group achieving at least a fixed level of proficiency in functional (a) literacy and (b) numeracy skills. The target age group for this indicator is the population of 15 years and older.

Translated into statistical terms, indicator 4.6 implies a need for:

- separate measures of literacy and numeracy;
- measures that are statistically representative of the adult population;
- measures that capture the full range of skills possessed by the adult population;
- measures that can be safely compared, at a point in time and over time;
- measures that are sufficiently precise to detect economically and socially meaningful change over the reference period.

Assessment programmes that address both national and international targets offer more value. As noted in *Table 1*, both uses imply a need for measures of literacy that can be compared over time to determine relative need and to track progress.

2.1 Topic for discussion

- Main uses of data which an assessment strategy should seek to meet

3 Criteria that define fitness for use

The data uses enumerated above provide a means to specify a set of criteria that define the statistical properties that any associated data system needs to generate and against which alternative assessment strategies may be judged.

This analysis identifies four criteria that must be met. Specifically, estimates of literacy and numeracy skill need to be:

- valid
- reliable
- comparable
- interpretable

Each of these criteria is detailed below.

3.1 Validity

Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment (Messick, 1989b).

Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores. These scores are a function not only of the items or stimulus conditions, but also of the persons responding and the context of the assessment. In particular, what needs to be valid is the meaning or interpretation of the score; as well as any implications for action that this meaning entails (Cronbach, 1971).

It is worth pointing about a philosophical aspect of the approach to measurement employed in the current set of international comparative assessments. These assessments set out to assess adults' ability to cope with unfamiliar reading and numeracy tasks as it is this ability that confers independence and agency. Independence and agency are keys to adapting to change, whether externally or internally imposed.

In the current context, the validity rests on reliably placing an individual on the proficiency scale and, by extension, identifying their learning needs. Accurate placement on the proficiency scale allows one to compute how many points they are away from the proficiency level needed to meet their own objectives and/or to meet collective social and economic goals.

For adults classified at Levels1 or Level 2 on the IALS/ALL/PIAAC/LAMP and STEP scales, it is far more difficult to assess what kind of instruction would need to be offered to move them up the scale. Analysis of the reading components data from the 2005 International Survey of Reading Skills (ISRS), PIAAC, LAMP and STEP studies data reveals that groups of learners can have quite different instructional needs despite being at the same place on the scale.

This finding raises fundamental questions about what inferences the framework must support. Apart from identifying where someone is on the literacy scale the proposed system should provide a clear indication about what type and amount of instruction would be needed to move each group up the scale.

3.2 Reliability

In this context, reliability denotes the 'consistency' or 'repeatability' of test results.

For practical purposes, reliability implies that, if the same individual was tested with the same test, or was tested with a different test that includes items that provide an equivalent sample of the determinants of item difficulty, one would get essentially the same result. In this context, the term 'essentially' is defined in terms of the precision of the two test results. Specifically, they do not need to be identical but need to offer a result that leads to the same decision or action. In this sense, reliability links back to validity since construct validity can only be judged in terms of the measure's ability to support a given action. More directly, to been judged reliable both measures must display the same magnitude of Type I and Type II classification errors.

Each of the data uses enumerated above place a distinct set of statistical demands on the measures.

Meeting the need to profile determinants and outcomes implies a need for the application of multivariate methods that demand relatively small sample sizes, i.e. roughly 60 completed cases in each cell to be included in the analysis.1

Meeting the objective of generating point estimates of average scores and numbers and proportions of the population at proficiency levels requires higher sample sizes, i.e. 100 to 400 completed cases per population subgroup for which data is needed by design. For this reason, international adult skill assessments have tended to field average samples large enough to yield completed cases for 5,000.

Meeting the objective of estimating the social gradient in literacy skill requires that an internationally comparable measure of socio-economic status be carried on the background questionnaire.

Meeting the objective of estimating the economic demand for literacy skill requires the collection of information that allows occupation to be coded to the four-digit level. Literacy demand levels are then assigned to each occupation and aggregated.

3.3 Comparability

In order to be useful, measures of literacy and numeracy have to be comparable. In fact, comparability is fundamental to the goals of assessment, as comparison allows one to identify which individuals and population subgroups are most at risk from skill-based disadvantage, to identify the relative level of need across countries and to monitor the rate at which the literacy 'problem' is getting better or worse.

The uses set out above imply a need for several dimensions of comparability.

First, results need to be **comparable within heterogeneous national populations**.

Second, results need to be **comparable across countries**.

Third, both nationally and internationally, results need to be **comparable over time.**

The fundamental issue facing policy-makers is whether the supply of skill is growing rapidly enough to reduce the size of literacy skill shortages and the levels of associated skill-based inequality and, prospectively, to meet social and economic objectives.

[1] Assuming a design effect of 2.

Comparability is not something to be assumed but is, rather, something that must be empirically demonstrated.

Current national and international literacy and numeracy assessments are designed to yield valid, reliable and comparable estimates of skill for population subgroups rather than individuals.

The introduction of computer-based, adaptive tests provides test-developers with a means to circumvent the problem of unacceptably high levels of test burden.

Fully adaptive tests also address one of the criticisms of current assessment practice, i.e. that unfamiliarity with the cultural content of test items introduces uncorrectable bias into the proficiency estimates.

3.4 Interpretability

To be useful, literacy and numeracy estimates need to be interpretable, which means that:

- differences in skill are associated with material differences in socially and economically valued outcomes;
- the observed differences in skill have been shown to be causal;
- skill levels can be improved through teaching and learning;
- the average level of skill and the distribution of skill can be influenced by policy.

Evidence summarized in *Chapter 3* confirms that literacy and numeracy are associated with large differences in individual, institutional and societal outcomes.

Causality has been established in several ways, including through the analysis of longitudinal data that repeated skill measures and a broad range of outcomes measures.

Several large-scale skill upgrading pilots undertaken in Canada establish that instruction can precipitate material skill gains in heterogeneous adult populations (DataAngel, 2017).

Unequivocal evidence of the causal relationship between literacy and numeracy skills and individual labour market outcomes and firm performance has been obtained through the conduct of a large-scale, two-stage randomized controlled skill upgrading trial in Canada (SRDC, 2014).

Macro-economic modelling undertaken with IALS, ALL and PIAAC data provides strong evidence in support of a causal link between key indicators of macro-economic performance – differences in long-term rates of GDP and labour productivity growth –and literacy skill (Coulombe, Tremblay and Marchand, 2007). Increases in average skill and reductions in numbers of adults with skills at levels 1 or 2 have been found to have a strong, positive impact on growth.

Analysis of PISA and IALS/ALL/PIAAC data provides clear evidence that policy can have a rapid and positive impact on the level and distribution of literacy and numeracy skill. The 2018 World Development Report includes examples of policy in less-developed countries precipitating rapid improvements in the skill levels of primary and secondary students (World Bank, 2018).

3.5 Other criteria

There are two additional criteria to be met:

Affordability, in the sense that the considerable design costs and implementation costs are amortized over a large number of participating countries.

Manageability, in the sense that the probability of experiencing catastrophic errors in implementation is within acceptable limits.

The PIAAC approach would tax the financial, operational and technical capacity of a significant minority of developing countries.

Given the relative importance of literacy skill and individual and collective outcomes and the cost of assessment programmes, current assessments, while expensive, offer good value for money.

Manageability is quite different. The PIAAC approach is more technically and operationally complex than can be managed by the bottom end of the OECD countries so the pretence that countries with lower levels of technical and operational infrastructure will be able to cope is absurd.

3.6 Topic for discussion

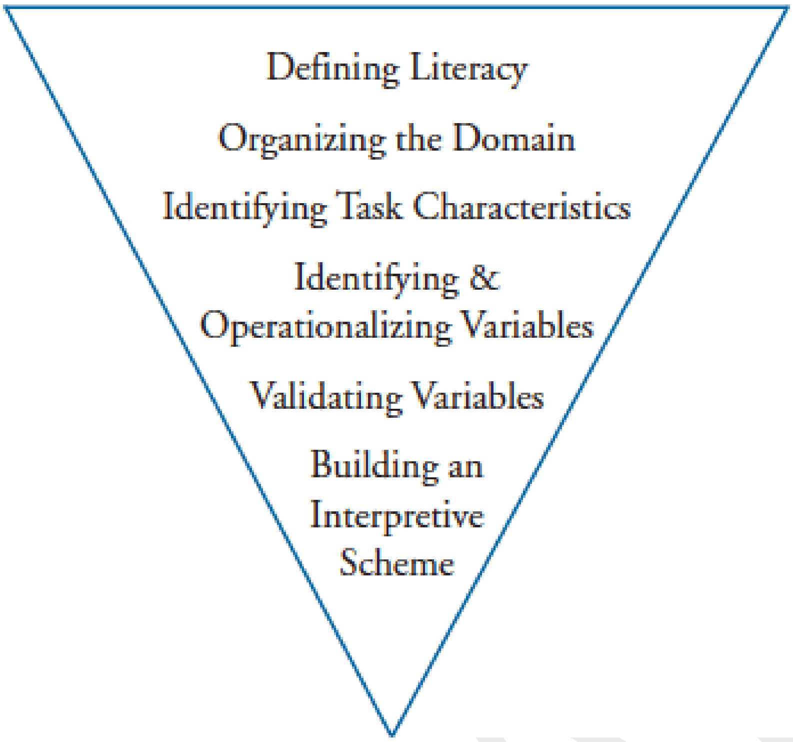
- Do the experts agree that the proposed assessment programme needs to satisfy these criteria?
- Are there criteria that could be dropped or relaxed?

4 A brief overview of how literacy is defined and measured internationally

The definition of literacy has evolved considerably over the past 40 years in response to theoretical advances that allow one to predict the relative difficulty of reading tasks to high levels of precision. At the extreme, the definition has shifted from being able to sign one’s name to being able to solve a broad range of tasks using the information gleaned from what one has read.

Current measures of literacy have been developed following a strict set of guidelines that ensure that the resulting measures are valid, reliable, comparable and interpretable. The development schema for a literacy framework is illustrated below.

Figure 1. A framework for assessing literacy


Defining Literacy
Organizing the Domain
Identifying Task Characteristics
Identifying & Operationalizing Variables
Validating Variables
Building an Interpretive Scheme

Once the variables that underlie the relative difficulty of tasks in the domain have been identified, researchers have developed pools of items that systematically sample the variables.

When administered in a test, these items afford a way to assess an individual's ability to cope with tasks that span the entire range of task difficulty.

As a last step, the probability that an individual can answer a particular item correct is estimated. Individuals are placed at the proficiency level as a result of meeting some threshold, either an additive score or a probability threshold for getting items at that proficiency level correct.

It is important to note that a defining feature of successful assessment programmes is the coherence between the conceptual aspects of the design, how the data is collected and processed, and how results are reported.

The key insight afforded by these theoretical advances is that attributes of the text being read have very little impact on the relative difficulty of tasks, explaining only 15 per cent of observed variance in task difficulty in the range from 180 to 500 points on the international proficiency scale. Rather, the cognitive demands of the reading task explain the overwhelming majority of differences in task difficulty associated with the emergence of fluid and automatic reading.

4.1 Definitions of literacy

The studies reviewed are all, implicitly or explicitly, based on a complex conception of literacy.

The first fundamental insight is that literacy involves both **learning to read** – the acquisition of the component skills that underlie the emergence of fluid and automatic reading – and **reading to learn**, the act of applying information gleaned through the application of fluid and automatic reading.

Figure 2, drawn from *Learning Literacy in Canada: Evidence from the International Survey of Reading Skills* (Statistics Canada and HRSDC, Ottawa, 2007) provides information on how IALS/ALL/PIAAC conceive proficiency transitions in literacy.

Figure 2

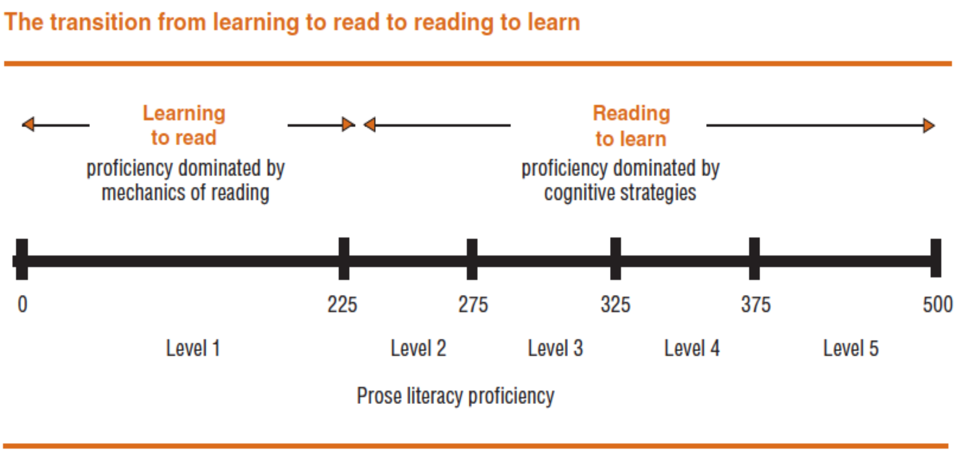

The transition from learning to read to reading to learn
Learning to read
proficiency dominated by mechanics of reading
Reading to learn
proficiency dominated by cognitive strategies
0
225
275
325
375
500
Level 1
Level 2
Level 3
Level 4
Level 5
Prose literacy proficiency

Figure 2 suggests that there are three groups of learners:

Below approximately 250 on the proficiency scale, adults are still in the process of learning to read, i.e. of acquiring the level of mastery of the reading components that allows them to become fluid and automatic readers who can devote most of their cognitive space to applying what they have read/building meaning.

Between 250 to 275, individuals continue to improve their application of the component skills but their position on the overall proficiency scale is determined by their probability of getting Level 3 tasks correct. Specifically, they do not have the transferable skill to get the 80 per cent or more of Level 3 items correct required to be classified at Level 3.

Above 275 points on the literacy scale, adults continue to improve their mastery of the reading components but are proficient enough that their performance is largely a function of their mastery of the cognitive strategies associated integrating and generating information.

The key difference in the studies reviewed in this report – IALS, ALL, PIAAC, IVQ, Skills for Life, the German LEO study, the Kenyan Adult Literacy Survey and the Bangladesh Literacy Assessment Survey – is not, therefore, their conception of literacy but the part of the scale on which they focus their attention, the methods they apply to derive scores and how they chose to define proficiency levels.

4.2 IALS and ALL

The IALS/ALL framework begins by **defining literacy** as ‘understanding, evaluating, using and engaging with written texts to participate in society, to achieve one’s goals and to develop one’s knowledge and potential’.[2]

This definition implies far more than just reading the words of the text. It includes an emphasis on how the information gathered from this encounter with written materials is used and influences one’s thinking.

IALS and ALL chose to focus all of their measurement on reading to learn so the test items are heavily focused on levels 2, 3, 4 and 5 on what has become the PIAAC literacy scale.

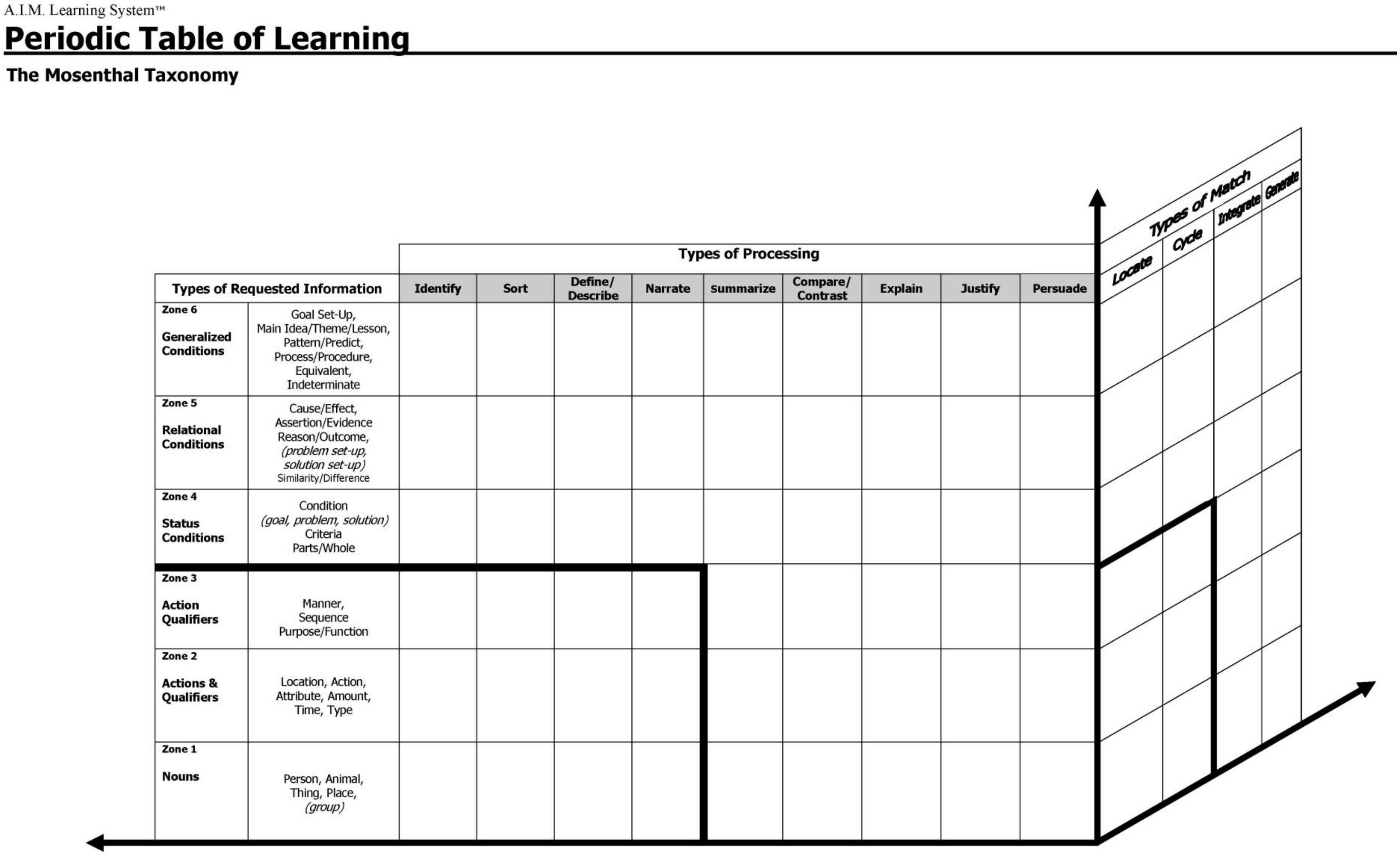
To answer test items correctly adults needed to be able to:

- read and understand the question being asked;
- apply the appropriate cognitive strategy to find the correct answer;
- understand and provide the appropriate type of response.

[2] Ibid (p. 20)

Figure 3 illustrates the IALS and ALL framework used to estimate the relative difficulty of reading tasks.

Figure 3. The variables that predict the relative difficulty of reading tasks



A.I.M. Learning System™

Periodic Table of Learning

The Mosenthal Taxonomy

		Types of Processing									Types of Match			
Types of Requested Information		Identify	Sort	Define/ Describe	Narrate	Summarize	Compare/ Contrast	Explain	Justify	Persuade	Locate	Cycle	Integrate	Generate
Zone 6 Generalized Conditions	Goal Set-Up, Main Idea/Theme/Lesson, Pattern/Predict, Process/Procedure, Equivalent, Indeterminate													
Zone 5 Relational Conditions	Cause/Effect, Assertion/Evidence Reason/Outcome, *(problem set-up, solution set-up)* Similarity/Difference													
Zone 4 Status Conditions	Condition *(goal, problem, solution)* Criteria Parts/Whole													
Zone 3 Action Qualifiers	Manner, Sequence Purpose/Function													
Zone 2 Actions & Qualifiers	Location, Action, Attribute, Amount, Time, Type													
Zone 1 Nouns	Person, Animal, Thing, Place, *(group)*													

4.2.1 Plausibility of distractors

The IALS/ALL frameworks also include a construct known as 'plausibility of distractors', which captures the presence or absence of competing information that distracts weak readers from the correct answer.

Systematic sampling of this design matrix allowed the IALS, ALL, PIAAC, LAMP and STEP proficiency scores to be interpreted as reliable indicators of general proficiency, again in the range assessed by the overall item pool.

Collectively, these variables explain 85 per cent of the total variance in task difficulty in the range of 180 to 500 on the international scale, a high enough percentage that leaves very little room for other variables to have an impact.

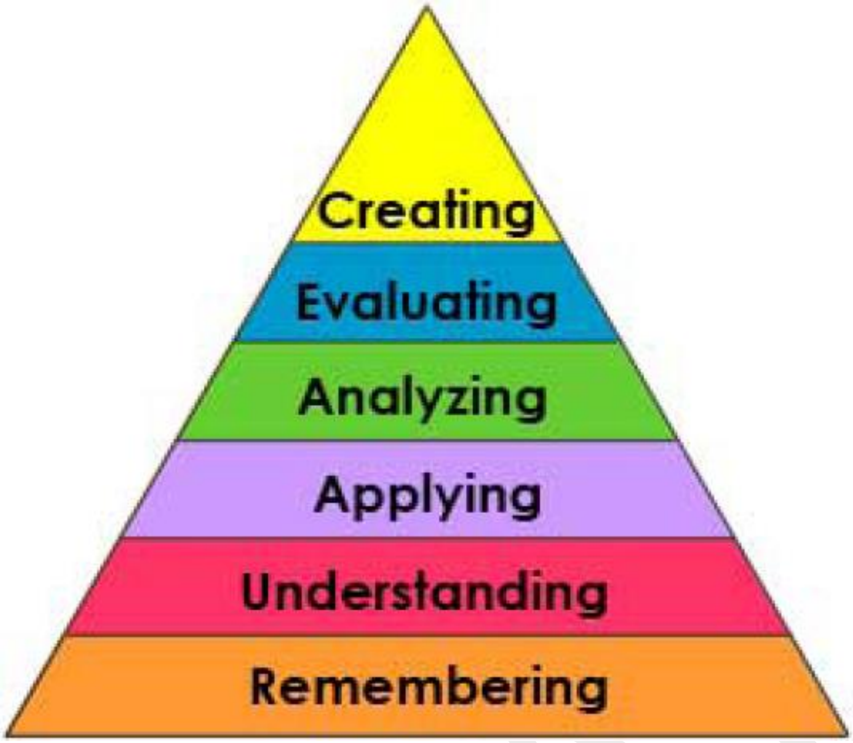
A non-trivial amount of time and money would have to be invested to improve upon the performance of these models.

Types of match explain fully 85 per cent of predicted task difficulty in the range of 180 to 500 on the international scale and engage universal cognitive strategies that operate in the prefrontal cortex. The other two dimensions explain the remaining 15 per cent of explained task difficulty. Importantly for current purposes, these strategies are linguistically and culturally independent. A fourth dimension, plausibility of distractors, provides additional predictive power associated with the fact that proficient readers are able to identify and ignore incorrect information that is in close proximity to correct answers.

The heavy black bar denotes the boundary between literacy tasks at levels 2 and 3. Level 2 tasks involve the routine application of procedural knowledge and facts. Importantly, this knowledge can be gained by means other than reading, so complicates the act of measurement. More directly, weak readers may get items correct because of what they know rather than from what they learn from reading a text.

As illustrated below, the boundary between levels 2 and 3 also corresponds to an important threshold identified in the curricular frameworks that underpin instruction in the world's education systems, including Bloom's revised taxonomy, i.e. the boundary between applying and analysing.

Figure 4. The levels in Bloom's revised taxonomy



This alignment is crucially important because the literacy measures need to speak to educators in a way that is easy for them to understand. In the range covered by the PIAAC literacy scale, the framework offers unambiguous insight into what amount and types of instruction are needed to move adults from Level 2 to Level 3, from Level 3 to Level 4, and from Level 4 to Level 5.

Almost all studies that will be discussed in the next chapter, i.e. PIAAC, LAMP, STEP, the UK Skills for Life study, the French IVQ study and the German LEO study, have either implicitly or explicitly accepted the IALS/ALL definition of literacy and the underlying predictors of task difficulty. The exception is the Bangladesh assessment that did not attempt to measure the upper 'reading to learn' regions of the literacy scale.

5 Definition of proficiency levels

Several studies were reviewed, including PIAAC, LAMP, STEP, the UK Skills for Life, the French IVQ, the German LEO study, the Kenyan National Adult Literacy Study and the Bangladesh Adult Literacy study. These included measures that were designed to improve measurement in the lower 'learning to read' regions of the literacy scale where IALS and ALL offered little information.

Each of these assessments chose, however, a different approach to getting more information about adults in the lower regions of the literacy scale.

Each assessment also chose to define and report proficiency levels in different ways that carry important implications for the comparability of results across languages and countries.

5.1 IALS, ALL and LAMP

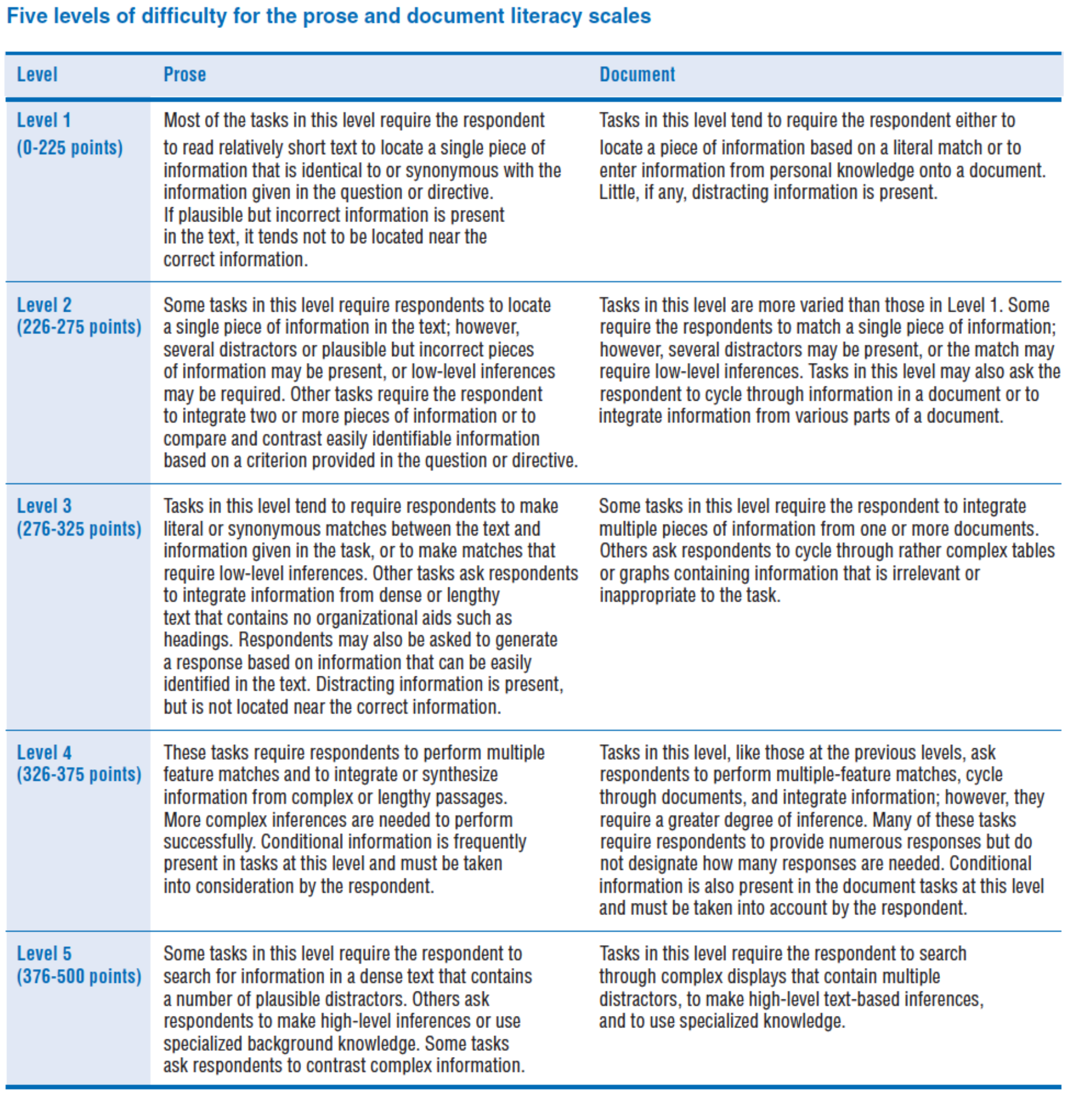
IALS, ALL and LAMP all estimate respondent's scores on a 500-point scale. Respondents are then assigned to one of five proficiency levels based on score thresholds and the imposition of a mastery standard that requires respondents to have an 80 per cent or better probability of getting items at the assigned level correct.

The IALS/ALL/LAMP and STEP proficiency levels are in the first instance defined to represent points along the literacy continuum where shifts occur in the essential nature of the skills required to get a task correct. In IALS and ALL, the predicted item difficulty is compared to the empirically observed item difficulty and is shown to be in very close agreement. This implies that the levels that are defined on the scale can be interpreted as a reliable indicator of proficiency.

5.2 PIAAC

Table 2 documents the PIAAC levels and their descriptions of what adults at each level can do.

Table 2. PIAAC levels and descriptions



Five levels of difficulty for the prose and document literacy scales

Level	Prose	Document
Level 1 (0-225 points)	Most of the tasks in this level require the respondent to read relatively short text to locate a single piece of information that is identical to or synonymous with the information given in the question or directive. If plausible but incorrect information is present in the text, it tends not to be located near the correct information.	Tasks in this level tend to require the respondent either to locate a piece of information based on a literal match or to enter information from personal knowledge onto a document. Little, if any, distracting information is present.
Level 2 (226-275 points)	Some tasks in this level require respondents to locate a single piece of information in the text; however, several distractors or plausible but incorrect pieces of information may be present, or low-level inferences may be required. Other tasks require the respondent to integrate two or more pieces of information or to compare and contrast easily identifiable information based on a criterion provided in the question or directive.	Tasks in this level are more varied than those in Level 1. Some require the respondents to match a single piece of information; however, several distractors may be present, or the match may require low-level inferences. Tasks in this level may also ask the respondent to cycle through information in a document or to integrate information from various parts of a document.
Level 3 (276-325 points)	Tasks in this level tend to require respondents to make literal or synonymous matches between the text and information given in the task, or to make matches that require low-level inferences. Other tasks ask respondents to integrate information from dense or lengthy text that contains no organizational aids such as headings. Respondents may also be asked to generate a response based on information that can be easily identified in the text. Distracting information is present, but is not located near the correct information.	Some tasks in this level require the respondent to integrate multiple pieces of information from one or more documents. Others ask respondents to cycle through rather complex tables or graphs containing information that is irrelevant or inappropriate to the task.
Level 4 (326-375 points)	These tasks require respondents to perform multiple feature matches and to integrate or synthesize information from complex or lengthy passages. More complex inferences are needed to perform successfully. Conditional information is frequently present in tasks at this level and must be taken into consideration by the respondent.	Tasks in this level, like those at the previous levels, ask respondents to perform multiple-feature matches, cycle through documents, and integrate information; however, they require a greater degree of inference. Many of these tasks require respondents to provide numerous responses but do not designate how many responses are needed. Conditional information is also present in the document tasks at this level and must be taken into account by the respondent.
Level 5 (376-500 points)	Some tasks in this level require the respondent to search for information in a dense text that contains a number of plausible distractors. Others ask respondents to make high-level inferences or use specialized background knowledge. Some tasks ask respondents to contrast complex information.	Tasks in this level require the respondent to search through complex displays that contain multiple distractors, to make high-level text-based inferences, and to use specialized knowledge.

PIAAC chose to adopt the IALS, ALL and LAMP proficiency definitions but introduced two changes.

First, Level 1 was divided into two Levels:

- Level 1
- Below Level 1

The Below Level 1 level was defined to separate out those respondents for whom there was no measurement. PIAAC also chose to adjust the descriptions of what individuals could do at each level. Implicitly, this change involves a reduction of the IALS/ALL/LAMP mastery standard from 80 per cent to 62.5 per cent. This has the impact of moving respondents who were close to a level score threshold down into the next lower level.

5.3 Skills for Life

The Skills for Life assessment chose to add levels to PIAAC level 1 that are defined by score thresholds on the overall literacy scale.

As noted earlier, this classification is based on a much more reliable estimate of low-skilled adults' score but offers little insight into their learning needs. More directly, one knows how far one is away from the next PIAAC level but not what it would take to move up the scale.

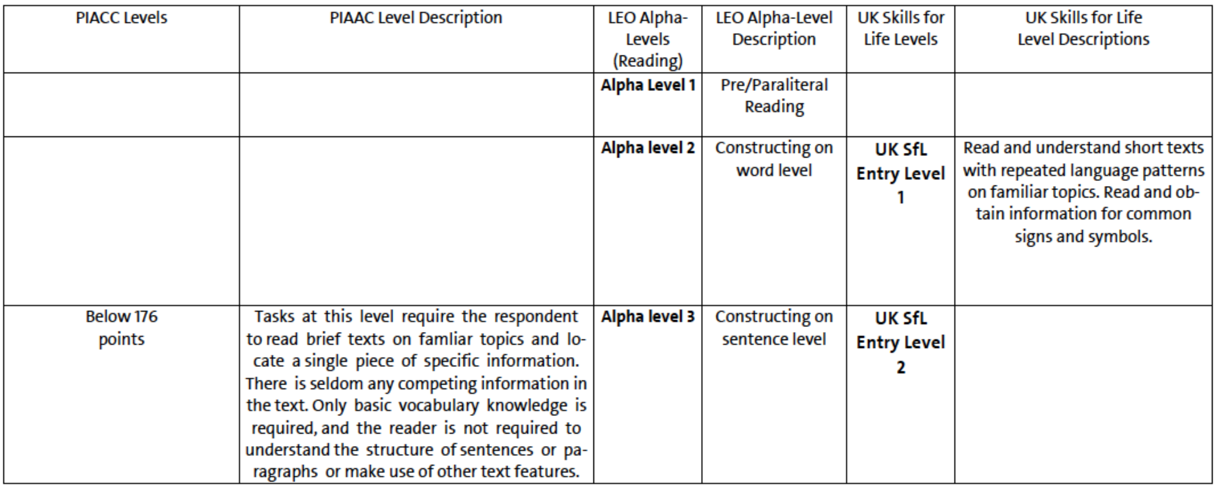
5.4 LEO

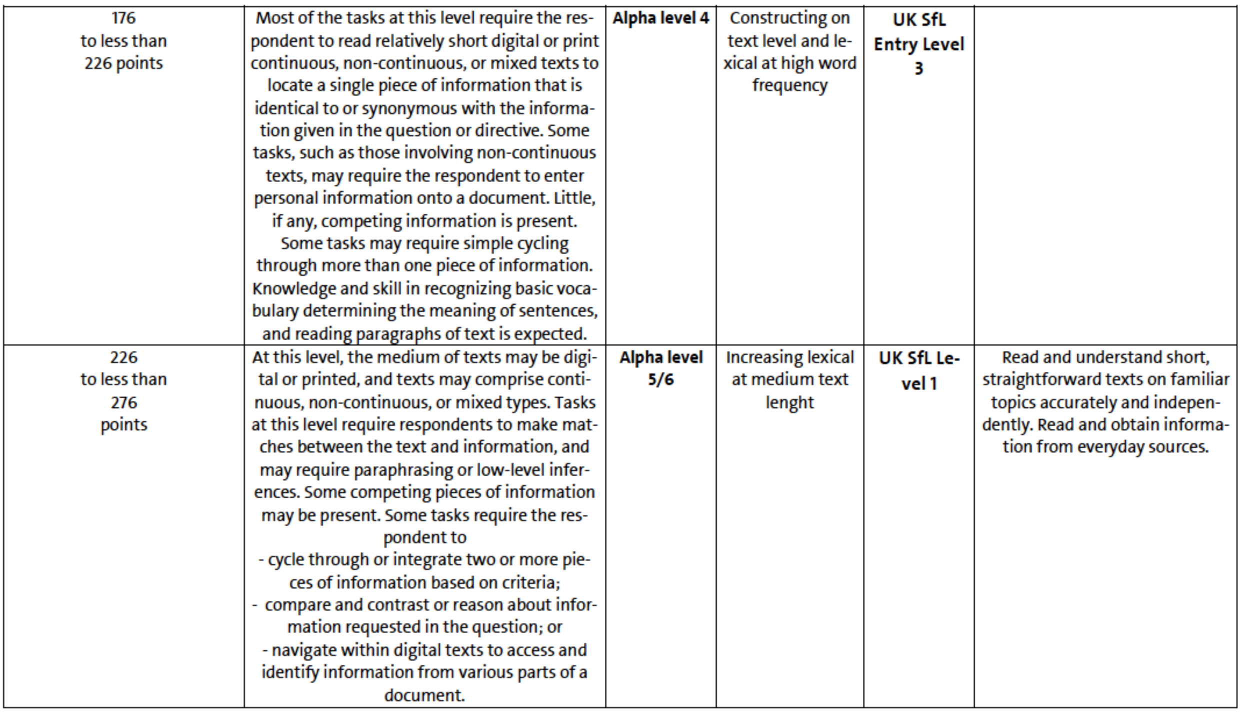
As illustrated above, LEO chose to apply a 62 per cent mastery standard that divides adults at Level 1 into equally sized groups, their so called Alpha levels. Implicitly, LEO also measures writing in the same way that IALS/ALL/PIAAC/LAMP and STEP do in the sense that respondents are required to enter their answers. In the case of LEO and PIAAC, this entry is computer based; in IALS, ALL, PIAAC, STEP and LAMP entry uses paper and pencil. As noted for the Skills for Life assessment, this classification is based upon a much more reliable estimate of low-skilled adults' score but offers little insight into their learning needs. More directly, one knows how far one is away from the next PIAAC level but not what it would take to move up the scale.

Table 3[3] provides a useful alignment of the LEO and Skills for Life lower levels with the overall PIAAC scale.

[3] Anke Grotluschen, 2018

Table 3. Alignment of LEO and Skills for Life lower levels with the overall PIAAC scale





PIACC Levels	PIAAC Level Description	LEO Alpha-Levels (Reading)	LEO Alpha-Level Description	UK Skills for Life Levels	UK Skills for Life Level Descriptions
		Alpha Level 1	Pre/Paraliteral Reading		
		Alpha level 2	Constructing on word level	**UK SfL Entry Level 1**	Read and understand short texts with repeated language patterns on familiar topics. Read and obtain information for common signs and symbols.
Below 176 points	Tasks at this level require the respondent to read brief texts on famliar topics and locate a single piece of specific information. There is seldom any competing information in the text. Only basic vocabulary knowledge is required, and the reader is not required to understand the structure of sentences or paragraphs or make use of other text features.	**Alpha level 3**	Constructing on sentence level	**UK SfL Entry Level 2**	
176 to less than 226 points	Most of the tasks at this level require the respondent to read relatively short digital or print continuous, non-continuous, or mixed texts to locate a single piece of information that is identical to or synonymous with the information given in the question or directive. Some tasks, such as those involving non-continuous texts, may require the respondent to enter personal information onto a document. Little, if any, competing information is present. Some tasks may require simple cycling through more than one piece of information. Knowledge and skill in recognizing basic vocabulary determining the meaning of sentences, and reading paragraphs of text is expected.	**Alpha level 4**	Constructing on text level and lexical at high word frequency	**UK SfL Entry Level 3**	
226 to less than 276 points	At this level, the medium of texts may be digital or printed, and texts may comprise continuous, non-continuous, or mixed types. Tasks at this level require respondents to make matches between the text and information, and may require paraphrasing or low-level inferences. Some competing pieces of information may be present. Some tasks require the respondent to - cycle through or integrate two or more pieces of information based on criteria; - compare and contrast or reason about information requested in the question; or - navigate within digital texts to access and identify information from various parts of a document.	**Alpha level 5/6**	Increasing lexical at medium text lenght	**UK SfL Level 1**	Read and understand short, straightforward texts on familiar topics accurately and independently. Read and obtain information from everyday sources.

5.5 IVQ

The IVQ was designed to assess skill in three sub-domains – reading, comprehension and writing. Respondents were assigned a percentage correct in each sub-domain. The writing component of IVQ is slightly more demanding than the entry requirements used in ALS, ALL, PIAAC, LAMP and STEP so there is a possibility that the results may not be directly comparable.

IVQ then constructed a composite classification across the three sub-domains.

- Individuals were labelled as having ‘no difficulties’ if they scored 80 per cent on all three of the sub-domains.
- Individuals were labelled as having ‘difficulties’ if they scored between 60 per cent and 80 per cent on one of the three of the subdomains, but no score lower than 60 per cent on any of the sub-domains.
- Individuals were labelled as having ‘considerable difficulties’ if they scored between 40 per cent and 60 per cent on all three of the sub-domains, but with no score lower than 40 per cent on any of the sub-domains.
- Individuals were labelled as having ‘serious difficulties’ as a result of not being classified in one of the foregoing groups i.e. they have a success rate below 40 per cent in at least one of the sub-domains.

5.6 KNALS

The Kenyan National Adult Literacy Survey assessed a range of literacy skills in two national languages (English and Kiswahili) and 18 regional languages. Seventy per cent of respondents took the test in either English of Kiswahili.

KNALS assessed three skills – reading, writing and numeracy – in the population 15 years of age and up.

Proficiency was measured with a mix of narrative, expository and document texts – the same mix as in PIAAC.

The study attempted to assess a broad range of literacy skills but most of the assessment time was devoted to items that would be classified at levels 1, 2 and 3 on the PIAAC reading scale. The 18 literacy items administered were classified into the levels shown in *Figure 5*.

Figure 5: KNALS competency skills levels

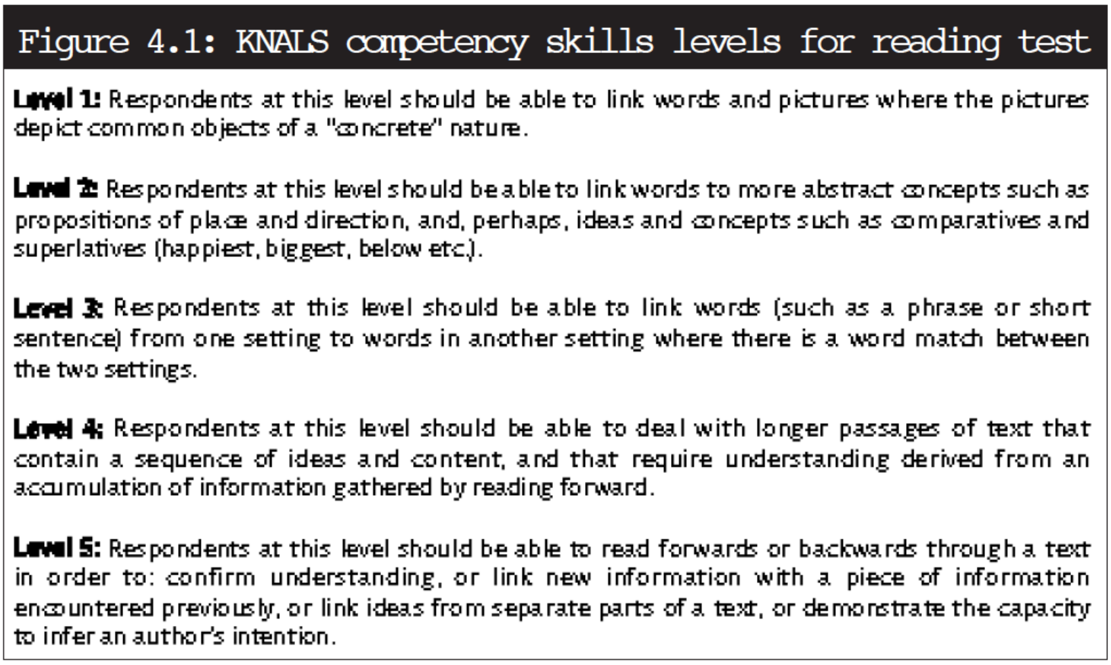


Figure 4.1: KNALS competency skills levels for reading test

Level 1: Respondents at this level should be able to link words and pictures where the pictures depict common objects of a "concrete" nature.

Level 2: Respondents at this level should be able to link words to more abstract concepts such as propositions of place and direction, and, perhaps, ideas and concepts such as comparatives and superlatives (happiest, biggest, below etc.).

Level 3: Respondents at this level should be able to link words (such as a phrase or short sentence) from one setting to words in another setting where there is a word match between the two settings.

Level 4: Respondents at this level should be able to deal with longer passages of text that contain a sequence of ideas and content, and that require understanding derived from an accumulation of information gathered by reading forward.

Level 5: Respondents at this level should be able to read forwards or backwards through a text in order to: confirm understanding, or link new information with a piece of information encountered previously, or link ideas from separate parts of a text, or demonstrate the capacity to infer an author's intention.

Items were scaled using a Rasch model, a one-parameter variant of the three-parameter item response model used to scale PIAAC results. This model assumes that items discriminate perfectly and pushes any error associated with this assumption not being true into the error terms. Proficiency levels were defined by defining Rasch model score ranges. The categorization included defining a category for people without any literacy skill.

Table 4. Literacy competency scores

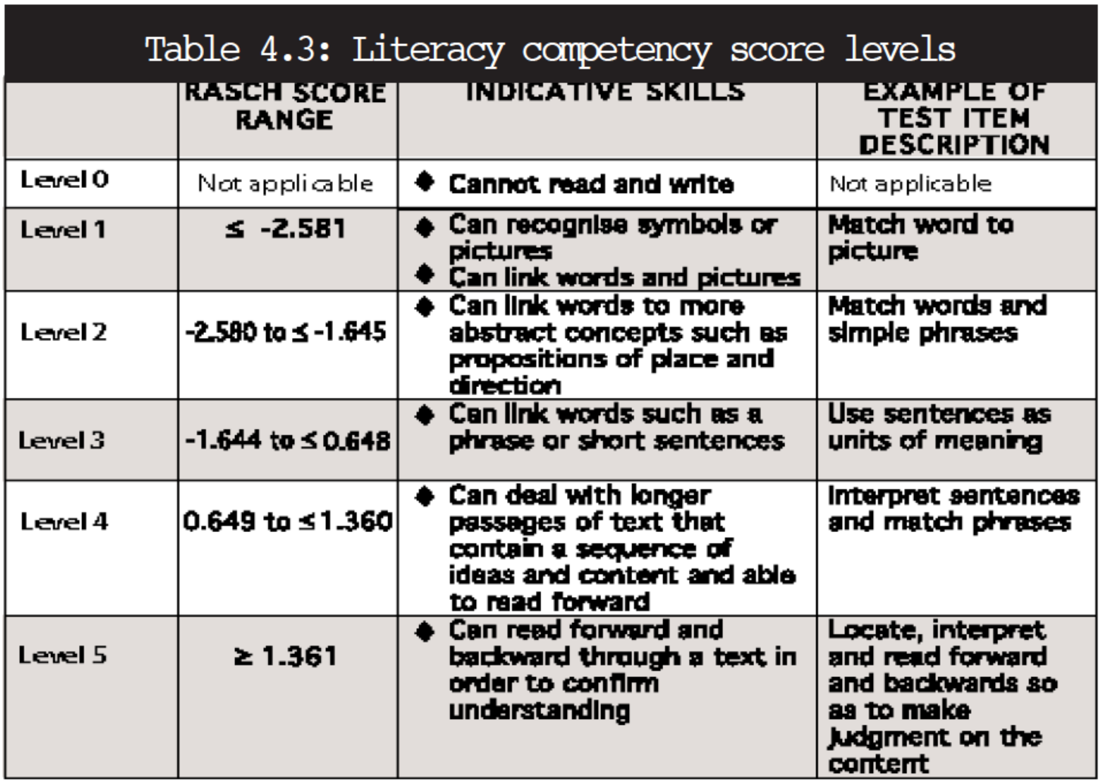


Table 4.3: Literacy competency score levels

	RASCH SCORE RANGE	INDICATIVE SKILLS	EXAMPLE OF TEST ITEM DESCRIPTION
Level 0	Not applicable	◆ Cannot read and write	Not applicable
Level 1	≤ -2.581	◆ Can recognise symbols or pictures ◆ Can link words and pictures	Match word to picture
Level 2	-2.580 to ≤ -1.645	◆ Can link words to more abstract concepts such as propositions of place and direction	Match words and simple phrases
Level 3	-1.644 to ≤ 0.648	◆ Can link words such as a phrase or short sentences	Use sentences as units of meaning
Level 4	0.649 to ≤ 1.360	◆ Can deal with longer passages of text that contain a sequence of ideas and content and able to read forward	Interpret sentences and match phrases
Level 5	≥ 1.361	◆ Can read forward and backward through a text in order to confirm understanding	Locate, interpret and read forward and backwards so as to make judgment on the content

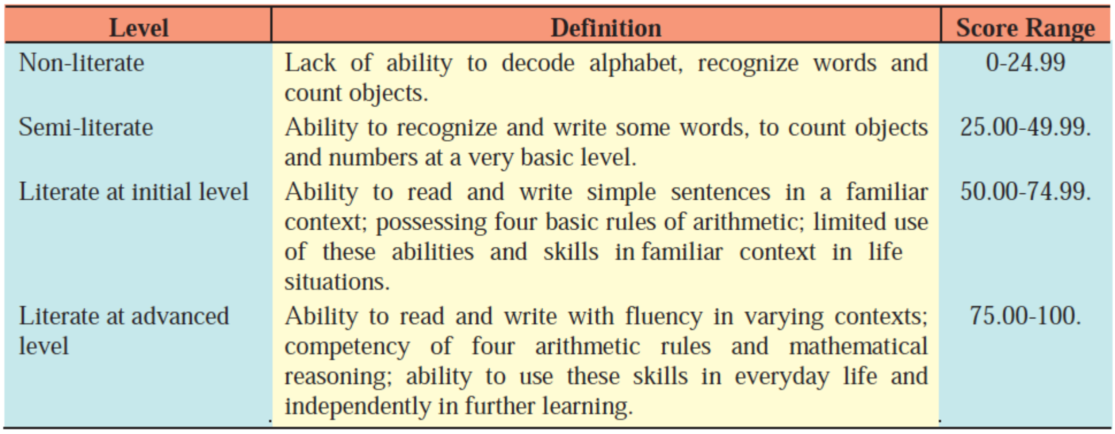
A comparison to the PIAAC levels suggest that the items fall exclusively into PIAAC levels Below Level 1, Level 1 and Level 2. Level 5 items fall into PIAAC Level 3.

The study went on to define Rasch Level 3 as the minimum mastery level and Rasch Level 4 as meeting the desired level of mastery. This classification stands as an example of a national standard being applied to an objective measure of skill. The defined levels of mastery are lower than those applied in the IALS and ALL studies. PIAAC avoided setting a minimum mastery level out of concern that such classifications are subjective and depend on national priorities, goals and expectations.

5.7 Bangladesh Literacy Assessment Survey

The BLAS study defines four proficiency levels based on score thresholds on the scale that is defined by the number correct out of 100, as illustrated in *Table 5*:

Table 5. BLAS proficient levels based on score thresholds



Level	Definition	Score Range
Non-literate	Lack of ability to decode alphabet, recognize words and count objects.	0-24.99
Semi-literate	Ability to recognize and write some words, to count objects and numbers at a very basic level.	25.00-49.99.
Literate at initial level	Ability to read and write simple sentences in a familiar context; possessing four basic rules of arithmetic; limited use of these abilities and skills in familiar context in life situations.	50.00-74.99.
Literate at advanced level	Ability to read and write with fluency in varying contexts; competency of four arithmetic rules and mathematical reasoning; ability to use these skills in everyday life and independently in further learning.	75.00-100.

The levels themselves and, by extension, the underlying tasks, are conceptually very similar to the reading component measures derived for the International Survey of Reading Skills (ISRS) study upon which the PIAAC, LAMP and STEP component measures are based.

Proficiency in the BLAS study will, however, be somewhat overestimated because of its reliance on a small number of items that are assumed to be equally familiar to all respondents.

5.8 Topics for discussion

- What domains and sub-domains should be included in a literacy assessment framework?
- Should the PIAAC feature of combing prose literacy and document literacy be adopted?

6 A comparison of approaches to measurement in the lower regions of the proficiency scale

This chapter compares the approach a range of studies have taken to improving the amount of measurement devoted to the lower regions of the literacy scale.

6.1 PIAAC

The PIAAC assessment chose to administer a variant of the reading component measures administered in the ISRS survey described in *Annex B*.

The ISRS study was designed to assess the component reading skills thought to underlie the emergence of fluid and automatic reading that is needed to master Level 3 and above literacy and numeracy tasks i.e. letter and number recognition, receptive vocabulary, decoding fluency and accuracy and passage fluency.

The availability of these measures provided deep insight into the learning needs of adults at levels 1 or 2, a part of the IALS/ALL proficiency distribution about which little was known. When analysed with complex methods these measures identify groups of learners in the lower range of the scale who share common patterns of strengths and weaknesses that imply a need for a distinct instructional response.

6.2 LAMP and STEP

The LAMP and STEP programmes chose to include a variant of the PIAAC reading component assessments and to develop additional items with very low difficulties.

Analysis of the data concerning groups of learners defined by shared patterns of strengths and weaknesses in the reading components reveals significant differences among languages that appear to be a function of the relationship between the orthographic structure of the written word and the spoken word. In languages such as Spanish, where one observes a one-to-one match between the spoken and written word, the process of decoding is simpler than in English where many phonemes remain unspoken. This insight carries direct implications for the proposed global framework.

First, reading component measures would need to be developed for each language.

Second, the available data suggest that one cannot undertaken first order comparisons of results across languages. For example, one would not be able to compare the proportion of adults on the number of letters recognized because the number of symbols differs across languages. One could, however, safely compare the proportion of adults able to recognize 80 per cent or more of the symbol set, a threshold that is needed to support fluid and automatic reading.

Third, notwithstanding the difficulties involved in direct comparison, adults unable to identify a single letter of the alphabet/symbol set can safely be classified as having no literacy skills.

LAMP and STEP also developed and administered additional literacy items with very low difficulties on the overall literacy proficiency scale. The inclusion of these items did increase the amount of measurement in the lower regions of the literacy scale but added little to the instructional prescription.

Importantly for current purposes, it proved to be much more difficult than expected to develop test items in the easiest range of the scale. The psychometric performance of the majority of such items was poor because the proportion of respondents getting the item right unexpectedly, given their proficiency level, rose to unacceptably high levels. Analysis suggested that enough low-level readers were getting the item correct because they were familiar with the content rather than through the application of their reading skills per se.

It is likely to be even more difficult to develop additional very simple test items that display the level of stable psychometric performance needed to support comparability in increasingly heterogeneous populations.

This finding suggests that it is likely that the 'lower rungs' approaches adopted in the Skills for Life, LEO and IVQ studies, Kenyan and Bangladesh assessments, would allow adults to be placed more precisely on the literacy scale but that their approach to defining additional levels would not yield statistically 'clean'' groups of learners.

Unfortunately for policy-makers, the methods used to analyse the reading components measures in the ISRS have not been applied to the PIAAC, STEP, LAMP or LEO data. As a result, data-users have not had access to a reliable way of classifying groups of learners that share common learning needs. Without this information, policy-makers do not have a way to do the basic cost-benefit rate of return analysis needed to argue for and to allocate funds.

6.3 Skills for Life

The UK Skills for Life Surveys (SfL), conducted in 2003 and 2011, borrowed heavily from the IALS design and adopted the same overall definition of literacy as PIAAC. Additionally, the study designers developed 25 very easy items in a bid to increase the amount of measurement in the lower regions of the overall literacy scale. The underlying goal was to provide a more fulsome description of what individuals in the 'lower rungs' of the scale were, and were not, able to do.

Essentially, these items require test takers to locate information in simple texts and, thus, allow one to come up with a more precise estimate of how far away someone is from the important boundary between Level 2 and 3.

These items allow respondents to be situated much more precisely on the lower regions of the proficiency scale. The SfL items did little, however, to provide additional insight into what sort of instructional response would be needed to move these learners up the scale.

6.4 The French IVQ

The French IVQ, administered by INSEE in cooperation with ANCLI, employed similar assessment methods to improve measurement in the lower regions of the scale. These measures were not designed to measure the components of reading that explain the emergence of fluid and automatic reading that characterizes performance in the upper regions of the proficiency scale. The IVQ also included measures designed to capture information on the coping mechanisms employed by poor readers.

Importantly, the IVQ was not designed to yield estimates that could be compared across countries but did assume that the measures were reliable within French populations.

The IVQ used classical test theory to summarize test scores. Classical test theory assumes that each item discriminates perfectly and does not offer an easy way to confirm empirically that test items are performing in the same way in different sub-populations, including those based on gender. Differential item functioning among men and women will not be evident.

The IVQ used score thresholds to define proficiency levels. Since classical test theory treats each item as equally informative adults are classified as being functionally illiterate based upon quite different patterns of incorrect items.

The IVQ also measured test-takers’ ability to write. In our view, writing emerges naturally as a function of learning to read for most people. Writing is also somewhat problematic because adults often choose to communicate simple information verbally.

6.5 The German LEO

The German LEO study is best thought of as a hybrid of the ‘lower rungs’ and ‘reading components’ approaches. The low-level reading measures administered in LEO tapped most of the skills assessed by the ISRS, PIAAC, LAMP and STEP reading components but the analysis undertaken did not attempt to identify patterns of strengths and weaknesses across the components. As illustrated in *Figure 6*, the LEO measures allow one to place people on the overall literacy proficiency scale much more accurately.

Figure 6. Alpha levels in the LEO study

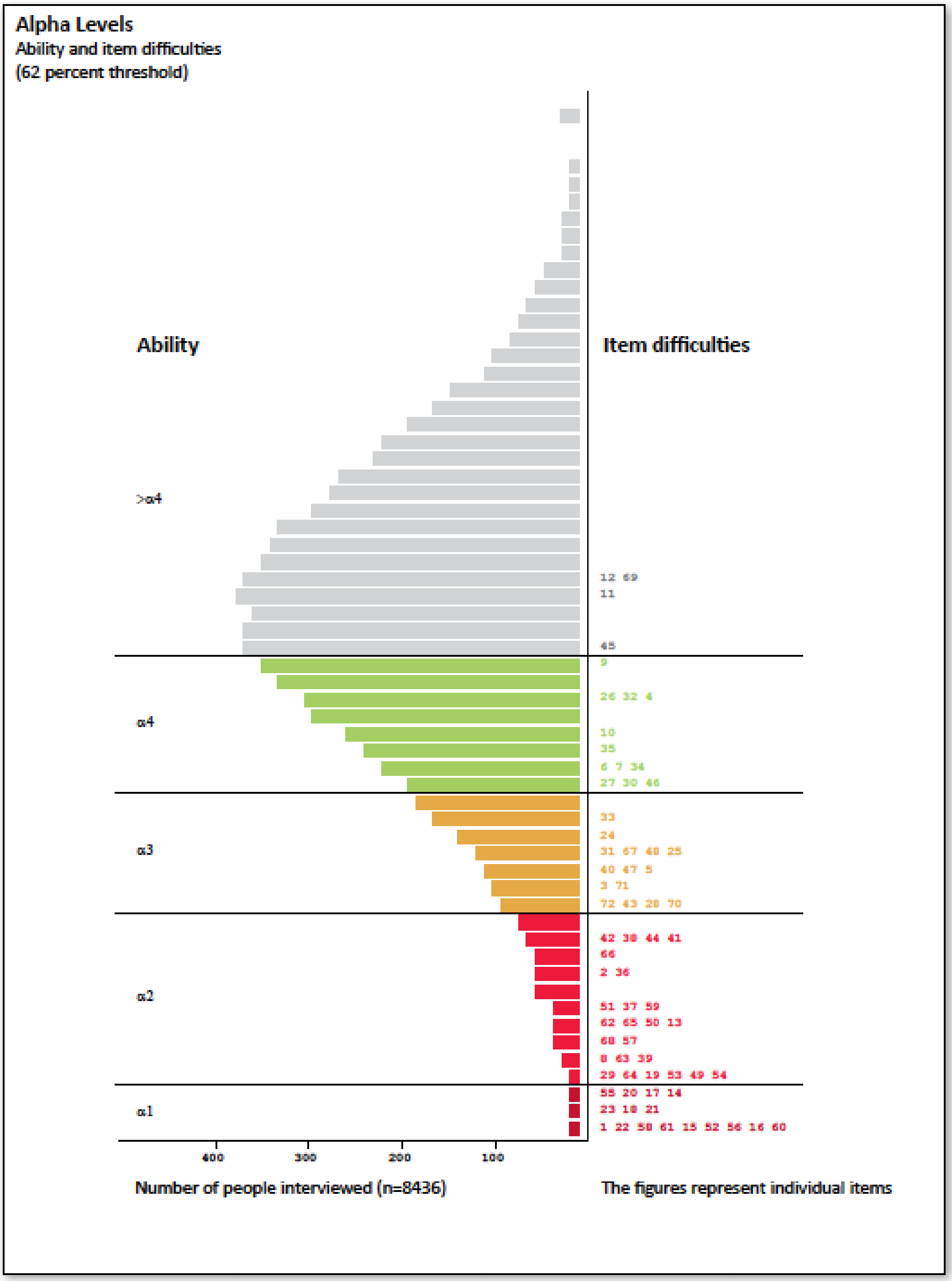


Figure 1: level of item difficulties and ability

Text length and word frequency predict the relative difficulty of LEO items quite well but this predictive value is a function of the relationship of these variables to the underlying processes assessed in the ISRS/PIAAC/LAMP/STEP component measures. Moreover, they do not represent things that one would teach to impart higher skill levels.

6.6 The Kenyan National Adult Literacy Survey (KNALS)

The Kenyan Government identified an urgent need for data on literacy and numeracy skill distributions. The national assessment team reviewed the LAMP assessment method and items and determined that the low-level items did not reflect Kenyan culture and context.

As a result, the Kenyan National Adult Literacy Survey adopted the conceptual framework that underpins LAMP but chose to develop assessment items that reflect Kenyan culture and context in a large number of indigenous languages. This approach rests upon the unproven assumption that Kenya is culturally homogeneous across a broad range of languages, population density and significant economic disparity.

This assumption could be tested by applying the statistical methods that were applied in the IALS/ALL/PIAAC/STEP and LAMP assessments. These methods identify items, individuals and population subgroups which are not performing in the predicted, stable way. In the latter two cases, comparisons are made that adjust for known differences in the full array of background characteristics.

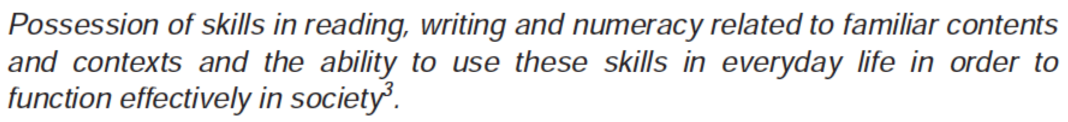
6.7 The Bangladesh Literacy Assessment Survey

The Bangladesh Adult Literacy Survey (BLAS) assessed the skills of the adult population 11 years of age and above.

Bangladesh's Non-Formal Education Policy adopted a definition of literacy that was very close to the UNESCO definition:

Literacy is defined as the ability to identify, understand, interpret, create, communicate and compute using printed and written materials associated with diverse contexts. Literacy involves a continuum of learning in enabling individuals to achieve their goals, develop their knowledge and potential and participate fully in community and society. (UNESCO. 2005. Aspects of Literacy Assessment: Topics and issues. [http://unesdoc.unesco.org/images/0014/001401/140125eo.pdf.)](http://unesdoc.unesco.org/images/0014/001401/140125eo.pdf)

The BLAS study chose, however, to adopt another definition of literacy:



Possession of skills in reading, writing and numeracy related to familiar contents and contexts and the ability to use these skills in everyday life in order to function effectively in society[3].

This definition introduced the requirement that test items needed to be familiar.

This design feature precludes comparison of results to the other studies reviewed in this report. The assessment used a very small number (5) literacy and numeracy questions to evaluate an individual's proficiency level. The fundamental problems with assessments that use small numbers of test items to determine skill levels is that they offer little discrimination along the continuum of skill and offer too little information to support any statistically defensible generalization of proficiency level. These studies also are much more susceptible to cheating. As a result, these assessments are of little use in defining thoughtful policy responses.

Table 6 provides a comparative summary overview of the different attributes and components measured in selected assessments (global and national).

Table 6. Summary of attributes and components measures in assessments

Attributes	IALS	ALL	PIAAC/STEP	LAMP	Skills for Life	IVQ	LEO	KNALS	BLAS
Defines literacy as a continuum	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Literacy mixed with other domains	No	No	No	No	No	Yes, writing	No	Yes	Yes
Combine prose and document literacy	No	No	Yes	No	No	Yes	Yes	Yes	Yes
Approximate proficiency range covered by main assessment	180-500	180-500	180-500	150 -500	150-500	125-225	100--225	150- 375	100 - 180
Scaling	3 parameter IRT	3 parameter IRT	3 parameter IRT	3 parameter IRT	3 parameter IRT	CTT	1 parameter IRT	1 parameter IRT	CTT
Components measured	none	Letter recognition	Letter recognition	Letter recognition	none		Letter recognition		
		decoding fluency and accuracy	decoding fluency and accuracy	decoding fluency and accuracy			Word recognition		Word recognition

Attributes	IALS	ALL	PIAAC/STEP	LAMP	Skills for Life	IVQ	LEO	KNALS	BLAS
		Working memory	Working memory	Working memory			Sentence processing		
		Receptive vocabulary	Receptive vocabulary	Receptive vocabulary					
			Sentence fluency						
			Passage fluency						
		Spelling		Spelling					
Zero established	No	Yes	Yes	Yes	No	No	No	Yes	Yes
Components analyzed to reveal patterns	N/A	Yes	No	No	N/A	No	No	No	No
Proficiency levels	N/A	Latent class	No	No	N/A	Multi domain conditional raw scores	Quantiles based on Rasch Scores	Quantiles based on Rasch scores	Quantiles based on raw scores
Accurate placement on scale	N/A	No	No	Yes	Yes	No	Yes	Yes	No

In sum, 'lower rungs' approaches allow low-level readers to be placed more accurately on the overall literacy scale.

'Lower rungs' approaches assume, however, that an individual's score on the overall scale means the same thing, in the sense that it implies a clearly defined instructional prescription. Thus, these approaches do not satisfy Messick's notion of validity. Because the distribution of correct items will vary by language in ways that do not reflect differences in proficiency, these measures do not satisfy the need for comparability.

Analysis of the ISRS and PIAAC reading components reveals that this assumption does not hold i.e. a latent class analysis provides much more homogeneous classification of learners needs. The 'lower rungs' approaches adopted in LEO, IVQ, Skills for Life, KNALS and the Bangladesh assessments do not yield lower-level proficiency levels that define groups of learners with homogeneous learning needs.

We would like to recommend that the PIAAC proficiency levels be adopted but that the reading components data for each country be analyzed to reveal groups of learners sharing common patterns of strength and weakness in the target languages. The numbers and proportions of adults in each such group should also be reported annually.

6.8 Topics for discussion

Please discuss the following regarding possible approaches to lower rungs of literacy:

- Adopting PIAAC proficiency levels to provide a comparative context for additional measures in the lower regions of the literacy scale? If so, using the PIAAC reading components approach to improving measurement in the lower regions of the literacy scale?
- Using the lower rungs approach as implemented in Skills for Life and LEO?
- Using a combination of the reading components and lower rungs approaches?
- Abandoning any attempt to assess lower regions of the scale in an internationally comparative way?

6.9 Face validity/familiarity

All of the assessments reviewed, save Bangladesh's assessment, assume that test items provide test takers with all the information they need to respond correctly to an item i.e. background knowledge plays a very limited role in determining the likelihood of correct response.

By extension, performance is dominated by test takers ability to read rather than by what they know. More directly, this approach to assessment assumes that good readers will be able to get items correct despite the fact that they are culturally unfamiliar.

The Bangladesh assessment chose to include items that are, by definition, familiar.

The studies reviewed differ in a fundamental way.

Notwithstanding this dichotomy, the studies reviewed can be classified into two slightly different groups, as follows:

- The IALS/ALL/PIAAC, LAMP and STEP studies that assume that low-difficulty items are culturally neutral and, by extension, that estimated item difficulties are stable across languages and countries and that estimated scores may be compared.
- The Skills for Life, IVQ, LEO, KNALS and Bangladesh studies assume that low-difficulty items are culturally biased among countries but are neutral within countries.

PIAAC-style assessments implicitly assume that adults need to be able to use reading to read and understand things that are outside their experience since many of the world's problems are global.

The IALS/ALL/STEP/LAMP/PIAAC approach is preferable as it would be difficult, or impossible, to determine if idiosyncratic national items are providing comparable measures across countries without applying the full psychometric analysis that is used to scale items in these studies. These methods provide empirical confirmation that items are working and allow designers to set empirical thresholds to exclude non-performing items.

The assumption that idiosyncratic national items function in a stable way within heterogeneous populations, such as the UK Skills for Life, the French IVQ and the Kenyan populations, that include PhDs, blue collar workers and pastoralists, was not empirically confirmed so it remains an open question as to whether these results may be safely compared within countries.

The performance of the items in these latter assessments could be empirically confirmed in several ways.

For example, coding the KNALS items with the PIAAC predictive-difficulty grammar would provide a sense of where the KNALS items would fall on the PIAAC scale.

Scaling the KNALS items using the PIAAC IRT models would confirm the placement of the KNALS items on the PIAAC scale and provide empirical confirmation that the items match their predicted difficulty and that the items are performing in a stable way across population subgroups and languages within Kenya once one has controlled for all background covariates.

Conducting a small equating study in which small samples take both assessments would confirm that the data are generating comparable results.

The LEO study represents a mid-point between the two approaches in the sense that it assumes that the items are performing in a stable way within the German low-skilled population but then goes on to try to predict that factors that underlie the relative difficulty of the tasks.

A second dimension of face validity that differs among the studies is their approach to response modes employed.

The IALS/ALL/PIAAC/LAMP/STEP approach is also very inclusive of all kinds of text – not just words on a page, but also visual displays, graphics, as well as materials that are presented through a digital medium.

The choice of **texts** is also wide ranging and is categorized according to six **characteristics**:

- Medium – print and digital

- Format – continuous and non-continuous
- Type – rhetorical stance (i.e., description, narration, exposition, argumentation, instruction, and records)
- Physical layout – type of matrix organization
- Features unique to digital text – hypertext, interactive, other navigation features (use of scroll bar, utilization of mouse).

There are three **task aspects** that are specified in the PIAAC framework. These three broad cognitive strategies are designed to achieve a full understanding of the texts:

- Access and identify tasks require the reader to locate information in a text. These items include both locating a single piece of information and cycling through the text when more than one piece of information is necessary.
- Integrate and interpret tasks require the reader to understand the relationship between parts of the text (i.e. problem/solution, cause/effect, etc.).
- Evaluate and reflect tasks call on the reader to draw on information, understandings, and knowledge that is external to the text. These tasks include such things as determining the relevance, credibility, etc. of a text. It also includes tasks that look at the register, structure of the text as well as the types of evidence that are provided.

The underlying design consideration in these assessments is that the test should be equally unfamiliar to respondents. The statistical methods used to summarize test results test this assumption explicitly by identifying individuals and items that do not perform within an acceptable range once one has controlled for the full array of background covariates.

The Skills for Life, IVQ, LEO and KNALS studies included items drawn from a broad range of contexts and content in an attempt to ensure that they offered all respondents an equal level of familiarity.

The BLAS study did not include a sufficient number of items to offer coverage of the all content and context dimensions. As a result, the assessment is likely to disadvantage some groups of respondents.

The Skills for Life, IVQ and LEO studies also all included a broader range of test stimuli, question prompts and response modes that were designed to reduce the reading demand associated with understanding the question being asked and to minimize the impact of written response modes on performance. This approach was adopted in order to provide adults with weak reading skills to full opportunity to demonstrate their skills.

While well-meaning, this approach is somewhat misguided if one accepts that literacy skill is the ability to read, understand and apply information. Removing the need to read to acquire the information needed to solve a problem is a deeply ironic accommodation.

7 Defining fixed proficiency levels

In the NALS/IALS/ALL/PIAAC/LAMP/STEP framework proficiency levels are not defined by the imposition of arbitrary cut-off points on the scale but by points at which the underlying cognitive processes shift. Individuals are then placed at a given proficiency level probabilistically by meeting or exceeding a set proportion of items at that level. Provided that predicted and empirically observed item difficulty are in close agreement then placement can be interpreted as a general indication of proficiency. This approach works well in the upper regions of the scale where cognitive strategies explain virtually all of the observed difficulty of items. The available data suggest that these strategies are universal.

Placement in a level in the lower range of the scale is more complex as the predictors of task/item difficulty explain much less of task difficulty.

Two distinct approaches, outlined above, have been made to meet the emerging data need of profiling the skill levels, and implied learning needs, of adults with low levels of reading proficiency.

The first approach, applied in several national studies, involved the development and administration of test items that improve the number of items at the lower end of the IALS/ALL/PIAAC literacy scale(s). These studies include the UK's Skills for Life study, the French IVQ, the German LEO study and the Kenyan KNALS study.

The second approach, initially pioneered by the Canadian and US governments, developed and implemented an assessment of reading components, the skills that must be mastered in order to become a fluid and automatic reader. This approach was subsequently adopted by PIAAC, STEP, LAMP and in the Government of Canada's Test of Workplace Essential Skills – Prime assessment.

Basically, the studies that apply a 'lower rungs' approach allow one to place individuals on the overall literacy proficiency scale much more accurately than the general IALS/ALL/PIAAC/LAMP and STEP assessments. This is useful as it provides an estimate of how far different population subgroups are from any specified target proficiency level.

A weakness in these approaches is that individuals can have the same estimated score on the overall scale but have quite different learning needs and costs.

Properly analysed, reading-components approaches yield a statistically clean categorization of learners that is based on shared patterns of strengths and weaknesses across the reading components that imply a need for a distinct instructional response.

The LEO and IVQ studies include analysis that detects some of the same skills that underlie the groups that are identified by a latent class analysis of the reading-components data.

For example, the LEO analysis of text length and word frequency, both of which figure in the reading-components taxonomy. Importantly, however, the reading components also include measures of decoding fluency and accuracy of pseudo-words and of working memory size that have been shown to have a profound impact on the estimated score of some weak readers, but not others.

8 Additional conceptual issues in design impacting the ability to monitor literacy globally

This chapter sets out a series of additional design decisions that need to be made for the proposed assessment programme if the framework and associated approach to measurement and reporting is to meet the goal of monitoring the Sustainable Development Goals related to literacy.

8.1 How low is 'low'?

If literacy is defined, in the first instance, as the act of reading then it is possible to define someone who is lacks any literacy skill whatsoever.

The ISRS study conducted in Canada and the US and adapted for use in PIAAC, LAMP, STEP and TOWES-Prime provides a tentative answer to this question.

The ISRS study revealed the inclusion of a letter recognition test allowed for the identification of the true absence of literacy skill i.e. adults who were unable to identify a single letter of the alphabet in the language of the test.

This innovation transformed the IALS/ALL proficiency scale from an ordinal scale with an indeterminate lower bound into a true interval scale in which score points along the 500-point scale were, by definition, of equal size.

Prior to this, the IALS/ALL/PIAAC literacy scales were without an absolute zero and the region between zero and 180 points – the value of the simplest 'locate' was essentially indeterminate and without any measurement. In simpler terms, adults can be placed on the scale and safely compared wherever they fall on the scale.

The transformation to an interval scale implies that, even though there may be limited measurement in the lower regions of the literacy scale, the range covered can be compared.

8.1.1 Topic for discussion

- Do the experts agree that the lowest level of literacy can be described as someone who can only recognize a single letter of the symbol set?

Target 4.6 and indicator 4.6.1 specifications also include several subjective elements that require definition including:

- What constitutes a 'substantial' proportion of adults and youth?
- What is the underlying meaning of 'achieve' literacy and numeracy?
- What is the definition of 'functional' relative to literacy or numeracy?

8.2 Defining 'substantial'

In this context, it is difficult to determine what the term 'substantial' should mean.

The main issue lies with the fact that the definition of substantial will depend on the level of economic and educational development and national economic and social objectives. For example, in the economically and educationally advanced countries of the OECD, 'substantial' in 1950 would have implied a need for between 60 per cent and 75 per cent of the youth cohorts graduating from secondary school and 15 per cent having the skills to go on to post-secondary study; targets that imply two levels of literacy and numeracy. Now, the norm in these same countries demands 95 per cent-plus secondary graduation rates and up to 85 per cent of youth cohorts going on to some form of post-secondary study.

In this sense, 'substantial' will vary from country to country in ways that reflect their underlying social, educational and economic objectives.

8.2.1 Topic for discussion

We believe that international reporting should refrain from defining 'substantial' in a single way. This implies publishing a range of estimates of average score and proportions by proficiency level that support a range of comparisons. By extension, each national government should be asked to define a target proportion for 2030 against which progress may be measured.

Do the experts agree that international bodies should refrain from defining 'substantial'?

8.3 Defining 'achieve'

The term 'achieve' carries many meanings, but, in this context, it is taken to imply a level of mastery.

The OECD PISA study chose to adopt a probabilistic mastery standard commonly used in school settings, i.e. individuals getting 62.5 per cent of items at a given proficiency level are placed at that level. In practical terms, this standard is quite low as individuals can be deemed competent even if they get 37 per cent of items at the target proficiency level incorrect. Ironically, the cause of adult literacy and numeracy shortages may lie in the low standards imposed by the primary and secondary education systems.

The IALS and ALL study chose to adopt a more demanding mastery standard, one borrowed from the trades in North America. To be judged to be proficient, tradespeople needed to demonstrate mastery of 80 per cent of the required content, including any tasks requiring reading and numeracy.

PIAAC chose to relax this mastery standard to that applied in the OECD PISA study. While this makes the proficiency estimates from PISA and PIAAC more aligned, it does a disservice to the data, in the sense that very few employers would tolerate employees that got 37 per cent of their reading-based decisions wrong. In addition, recent analysis by the author suggests that the low levels of mastery demanded by the education system in Canada actually induce employers to lower the technical and cognitive demands of their jobs to cope with high proportions of workers with unreliable skill levels. This would not be a problem were it not for the facts that this behavior is associated with a significant loss of output per hour worked and massive skill loss on the part of the workforce that initially had the required ere, skill level.

KNALS and BLAS all rely on thresholds defined by a percentage of correct responses. IVQ uses a more complex formula that employs a percentage of correct responses over multiple sub-domains to yield a nuanced profile of need. LEO and Skills for Life adopt the PIAAC mastery standard.

In reality, however, both of these mastery standards do a disservice to the underlying data if one considers the literacy demands of work. An empirical analysis by the author identified the mastery level that yielded the most reliable classification of proficiency levels in different occupations. The analysis suggests that the optimal mastery level – i.e. the one that yields the cleanest classification of individuals by level – actually varies by occupation. For occupations with high knowledge and skill intensity where the costs of error are high – such as neurosurgeons and other medical specialties – the optimal mastery standard was 95%. For occupations at the other end of the spectrum, where the costs of error were relatively low – such as labourers in agriculture - the optimal mastery level is 5%. Workers in these latter occupations could afford to get 95% of their reading tasks incorrect and still be judged competent.

Since no country collects systematic information on the costs of literacy and numeracy errors, there is little choice but to adopt the prevailing PIAAC mastery standard as a prerequisite to comparison to PIAAC proficiency estimates. Nonetheless, care will have to be taken in making definitive statements about the adequacy or functionality of any given national proficiency distribution.

8.3.1 Topic for discussion: What mastery standard should be applied for 4.6.1?

- Do the experts agree that the mastery standard for 4.6.1 measures should be 62.5 per cent?.

8.4 Defining 'functional'

A second related implication of the NALS/IALS/ALL/PIAAC approach to measurement is a shift from an arbitrary definition of what constitutes being functionally and fully literate. In the past, whether someone was functionally or fully literate depended on the imposition of arbitrary score cut points on an arbitrary proficiency scale. UNESCO defines 'functionally literate' as follows:

A person is defined as 'functionally literate who can engage in all those activities in which literacy is required for effective functioning of his [or her] group and community and also for enabling him [or her] to continue to use reading, writing and calculation for his [or her] own and the community's development'. (UNESCO. 2006. EFA Global Monitoring Report 2006 – Literacy for Life, p. 154 http://www.unesco.org/education/GMR2006/full/chapt6_eng.pdf.)

Several of the studies reviewed, including the IVQ, Kenyan and Bangladesh studies follow earlier practice of defining thresholds that distinguish various levels of functionality.

In contrast, PIAAC, STEP, LAMP, LEO and Skills for Life rely on labels for levels that avoid any implication of functionality.

In the current generations of international comparative assessments, functionality is defined in a variety of ways, including as relative to the reading demands of people's jobs, in terms of the level at which the probability of experiencing poor outcomes is reduced and the level at which people are able to achieve their goals.

Judging whether someone is functionally or fully literate is, by definition, a relative assessment that involves the comparison of the level of reading demands faced by individuals and the skill level they have. This comparison allows individuals to be classified as being in literacy skill shortage, balance or surplus and for an analysis of the impact that each has on economic and social outcomes observed at various levels.

Increasingly, life requires individuals to read, understand and apply information derived from print. Individuals who have the skill levels needed to cope reliably with familiar demands they encounter gain agency and independence. Those who lack the level of mastery needed to cope with unfamiliar reading tasks at a given level of difficulty risk making the wrong decision or of relying on others. In the former case the cost of error may be high, in the latter dependence creates a power relationship in which one cannot assume that interests are perfectly aligned.

So, the first level of functionality is determined by whether individuals have the skill to meet the current reading demands they face in their daily life. Important to note that the proficiency level demanded will vary from individual to individual and some individuals manage to live perfectly satisfying and productive lives without any reading whatsoever.

The second level of functionality is related to literacy as a tool for individuals to deal with unfamiliar reading demands, either proactively or reactively.

Reactive uses of literacy involve individuals having the proficiency level to cope with reading demands that are imposed by external forces – getting a disease and having to understand and apply dosage and contraindications on medications, learning to use a new piece of machinery safely, etc. The available evidence suggests that Individuals who do not have the required level of proficiency bear a significant burden judged in terms of poorer outcomes. They are less employable, work less, earn less, are in poor health, more at risk of experiencing a workplace illness or accident and are less socially engaged.

Proactive uses of literacy involve applying literacy skill to achieve one's goals – the definition of literacy that is embedded in the all of the international comparative assessments of adult literacy, including the OECD's 2012/2014 PIAAC assessment cycles.

In this version of functionality, individuals need a given level of reading proficiency to give them a reasonable probability of realizing their goals. For example, college programmes require a minimum of Level 3 reading proficiency to ensure students get full value out of the experience and have a reasonable probability of persisting to the point of graduation. Individuals without this proficiency level realize demonstrably poorer educational outcomes.

The third level of functionality moves up from the individual to the level of social organizations; for example, firms, education and training providers and hospitals. At this level, there are two sides to functionality. One side is the level of literacy skill assumed by these institutional providers of goods and services. The second side of this level of functionality is the literacy proficiency level possessed by the clients being served by each institution. Inevitably, the literacy proficiency assumed by the providers will be above the proficiency level possessed by some proportion of the clientele.

These misfits between skill supply and demand impose costs on the institutions and on the individuals. For example, firms whose workers lack the literacy skill needed to perform at the production frontier will be less productive and profitable than their competitors who have smaller literacy proficiency gaps. Similarly, providers of health services that are largely publically funded will face higher levels of demands for their services and higher costs of treatment than providers where skill supply and demand. In these cases, both the public and private rates of return will be lower than they might otherwise be. This creates a public interest in policy measures that serve to improve the fit between literacy demand and supply. These measures include both measures to reduce the level of literacy proficiency demanded, for example plain language initiatives, of measures that reduce the skill gap by increasing the supply of literacy skill, or of measures to improve the efficiency of the markets that match skill supply and demand. Thus, by way of example, the functional level of literacy needed to maximize private and public returns on investments in the provision of public goods and services might be higher or lower than that level demanded to maximize individual level returns on literacy acquisition and application.

The fourth and final level of functionality moves the standard up to the societal level. A fundamental role of every government is to seek to improve the welfare of their citizens and to set goals for their improvement be it income, health, social welfare or social equity.

Research clearly shows that average literacy skill proficiency level and the proportion of adults with low literacy skills are the single most important determinants of differences among countries in long-term rates of GDP and labour-productivity growth.

Aggregate differences in the skill levels of employees of social institutions such as firms, hospital and schools, and in the skill levels of these institutions' clienteles, have been shown to influence both public and private rates of return on investment.

Individual skill differences have also been shown to explain most of the social inequality observed in a broad range labour market, health, educational and social outcomes. The net result is that low-skilled adults bear a disproportionate share of poor outcomes. By way of example, the probabilities of Canadian adults with Level 1 or Level 2 literacy being in fair or poor health, of experiencing a spell of unemployment, of being poor and of being socially disengaged are roughly 2.5 times higher that their more skilled peers, even after adjusting for the impact of a broad range of related characteristics such as education, age, gender, language, ethnicity, occupation and immigrant status. These probabilities are amplified in less economically developed and educated countries. This level of skill-based inequality is unfair and places a moral obligation on governments to increase the skill levels of low-skilled adults, particularly because these relationships have been shown to be causal (i.e. low literacy skill causes poor health, labour market, educational and social outcomes) and the instructional approaches exist to raise adult skill level rapidly and at low cost.

By way of summary, functionality can be only be defined in relative terms based on whose interests are being served. More directly, the level of literacy needed for an individual to cope with the reading demands that they confront in their daily lives will differ from the level needed to realize their goals, from the level needed to for social institutions to be efficient and effective and for societies to realize their social and economic objectives.

One of the key objectives to be served by the assessment system is to monitor progress related to Sustainable Development Goal 4.6.

In theory, one could simply publish the proportion of adults falling below a given threshold – for example Level 3 on the PIAAC literacy scale – and the observed change in this percentage over time to meet this objective. Level 3 is an important threshold as it represents where processing moves from the recall processes in the back of the brain to the reasoning processes in the pre-frontal cortex.

In practice, this approach would not satisfy the political need for evidence of progress.

As noted above, the Level 3 threshold is too demanding for the majority of non-OECD countries. Material reductions in the proportions below literacy Level 2, and reductions in the size of key literacy market segments, merit political focus.

Along the same lines, at current and expected rates of investment the proportion of adults below Level 3 will change very slowly with time. Conversely, a country could have realized significant improvement in their average score without having any meaningful impact on the proportion below Level 3 skill.

Most importantly for current purposes, the standard against which countries will judge their own performance, and against which they are judged, will vary with their level of economic, educational and social development and their own economic and social goals. For example, in the first half of 20th Century North America achieved its economic goals by increasing the proportion of the workforce with Level 2 skills. In the second half of the twentieth century the focus turned to increasing the supply of workers with Level 3 skills. Currently, the focus has shifted to increasing the supply of Level 4 and Level 5 skills.

This analysis suggests a need for a nuanced reporting strategy, one that presents:

- Average scores and changes in average scores over time
- The distribution of scores by proficiency level, presented in several ways, i.e. numbers and proportions by proficiency level, the numbers and proportions below Level 2, the numbers and proportions below Level 3 and changes in these proportions observed over time.
- The distribution of literacy by market segments by country that defines the size and associated learning needs and cost of upgrading each segment. Over time the publication of these data will provide a differentiated profile of learners and a first order estimate of the size of the investment that would be needed to bring all adults to Level 2 or Level 3.

Over time, publication of these data will provide an average rate of improvement that, in turn, will allow countries to be classified as over-performing and under-performing.

8.4.1 Topic for discussion

We believe that this implies a need for each country to establish their own definition of what level constitutes the functional level(s), ones that reflect their tolerance for literacy skill-based inequality in individual outcomes, their targets for the performance of key social institutions, including firms and educational institutions, and their social and macro-economic goals.

Because some of the impacts of literacy skill are collective – either influencing the public's return on tax investments or the overall levels of macro-economic performance, social progress and population health governments need to implement this process.

Do the experts agree that international organizations should refrain from imposing a standard definition of functional?

9 Assessment methodology issues: Options for discussion

In an ideal world, every nation in the world would field the OECD PIAAC assessment, the World Bank's STEP assessment or the LAMP assessment as these studies provide data that meet all of the data needs set out above.[4] Conventionally, these assessments include:

- The administration of an extensive background questionnaire that identifies key population subgroups, documents the determinants of skill differences and allows one to explore the impact that skill differences have on individual outcomes.
- The administration of a direct test of adult literacy and numeracy that covers the full range of skill in the population.
- The administration of a direct test of the reading skills that support the emergence of fluid and automatic reading that characterize performance at Level 3 and above.

UNESCO's LAMP assessment was developed to better respond to the needs of less developed countries while maintaining the link to established proficiency scales. More specifically, the LAMP assessment:

- Includes a background questionnaire that has been adapted for use in less economically and educationally developed countries.
- Includes an item pool that includes more low level items that provide more discrimination in the lower regions of the scale.
- Includes a filter booklet that routes less skilled individuals to a less demanding test and the reading components.

[4] Although neither PIAAC nor STEP offer results that are reliable at the individual level, that is needed to support administrative or evaluative purposes the availability of tools such as ETS's partially adaptive literacy and numeracy assessment and Bow Valley College's fully-adaptive TOWES-Prime assessments support individual measurement on the same proficiency scales.

One option would be to resurrect the LAMP assessment programme and implement it in as many countries as possible. This would involve the administration of the LAMP instruments to 3,000 to 5,000 respondents aged 16 to 65.

The LAMP assessment was, however, paper and pencil-based so offers none of the benefits of computer-based, adaptive assessments. LAMP also imposed a significant financial, technical and operational burden on participating countries, one that is likely to exceed the capacities of the target countries.

More pointedly, experience demonstrates, however, that the PIAAC, STEP and LAMP assessments impose too high a financial, technical and operational burden for many less-developed countries. These burdens translate into unacceptably high levels of risk that the estimates of skill levels and distributions will be so biased and error-laden that they are unfit for use.

There is a need, therefore, to identify assessment options that serve to reduce the financial, technical and operational burden associated with fielding a skill assessment such as PIAAC, STEP or LAMP **without** sacrificing the ability to compare results over time and among population subgroups

Separate analysis, undertaken on behalf of UIS, identifies five options that meet these criteria:

1. Reduce the operational, financial and technical burden by administering a skill assessment to a sub-sample of respondents to an existing survey.
2. Reduce the operational, financial and technical burden by administering a skill assessment to a purposive sample of respondents that provides estimates of the probability of being a proficiency level.
3. Reduce the operational, financial and technical burden by administering a fully adaptive web-based skill assessment.
4. Reduce the operational, financial and technical burden by reducing the number of skill domains assessed.
5. Reduce the operational, financial and technical burden by administering a skill assessment whose sole purpose is to classify adults above or below a key threshold, e.g. above or below 275 points, the threshold between literacy Level 2 and Level 3.

Each of these options is described in a report prepared for the UNESCO Institute for Statistics and their advantages and disadvantages summarized. Each of these options could be fielded separately, but in reality, options 1 and 2 could be combined with any of options 3, 4 or 5 to compound the reductions in operational, technical and financial burden.

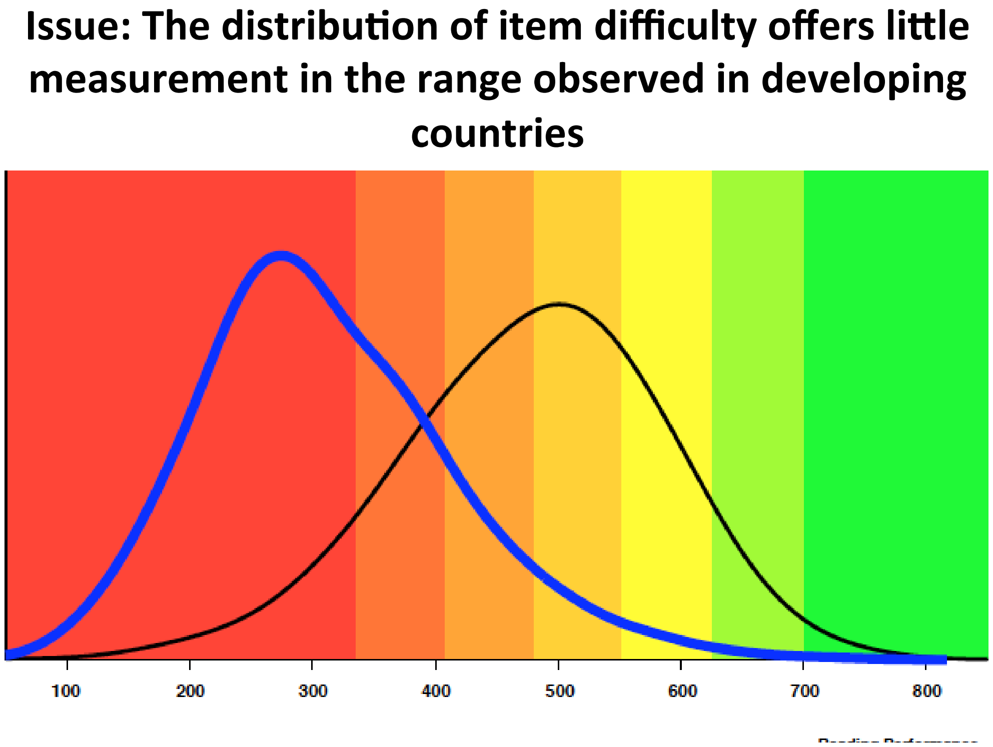
Option 5 is not recommended would involve imposing a single international standard and would not serve the broader national needs for data on literacy distributions.

Of all of the options reviewed, three merit discussion:

- Administering a web-based, fully adaptive assessment.
- Reduce the number of skill domains assessed.
- Move to a tablet platform that does not require access to the internet.

9.1 Administering a web-based, fully adaptive assessment

Figure 7 illustrates the how the fit between the distribution of test item difficulty in a paper and pencil test of the overall literacy scale reduces the information yield for countries with skill distributions below the assumed average. Essentially, countries with lower averages learn little other than that most of their population is in a range where little assessment has taken place.



Issue: The distribution of item difficulty offers little measurement in the range observed in developing countries

Figure 7.

Web-based, fully adaptive testing of the overall literacy scale(s) offers several linked benefits as outlined below it:

- Provides equally reliable results no matter where an individual is on the scale.
- Reduces the average duration of each test by a minimum of 40 per cent.
- Reduces the bias introduced by non-performing items. The system detects the deviation in real time, underweights the items impact on the estimated score and keeps administering test items until the pre-specified precision threshold is reached.
- Yields estimated scores that are reliable at the individual level (rather than the population subgroup.
- Improves respondent relations because most items administered are close to their level of skill.
- Reduces the operational burden of assessment as the system handles all aspects of delivery and yields results in real time. As a result analysis can begin immediately.
- Reduce the risk of cheating.
- Reduces the total cost of fielding an assessment cycle because it displaces a lot of costly and error-prone manual processes involved in test administration and processing e.g. booklet printing, shipping, booklet administration, data capture, scoring and scaling.

Web-based adaptive tests already exist that would serve the purposes laid out above.

The Government of Canada funded the development of the web-based, fully-adaptive Test of Workplace Essential Skills (TOWES-Prime). It measures prose literacy, document literacy and numeracy, as well as reading components, in 11 languages. The TOWES-Prime system allows users to chose one of four precision levels, so the tool is able to support programme evaluation and administrative uses.

ETS, the OECD and the EU have also produced a PIAAC-based, partially adaptive test that is commercially available. Both these tests require access to stable internet connections over a mobile network if required, so could be made available practically anywhere. Procedures have been developed that facilitate adaptation to additional languages rapidly.

9.1.1 Topic for discussion

- Should the assessment programme move to a computer-based platform?

9.2 Reduce the number of skill domains assessed

The theory upon which PIAAC is based identifies separate sets of variables that predict the relative difficulty of prose literacy and document literacy. Both domains were assessed in IALS, ALL and LAMP. In order to reduce the amount of testing time devoted to assessing reading, PIAAC chose to combine the prose literacy and document literacy domains into a single literacy domain. Although scores on the two tests are highly correlated (95 per cent) one sees important gender differences, so combination demands a careful balancing of item types and difficulties over assessment cycles in order to preserve the ability to produce trend estimates overall and for men and women.

9.2.1 Topic for discussion

- Should the assessment programme combine prose literacy and document literacy into a single domain?
- Alternatively, should the assessment focus solely on prose literacy, or document literacy?

9.3 Move test administration to a tablet

It would be operationally useful if the computer-based platform did not require stable internet access, i.e. if it could run on a standalone tablet. This technology exists[5] and would have the added benefit of allowing the use of more visual stimuli, voice prompts and voice response.

[5] See, for example, analytic measures tests

9.3.1 Topic for discussion

Do the experts believe that test administration should move to a tablet? If so, do the experts believe that there is a need to expand the range of stimuli employed in the test? Specifically is there a need for:

- More visual stimuli, i.e. pictures?
- More oral questions?
- More oral responses?

10 Summary and conclusions

This paper has documented the need for estimates of literacy and set out the criteria that must be satisfied if the estimates are to be fit for their intended uses.

The paper also identifies a number of conceptual and operational choices to be made and their likely impact on the fitness of estimates.

This information is meant to facilitate and focus a discussion among experts with a view to reaching a consensus about key elements of design.

Annex A: The origins of the PIAAC reading components measures

As noted above the IALS/ALL/PIAAC item pool was designed to measure the full range of skill observed in the OECD economies so test items had to be developed to cover from roughly 180 to 500.

In this design, a relatively small proportion of the test items could be allocated to the lower regions of the literacy scale. In fact, the easiest IALs and ALL item had a score of 180 and very few items scaled below 225. This design feature initially did not have a material impact on the utility of the proficiency estimates for the OECD countries but, as the assessment programme attracted participating countries with lower average scores, an increasing percentage of the test items fell in parts of the scale where a relatively small proportion of the adult population were classified.

As a result, the relative efficiency of the design fell. For all intents and purposes, countries with averages significantly below the OECD mean skill level learned relatively little about their skill distribution other than that most of their population fell below the level where the design offered much measurement.

The further expansion of the assessment programme to a large number of countries with significantly lower levels of educational attainment and overall lower quality education translates into a need to devote much more attention to the lower regions of the scale.

As noted above, the OECD countries involved in the development and implementation of IALS were not particularly interested in what the data had to say about the learning needs of adults in the lower regions of the proficiency scale. They were much more preoccupied with how their average skill level and distribution of skill in the upper levels of the scale compared to their competitors and, given their levels of educational attainment, did not expect to find large proportions of their adults classified in the lower levels of the scales. The publication of the IALS data led many of these policy makers to begin re-evaluating their assumptions and eventually led to a renewed interest learning more about the skills and learning needs of adults in Levels 1 and 2. In part, this renewed interest was driven by the IALS result that showed the disproportionate share of individual social and economic disadvantage borne by adults in these proficiency levels. This interest crystalized in Canada and the US after publication of the ALL data in 2003 that showed that average skill levels declined between 1994 and 2003 despite the fact that average years of educational attainment increased significantly in both countries over the period. The International Survey of Reading Skills (ISRS) was developed and implemented in 2005 in Canada and the US by Statistics Canada, NCES and ETS in response to this shift in focus.

The ISRS study was designed to assess the component reading skills thought to underlie the emergence of fluid and automatic reading that is needed to master Level 3 and above literacy and numeracy tasks i.e. letter and number recognition, receptive vocabulary, decoding fluency and accuracy and passage fluency. The availability of these measures provided deep insight into the learning needs of Level 1 and 2 adults, a part of the NALS/IALS/ALL proficiency distribution about which little was known.

The LAMP programme adapted the ALL/ISRS methods and assessed prose literacy, document literacy and numeracy, as well as a variant of the ISRS reading components, in 7 countries and 10 languages. Participating countries developed additional test items to both acquaint them with the underlying framework and to add face validity to the item pools. Results were reported on the NALS/IALS/ALL prose literacy, document literacy and numeracy scales. Second order comparisons of reading components were undertaken.

Having established that literacy could be measured in a valid, reliable, comparable and interpretable way, and that differences in literacy skill had a material impact on key indicators of social and economic development, the attention of OECD countries turned towards what they might do to increase the skills of the population and where investments in skill upgrading might yield the largest rates of return on investment.

The work on assessing reading components can be traced back to work undertaken by the Educational Testing Service (ETS) in conjunction with John Strucker at Harvard. Building on Strucker's and Rosie Davidson's clinical assessment work, the Governments of Canada and the US partnered to design and implement the International Survey of Reading Skills (ISRS) expressly for the purpose of profiling the skills of test takers in the lower regions of the IALS/ALL/PIAAC literacy scale.

The ISRS administered a battery of reading component measures that had been shown in Strucker's clinical studies to be prerequisites to the emergence of fluid and automatic reading that characterizes readers at mid-level 2 and up.

The ISRS design was based upon the simple view of reading that posits that the emergence of fluid and automatic reading depends upon mastery of a set of component reading skills – rapid letter recognition, adequate receptive vocabulary, rapid decoding speed and accuracy and adequate working memory.

To measure these component skills the ISRS administered the following standardized clinical assessments:

- The Rapid Alphabetic Naming (RAN) test to gauge letter recognition speed and accuracy
- The Peabody Picture Vocabulary Test (PPVT) to measure vocabulary depth
- The Test of Word Recognition Efficiency Real Word (TOWRE) and the Test of Word Recognition Efficiency Pseudo- word to test decoding fluency and accuracy
- The Digit Span Test to test the size of the working memory. There is reason to believe that stimulatory and nutritional deficits in childhood impair the development of the working memory. By age 11 the working memory becomes fixed, so any deficit will impair adult reading.

The Spelling Test to test the accuracy of spelling. This test is best thought of as an output of having mastered the other component skills rather than as an input to fluid and automatic reading.

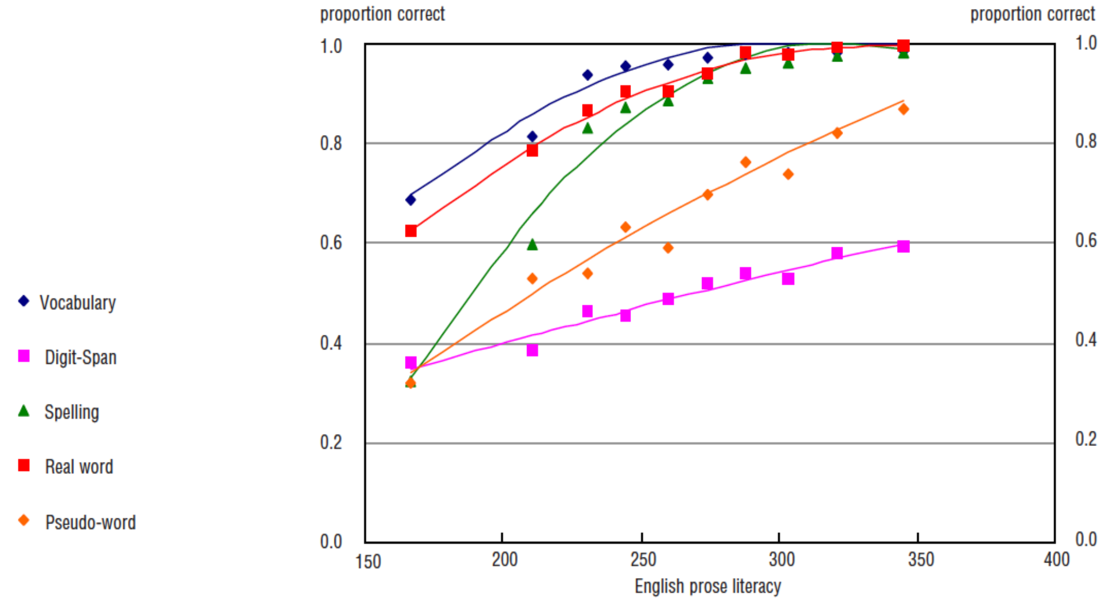
In beginning readers, these skills require the activation of the prefrontal cortex. With practice, the application of these skills shift to the recall processes at the back of the brain and consume significantly less energy to undertake. With the component processing shifted to the long-term memory processes at the back of the brain, virtually all of the available frontal cortex cognitive space can be devoted to applying what they have read.

Importantly for the current review, Canadian ISRS respondents were selected as a representative sub-sample of ALL respondents. In addition to the reading components, the Canadian respondents also had their speaking and listening skills assessed, a feature that was designed to reveal how much of observed weakness in reading skills was attributable to weak oral fluency skills in the languages of the tests. Finally, the Canadian components tests were administered in both English and French, feature that allowed for an exploration of linguistic differences in reading component scores.

The following chart reveals that scores on each of the reading component measures rise steadily with the proficiency score on the prose literacy scale.

Figure B.1

Line graph showing observed scores on each component by score on the IALSS prose scale, English, Canada excluding territories, population aged 16 to 65, 2005



UIS's Literacy Assessment and Monitoring Programme (LAMP) adopted a variant of the ISRS reading components assessments.

A variant of the LAMP reading component measures were then administered to selected respondents in the first and second waves of the OECD's PIAAC assessment programme.

As noted above these assessments are focussed on the skills that underlie the fluid and automatic reading that characterizes level 3 proficiency on the international literacy scales.

Importantly, the component data can only support second order comparisons. For example, they only allow a comparisons of estimates of the proportion of the target population that recognize 80% of the symbol set in a given language, not the absolute number of symbols.

The PIAAC reading component assessment was adapted from measures developed for the 2005 International Survey of Reading Skills (ISRS) jointly developed and implemented by the Canadian and US governments in collaboration with the Educational Testing Service.

PIAAC added tests of sentence and passage fluency to the reading components measures. In our view these measures are redundant, because these skills can be inferred from the fact that the individual has not mastered the skills needed to be placed at Level 3. As such, their administration imposes a response burden that has little

To offer in terms of an instructional prescription.

Importantly, PIAAC was the first adult assessment to include test administration on a computer platform. Analysis of the resulting data suggests that the paper and pencil and computer-based collections yield comparable results even for very low-skilled adults.

The STEP programme adapted the IALS/ALL/ISRS/LAMP methods and item pools for use in a group of less-educationally advanced countries. In order to reduce the operational and financial burden of fielding a comparative assessment, the STEP program the STEP sample was limited to urban areas and the coverage of the skill assessment limited to one skill domain, literacy. The STEP assessment also included the reading components assessment to improve the information yield of the assessment in the lower regions of the literacy scale.

Collectively, the YALS/NALS/IALS/ALL/ISRS/LAMP/PIAAC/STEP programmes have provided a wealth of valid, reliable, comparable and interpretable data that:

- Confirms the theories of task difficulties that underlie the items
- Provides empirical confirmation that the test is working in a stable way in heterogeneous populations within and between in a wide range of languages.
- Confirms the existence of meaningful differences in the average levels and distributions of skill both within and between countries
- Confirms that these differences in the level and distribution of skill are associated with meaningful differences in outcomes and underlying policy choices

The fact that these differences in the level and distribution of skill and outcomes have been shown to be causal and tied to broadly defined policy choices creates a moral/ethical obligation on the part of policy makers to focus their attention on fostering skill demand, in creating more and more equitably distributed skill, in improving the efficiency of markets that match skill supply and demand and in ensuring maximal skill utilization.

Having established that literacy could be measured in a valid, reliable, comparable and interpretable way across languages and cultures, and having documented that the observed relationships with outcomes were both material and causal, policy makers faced a moral obligation to improve average levels of literacy skills, to reduce the proportion of adults with skill below level 3 and to reduce levels of social inequality in skill.

Importantly, this obligation translates into a need to know much more about the lower end of the proficiency scale than the main assessment can provide.

The ISRS study revealed several other important facts for the current review including:

The inclusion of a letter recognition test allowed for the identification of true absence of literacy, i.e. adults who were unable to identify a single letter of the alphabet in the language of the test. This innovation transformed the IALS/ALL proficiency scale into a true interval scale in which score points along the 500-point scale were, by definition, of equal size. Prior to this the IALS/ALL/PIAAC literacy scales were without an absolute zero and the region between zero and 180 points – the value of the simplest "locate" was essentially indeterminate and without any measurement. In simpler terms adult's can be placed on the scale and safely compared wherever they fall on the scale.

Performance on the component measures differed significantly between English and French. These differences were determined to be functions of the differing orthographic structure of the languages, most particularly that the act of decoding was significantly easier in French. Subsequent testing in Spanish revealed that decoding in that language was even easier because of the fact that all letters are sounded so respondents can access their knowledge of the spoken word to help them decode unfamiliar words. These findings imply that one cannot compare reading component scores directly across languages. At best, one can make second order comparisons, such as the proportion adults who recognize 80% of the symbol set for their language rather than the actual number of letters recognized in a given period of time. Nevertheless, conventional definitions of fluid and automatic reading suggest that beginning readers need to achieve mastery levels of 80% or better in each of the reading component skills to become fluid and automatic readers.

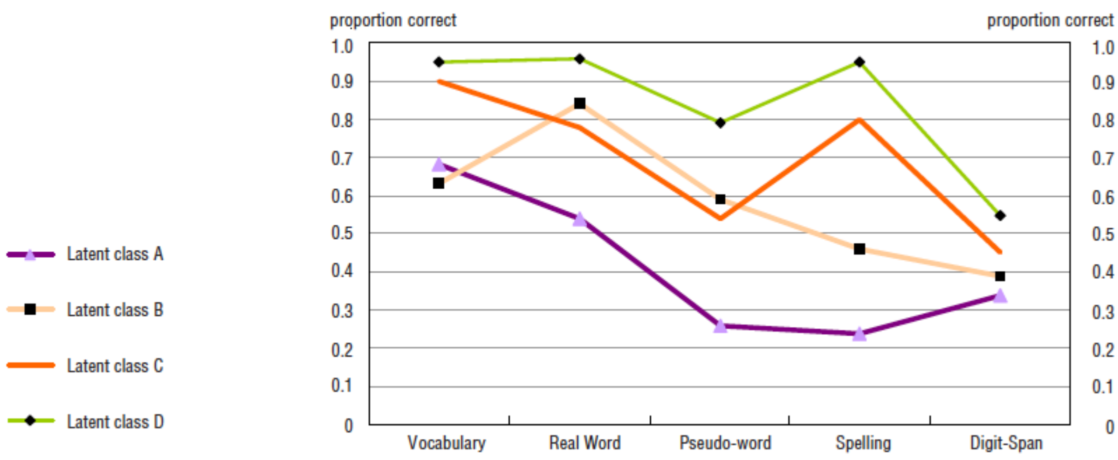
The oral fluency measures available for Canada revealed that a large proportion of adults with low literacy and numeracy scores, both immigrants and non-immigrants, also had weak speaking and listening scores in the language of the test. Most of the adults had reasonable levels of education and mother tongues other than the language of the test, a finding that suggests that they were literate in their own language but were still working on the transfer to English or French. A second group had a low education levels mother tongues other than the language of the test, so were unlikely to be literate in their own language.

A third, much smaller, group had the language of the test as a mother tongue, reasonable oral fluency scores and very weak component skills, a pattern that strongly suggests learning disabilities such as dyslexia. Being able to differentiate these groups is fundamentally important, as the instructional prescriptions for the groups are dramatically different. This insight is fundamentally important in developing countries many of which have significant linguistic diversity.

A latent class analysis of ISRS reading component data, shown in the following chart, revealed the presence of four distinct groups each of which shares a common pattern of strength and weakness across the components and, importantly, quite distinct instructional responses.

Figure B.2

Average proportion correct scores on each component displayed separately for each latent class, Canada excluding Territories, population aged 16 to 65, 2005

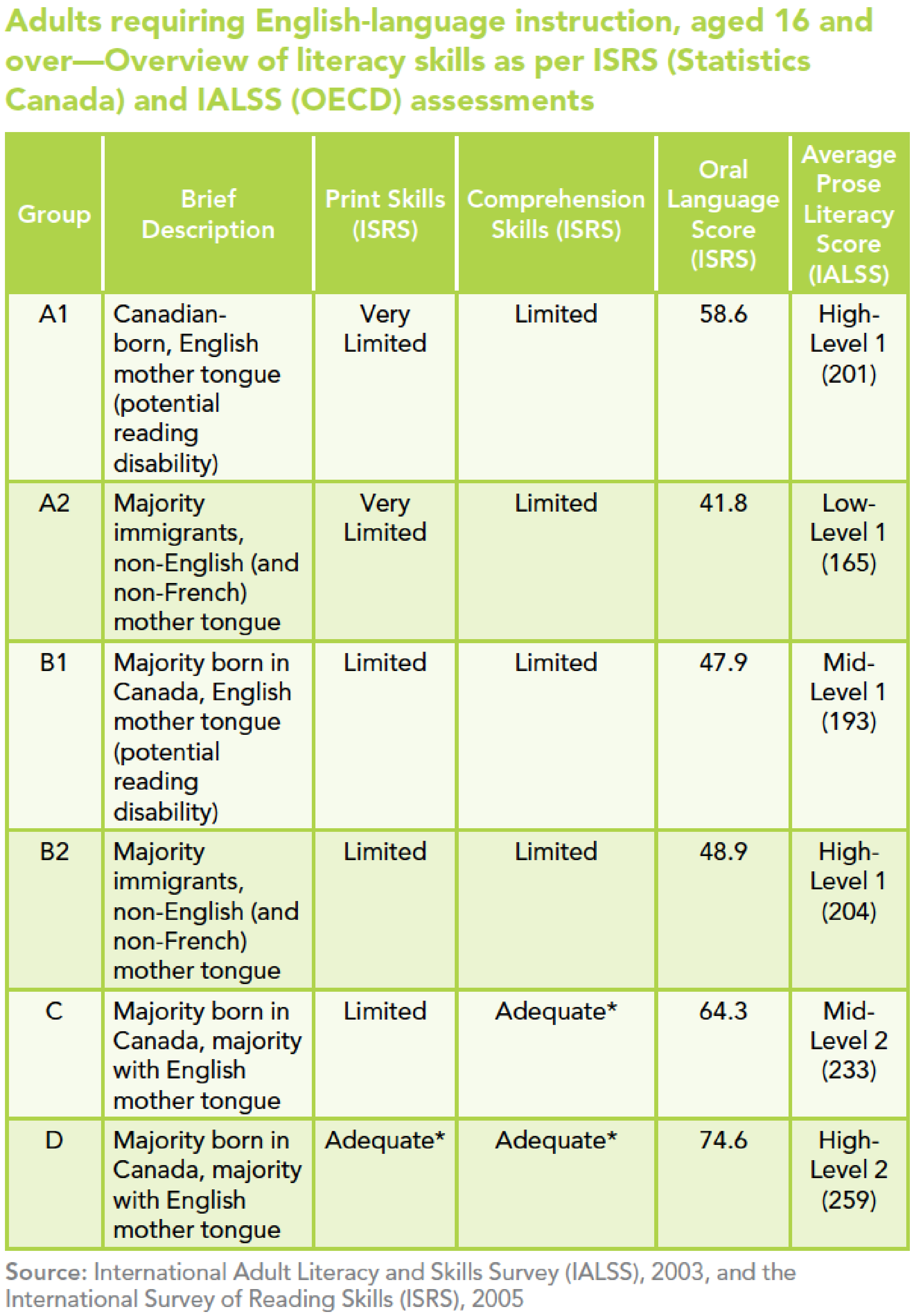


Note: See Table C.4.6 in Annex C.

-

As documented in the following table, subsequent analysis of the Canadian data lead to the definition of two distinct groups within both latent class A and Latent Class B based upon whether the individual was a non-official language immigrant or not. The final classification thus identifies 6 distinct groups. Each group shares a common set of strengths and weaknesses and who would benefit from a distinct instruction response to raise their skill level.

- Figure B.3



Adults requiring English-language instruction, aged 16 and over—Overview of literacy skills as per ISRS (Statistics Canada) and IALSS (OECD) assessments

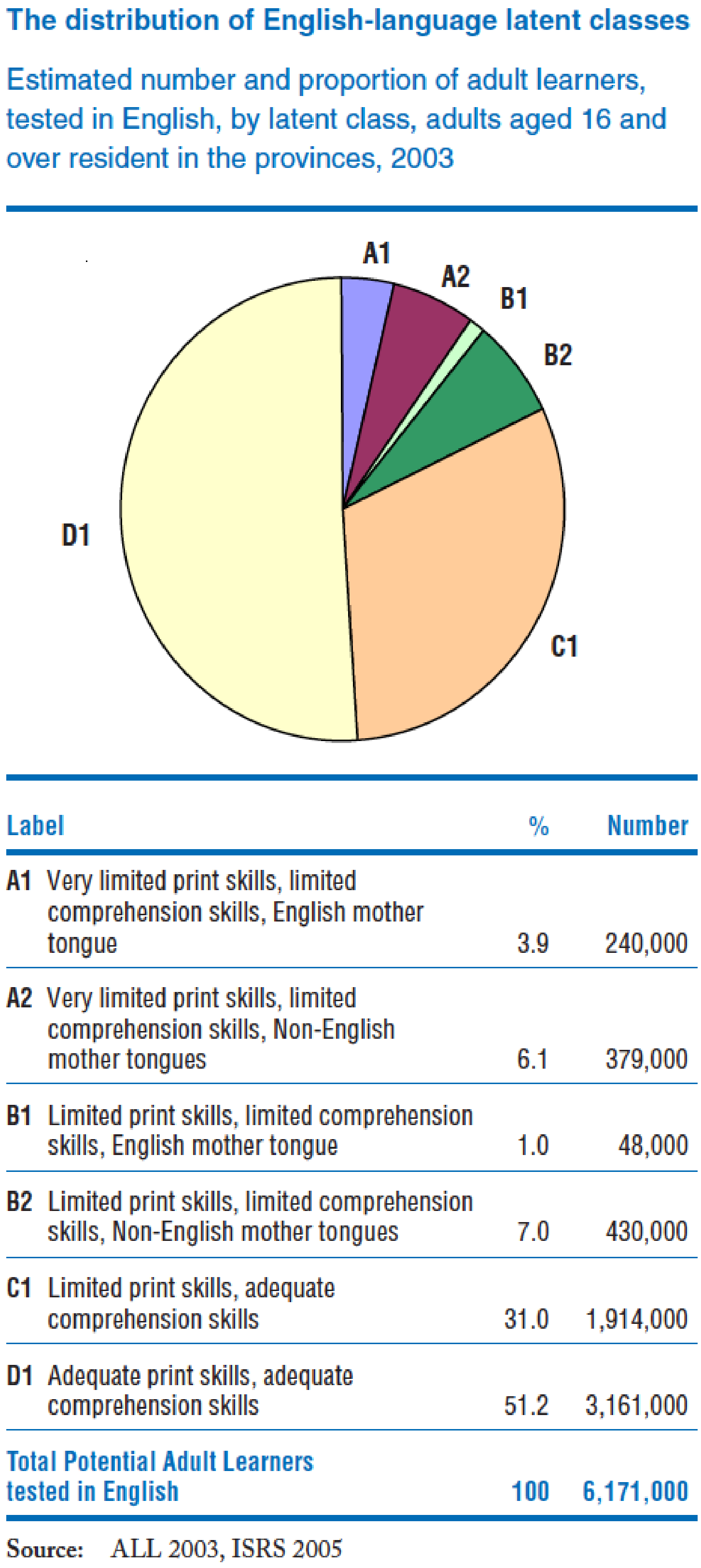
Group	Brief Description	Print Skills (ISRS)	Comprehension Skills (ISRS)	Oral Language Score (ISRS)	Average Prose Literacy Score (IALSS)
A1	Canadian-born, English mother tongue (potential reading disability)	Very Limited	Limited	58.6	High-Level 1 (201)
A2	Majority immigrants, non-English (and non-French) mother tongue	Very Limited	Limited	41.8	Low-Level 1 (165)
B1	Majority born in Canada, English mother tongue (potential reading disability)	Limited	Limited	47.9	Mid-Level 1 (193)
B2	Majority immigrants, non-English (and non-French) mother tongue	Limited	Limited	48.9	High-Level 1 (204)
C	Majority born in Canada, majority with English mother tongue	Limited	Adequate*	64.3	Mid-Level 2 (233)
D	Majority born in Canada, majority with English mother tongue	Adequate*	Adequate*	74.6	High-Level 2 (259)

Source: International Adult Literacy and Skills Survey (IALSS), 2003, and the International Survey of Reading Skills (ISRS), 2005

These differences are singularly important for policy as the number of hours of instruction required varies by a factor of 10, and the unit cost of instruction per learner varies by a factor of 20. Moreover, the expected benefits that would accrue to higher skill levels vary significantly by group, enough to imply very different rates of return on public and private investments in skill upgrading.

- Unfortunately, although the average literacy score of each group rises from a low of 165 to a high of 259 they do not do so in the monotonic way assumed in the "lower rungs" approaches. More specifically the "lower rungs" approach assumes that adults classified in a given level share common learning needs, common instructional responses, would require the same number of hours, would require the same level of investment and would yield the rates of return on investment. In reality, the available data reveal that individuals can arrive at the same estimated score on the scale despite having dramatically different patterns of reading component scores. Concomitantly, each latent class displays significant variation in the distribution of the constituent component measures. In sum, only patterns of strength and weakness in the component measures yield clear instructional guidance.
- To make matters worse for policy makers, the relative size of the groups varies significantly so that the aggregate costs and benefits of each group vary a great deal. The following charts present the distribution of adult's by latent class and the estimated costs of raising each group to Level 3 through "best practice" instruction.
- Thus, the proficiency levels provided by the "lower rungs" approaches are likely to mislead policy makers, causing them to underestimate the cost of moving low skilled populations up the proficiency scale and leading them to allocate skill upgrading investments sub-optimally, to groups that don't yield the highest rates of return.

Figure B.4



The distribution of English-language latent classes

Estimated number and proportion of adult learners, tested in English, by latent class, adults aged 16 and over resident in the provinces, 2003

Label		%	Number
A1	Very limited print skills, limited comprehension skills, English mother tongue	3.9	240,000
A2	Very limited print skills, limited comprehension skills, Non-English mother tongues	6.1	379,000
B1	Limited print skills, limited comprehension skills, English mother tongue	1.0	48,000
B2	Limited print skills, limited comprehension skills, Non-English mother tongues	7.0	430,000
C1	Limited print skills, adequate comprehension skills	31.0	1,914,000
D1	Adequate print skills, adequate comprehension skills	51.2	3,161,000
Total Potential Adult Learners tested in English		**100**	**6,171,000**

Source: ALL 2003, ISRS 2005

The size of the English Literacy market by market segment based on the estimated cost of raising all potential learners to prose Literacy Level 3, 2003

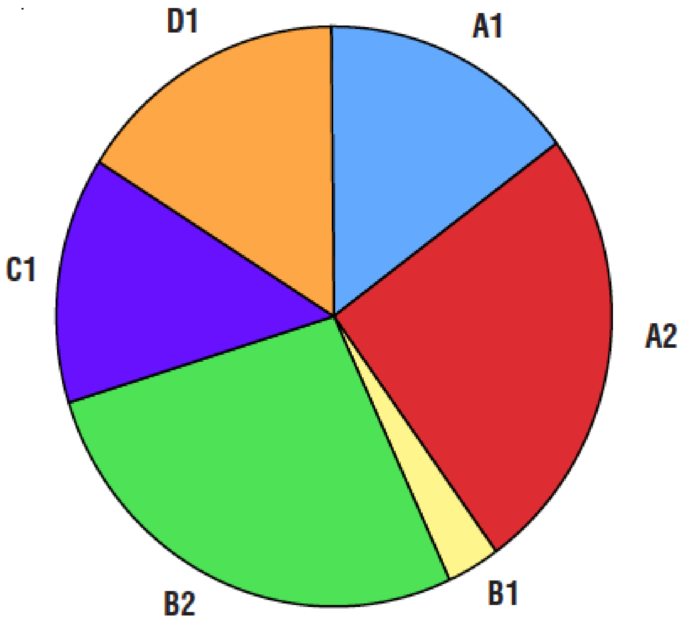
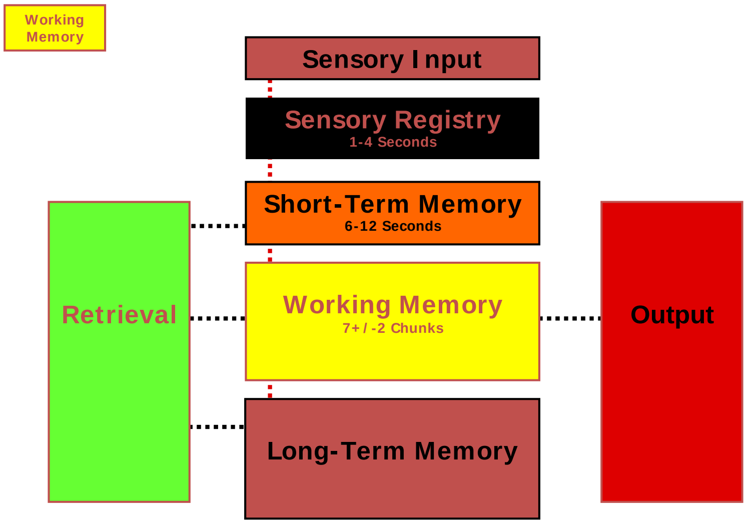


Figure B.5

Label		%
A1	Male dropouts, reading disabled	15.0
A2	Immigrant women, little education	25.0
B1	Male, high school only	3.0
B2	Educated immigrant women	27.0
C1	Slight problems with decoding and comprehension	14.0
D1	No problems with mechanics of reading, lack skill to get to Level 3	16.0

It is worth reflecting on what research says about the biological processes that underlie the acquisition and application of reading. The basic structures that supports the act of reading is shown below

Figure B.6



Retrieval and output processes engage five systems. individuals in the process of learning to read must devote their entire cognitive space to recognizing letters and words.

With practice, letter cognition, word recognition and decoding processes increasingly rely on the long-term memory. As illustrated below, less and less of the available cognitive space is used to support the act of reading, something that frees up space for applying what has been read.

Importantly, this process, known as pixilation, actually reduces the amount of energy needed to perform the task.

The process of pixilation continues into the higher literacy proficiency levels. Expert readers are able to retrieve large bodies of knowledge, hold the “chunks” of information in working memory and apply the information to solving problems.

As noted earlier, the component measures interact in different ways to support the emergence of fluid and automatic reading in different languages. Or example, the act of decoding is significantly simpler in Spanish, a language in which every letter is sounded, than English that has lots of silent letters. In Chinese, letter recognition and decoding are folded into receptive vocabulary. National standards in China stipulate that to become a fluid and automatic reader adults need to be able to recognize 1500 characters. Although fluency and automaticity emerge in different ways in different languages, processing at literacy levels 3 and above depends on adults having acquired this skill in all languages.

This analysis holds important implications for any global assessment of reading in the lower regions of the scale.

First, assessments will have to be developed for each language. The key difference among these measures will be the assessments of decoding fluency and accuracy that will vary depending upon the degree to which the orthographic structure of the written word matches the spoken word.

Second, because the component measures, and relationship among the component measures, will vary among languages, the statistical methods that are used to identify groups of learners that share common patterns of strength and weakness, and to identify associated instructional prescriptions, will need to be repeated for each language.

Third, sampling strategies will have to reflect the interactions of language of test with mother tongue and language(s) of instruction among immigrant and non-immigrant adult populations.

The fact that these latter two conditions were not implemented in PIAAC, LAMP or STEP greatly limits the utility of these data for policy formulation.

Analysis of reading components data for multiple languages suggests that performance depends on mastery of the reading components. Adults who have yet to master the reading components are still in the process of learning to read and must devote most of their cognitive resources to the act of reading. In contrast, adults that master all of the components are, by definition, fluid and automatic readers who can devote the bulk of their cognitive resources to building meaning.

Since the complexity of these component tasks varies by language, the component tasks can only be compared at the second level e.g. what proportion of the population can recognize 80% or more of the symbol set used to represent the language. Analysis of data for Canada identifies six specific groups of learners, each with a unique pattern of strength and weakness in their reading components and each of which demands a unique instructional response. In other languages analysis of the reading component data will reveal different groups.

In this sense the lower regions of the NALS/IALS/ALL/LAMP/PIAAC/STEP scales are not strictly uni-dimensional.