Int Rev Educ (2011) 57:199–217
DOI 10.1007/s11159-011-9203-2

Measuring the continuum of literacy skills among adults: educational testing and the LAMP experience

Cesar Guadalupe · Manuel Cardoso

Published online: 5 May 2011

Abstract The field of educational testing has become increasingly important for providing different stakeholders and decision-makers with information. This paper discusses basic standards for methodological approaches used in measuring literacy skills among adults. The authors address the increasing interest in skills measurement, the discourses on how this should be done with scientific integrity and UNESCO's experience regarding the Literacy Assessment and Monitoring Programme (LAMP). The increase in interest is due to the evolving notion of literacy as a *continuum*. Its recognition in surveys and data collection is ensured in the first commitment in section 11 of the Belém Framework for Action. The discourse on how measurements should be carried out concerns the need to find valid parsimonious approaches, also their relevance in different institutional, cultural and linguistic contexts as well as issues of ownership and sustainability. Finally, UNESCO's experience with LAMP shows how important addressing these different issues is in order to equip countries with an approach that is fit for purpose.

Keywords Literacy · Adult literacy · Literacy skills · Literacy measurement · Literacy survey · Literacy assessment · Educational testing · Reading skills · Numeracy skills

Resumé Mesurer le continuum de l'alphabétisme des adultes : évaluation des acquis et expérience LAMP – Le domaine de l'évaluation des acquis de

Both authors work for the UNESCO Institute for Statistics (UIS) at the unit responsible for the Literacy Assessment and Monitoring Programme (LAMP). The opinions expressed in this essay are the exclusive responsibility of the authors.

C. Guadalupe (✉) · M. Cardoso
UNESCO Institute for Statistics (UIS), Montreal, Canada
e-mail: c.guadalupe@uis.unesco.org

M. Cardoso
e-mail: m.cardoso@uis.unesco.org

l'apprentissage gagne aujourd'hui en importance car il convient de fournir les informations aux diverses parties prenantes et aux décideurs. Cet article présente des critères de base pour les approches méthodologiques utilisées pour mesurer les compétences de lecture et d'écriture chez les adultes. Les auteurs traitent l'intérêt croissant pour la mesure des compétences, les discours sur les moyens d'y procéder selon les règles scientifiques, et l'expérience de l'UNESCO avec le Programme d'évaluation et de suivi de l'alphabétisation (LAMP). Cet intérêt accru est dû à la nouvelle notion de l'alphabétisme considéré en tant qu'un continuum. La prise en compte de cette notion dans les enquêtes et collectes de données est assurée dans le premier engagement du point du Cadre d'action de Belém. Le discours sur la façon dont ces mesures doivent être effectueés renvoie au besoin de trouver des méthodes valables parcimonieuses, à leur pertinence dans différents contextes institutionnels, culturels et linguistiques, ainsi qu'aux questions d'appropriation et de durabilité. Finalement, l'expérience de l'UNESCO avec le Programme LAMP montre l'importance de répondre à ces différentes questions afin de doter les pays d'une approche qui puisse permettre d'atteindre les objectifs.

Zusammenfassung Wie misst man die Lese- und Schreibfertigkeiten von Erwachsenen als Kontinuum? Bildungstests und die Erfahrungen mit LAMP – Wenn es um die Versorgung von Interessenvertretern und Entscheidungsträgern mit Informationen geht, spielen Bildungstests eine immer wichtigere Rolle. In diesem Artikel werden grundlegende Standards für methodologische Ansätze zur Messung der Lese- und Schreibfertigkeiten von Erwachsenen erörtert. Die Autoren behandeln das zunehmende Interesse an der Messung von Fertigkeiten, die Diskurse, wie solche Messungen wissenschaftlich seriös durchzuführen sind, und die Erfahrungen der UNESCO mit dem Literacy Assessment and Monitoring Programme (LAMP). Das Interesse nimmt zu, weil Alphabetisierung heute immer mehr als Kontinuum begriffen wird. Die Anerkennung dieses Konzepts in Erhebungen und Datensammlungen wird im Aktionsrahmen von Belém in der ersten Verpflichtung unter Punkt 11 gewährleistet. Bei der Diskussion um die richtige Durchführung der Messungen geht es um die Suche nach fundierten und zugleich sparsamen Ansätzen, um ihre Relevanz in verschiedenen institutionellen, kulturellen und sprachlichen Kontexten sowie um Fragestellungen der Eigenverantwortung und Nachhaltigkeit. Die UNESCO hat mit LAMP die Erfahrung gemacht, dass diese verschiedenen Aspekte behandelt werden müssen, um Ländern Instrumente zur Verfügung stellen zu können, die ihren Zweck erfüllen.

Resumen Medir el continuum de las habilidades de alfabetismo en adultos: evaluaciones educativas y la experiencia con el programa LAMP – El campo de la evaluación educativa se ha vuelto cada vez más importante en cuanto a la información que suministra a los diferentes grupos de interés y tomadores de decisiones. Este trabajo se ocupa de estándares básicos en los enfoques usados para medir habilidades de alfabetismo en personas adultas. Los autores abordan el creciente interés en la medición de habilidades, los discursos sobre cómo ésta debe realizarse con integridad científica y la experiencia de la UNESCO con el Programa de Evaluación y Monitoreo de la Alfabetización (LAMP). El creciente interés se debe a la evolución que ha tenido la noción de alfabetismo en tanto continuum. Su

reconocimiento en estudios y generación de datos está asegurado como primer compromiso en el punto 11 del Marco de Acción de Belém. El discurso sobre cómo deberían realizarse las mediciones concierne la necesidad de encontrar enfoques parsimoniosos válidos, al igual que su relevancia en diferentes contextos institucionales, culturales y lingüísticos, as como aspectos de apropiación y sostenibilidad. Por último, la experiencia de la UNESCO con el programa LAMP muestra la importancia que tiene abordar estos temas a efectos de dotar a los países de un enfoque que sea suficiente para lograr su propósito.

Резюме Измерение уровня грамотности среди взрослых: образовательное тестирование и опыт программы LAMP – Область образовательного тестирования становится более и более важной для предоставления информации различным заинтересованным сторонам принимающим ключевые решения в области образования. В данной статье обсуждаются базовые стандарты методологических перспектив используемых для измерения уровня грамотности среди взрослых. Авторы статьи обращают внимание на растущий интерес к измерению уровня грамотности, а также на дискуссии о том как его следует проводить с научной точностью, с учетом опыта ЮНЕСКО относительно программы мониторинга и оценки уровня грамотности (LAMP). Рост интереса связан с развивающимся понятием грамотности как континуума. Его использование при анкетировании и сборе данных отмечено в первом пункте части 11 Программы действий, принятой в Белене. Дискуссия о том, как следует проводить измерения, затрагивает проблему поиска эффективных и подходящих подходов, а также их релевантность в разных институциональных, культурных и лингвистических контекстах, включая вопросы права собственности и устойчевого развития. И наконец, опыт ЮНЕСКО с LAMP показывает как важным является обсуждение этих различных вопросов для предоставления странам подхода который соответствует поставленным целям.

Introduction

For many decades statistical information on literacy has been restricted to what is known as *literacy rates*. These are usually computed on the basis of a single question put to individuals through a population census or household survey.[1]

[1] In 1958, UNESCO's General Conference approved a series of recommendations concerning the standardisation of educational statistics. These included a simple definition of literacy (see UNESCO 1958) that was later echoed by the United Nations Statistical Division (UNSD 1997, 2.145). These form the basis on which population census and household surveys have structured the questions they posed. In this context, literacy rates are defined as being the "total number of literate persons in a given age group, expressed as a percentage of the total population in that age group" (UIS 2010, p. 269). Nevertheless, the notion of literacy has evolved as shown in different UNESCO documents (UNESCO 1978, 2004, 2005) and this evolution was also echoed by the United Nations Statistical Division in the most recent revision of the above-mentioned document: "Literacy has historically been defined as the ability both to read and to write, distinguished between 'literate' and 'illiterate' people. A literate person is one who can both read and write a short, simple statement on his or her everyday life. (…) However, new understanding referring to a range of levels, of domains of application, and of functionality is now widely accepted. (…)

In contexts where literacy rates are extremely low, these figures seem good enough to identify literacy as a major policy priority, while under other circumstances this can actually veil literacy problems. In any case, the information they provide is too limited to inform the actual design of policy interventions.

This evidence is limited, given the nature of literacy and how this has been represented in its definitions over the past decades. While 50 years ago, literacy was understood as being the "ability to read and write", it has subsequently become evident that this "ability" is not a single trait that can simply be possessed or not by an individual. As recognised by the Belém Framework for Action (UIL 2010), it is a manifold phenomenon (including the abilities to read, write and compute as instantiated in different contexts) and its measurement should consider this complexity as well as the fact that each of those skills exists over a continuum ranging from very low levels, where individuals can hardly perform the most basic tasks, to higher levels, where individuals are able to perform complex tasks using complex written materials.

Thus compiling the numbers of individuals who report being able to read and write does not provide any significant evidence on their actual skills. As a result, measuring those skills appeared as an emerging and crucial issue.

Consequently, over the past decades, several countries have conducted specific literacy surveys which usually include a test of skills. The intentions behind these efforts are: (i) to generate evidence on the actual distribution of skills (rather than just formal credentials or self-reported literacy status) across the population, (ii) to have detailed evidence on the profile of those requiring policy interventions and (iii) to have a better understanding of the relationship between skills (and not formal credentials) and other socio-economic variables.

These efforts have been developed concurrently with other forms of educational testing intended to yield information on student achievement.[2] Certainly, over the past decades an increasing body of evidence shows that educational attainment[3] is not necessarily a good indication of what an individual knows and is able to do. The whole debate around the identification of educational quality and the level of student achievement shown using standardised tests rests upon this evidence.

Nevertheless, skills measurement is a highly controversial topic. Current debates refer both to the political agenda which standardised testing might serve and to the feasibility of measuring skills using standardised testing.

Footnote 1 continued
Nevertheless, administering a literacy test to all household members in the course of enumeration may prove impractical and affect participation, therefore limiting the utility of the results. Countries have regularly used simple self-assessment questions within a census to provide an indication of literacy rates at the small area level. An evaluation of the quality of statistics should be provided with census statistics on literacy" (UNSD 2008, pp. 147–148).

[2] At international level, the most important and pioneering efforts have been conducted by the International Association for the Evaluation of Educational Achievement (IEA, www.iea.nl) since the late 50s, and then followed by the Organisation for Economic Co-Operation and Development (OECD) since the late 90s as well as regional initiatives in Latin America and Africa. In any case, these international efforts rely on and/or promote national capacities usually organised in educational testing units in the Ministries of Education.

[3] Educational attainment is usually measured in number of years of schooling attended (excluding years spent repeating the same grade), highest level attended or completed, or highest certification acquired.

For some authors (Revell 2005; Mahony and Hextall 2000), standardised testing is just a mechanism of over-centralised control which has been put in place by undermining the local and professional autonomy that used to characterise some educational systems, in order to favour a centralised power structure (Power 1997). At the same time, governments implementing standardised testing mechanisms stress the importance of accountability when the citizenry's right to education and proper utilisation of public resources are at stake. In this regard, measuring what people know and are able to do is mostly intended to show the overall performance of the education system and the challenges it has to face, rather than inform on individual performance.

The appropriateness of standardised testing is contested on the basis of considerations regarding teachers' professional autonomy (they should be the only ones who appraise student achievement) and by others who are more concerned with the effects that local context plays in testing (Cooper and Dunne 1998). Ultimately, the latter form of criticism aims at contesting the actual feasibility of having standardised tests which yield comparable information across individuals.

At the same time, test respondents who have been schooled, and therefore exposed to testing, may perform in a different way than those lacking this experience. To what extent does the degree of readiness to take a test vary across individuals and groups? What is the relationship between classroom assessment and an external assessment intended to yield information for policy purposes? We will not address these debates here, but we are fully aware of them, for they underlie the design of educational tests and should not be overlooked.

This paper takes as its starting point the fact that several countries conduct standardised tests, and show increasing interest in measuring literacy skills among their youths and adults. This interest might become even more evident in the light of the Belém Framework for Action's treatment of literacy and its measurement and monitoring issues (UIL 2010, section 11)[4] Therefore, this paper intends to provide countries and practitioners in general with some guidelines and key elements to be factored into the national debates on literacy skills measurement.

This article extensively relies on what the UNESCO Institute for Statistics has learnt in the process of designing and validating the Literacy Assessment and Monitoring Programme (LAMP).[5]

Educational testing

Skills testing is an area that has been developed over the past decades in a significant way. From a situation where testing was basically one of the resources used by teachers to appraise student progress, educational tests have evolved into

[4] This framework expresses countries' commitments endorsed by delegations (UNESCO's 144 Member States, representatives of civil society organisations, social partners, United Nations agencies, intergovernmental agencies and private sector) at the Sixth International Conference on Adult Education (CONFINTEA VI, held in Belém, Brazil in December 2009). In the above-mentioned section (11a), countries committed to "ensuring that all surveys and data collection recognise literacy as a continuum".

[5] For additional information on LAMP see UIS (2009).

the realm of complex psychometric techniques and survey designs. This field has become extremely important to provide different stakeholders with instruments and information to make decisions in different areas such as: higher education admissions; hiring processes; overall levels of student achievement (usually referred to as information on the "quality" of education), etc.

A first element to be singled out here is purpose. Some tests are designed to yield individual ability measures, and therefore the emphasis is on producing accurate point estimates for each examinee. Others aim to portray a general picture of the educational situation; here the emphasis is on both characterising the typical skill levels and showing their variability across the population, including the performance of some sub-populations, in order to inform policy development; in this context, individual scores are less important. Therefore, the former must be administered to each individual in a given group (as if in a "census", of a classroom, for instance), while the latter are conducted more efficiently by using randomised samples.

T. Neville Postlethwaite (2004) presents a clear and didactic exposition of the main characteristics of this second group of studies. He identified some major "technical standards for sample survey work in monitoring educational achievement."[6] These standards, slightly rephrased, will be used here to organise the following discussion, even if the actual argument will depart from Postlethwaite's text so as to match this paper's focus more closely.[7]

1. A clear and explicit identification of the study's aims and purpose

This element might sound obvious but unfortunately that is not the case. Sometimes countries get involved in educational testing exercises without having a clear idea of what they are going to do with the resulting information. Actually, in an ideal situation a clear identification of what the use of the data is going to be should precede the study's development.

The purpose of the study is a major element in establishing its design and one of the most frustrating situations arises from a potential mismatch between expectations and what is feasible with the evidence a study generates. A latent "solution" to this situation is to *torture*[8] the evidence to provide (weak) answers to questions it was not designed to answer. This only transforms useless evidence into misleading evidence, which is worse.

The study's purpose must lead to a definition of its scope and main attributes. For instance, including "general knowledge" as an integral part of literacy may introduce noise into the construct to be measured, since "general knowledge" might be acquired and sustained outside the realm of the written word. A reading test, for

[6] This wording is the title of chapter V in Postlethwaite (2004).

[7] Postlethwaite's text is mainly focused on educational testing as conducted in schools. Measuring literacy skills of the youth and adult population entails some specific characteristics, given that it has to be conducted using a household survey platform and that the actual subject matter is embedded in a theoretical discussion about literacy, rather than, as it is usually the case in school-based testing, on the prescriptions of a given curriculum.

[8] This word of alert is usually attributed to the British economist Ronald Coase having said: "If you torture the data long enough, it will confess."

example, should focus on the examinees' ability to derive meaning from a given text, not on the knowledge they acquired prior to the test, by reading or by any other means. Thus, inclusion of "general knowledge" in a literacy test poses a validity problem: does the study measure what it intends to?[9]

A similar problem arises regarding the inclusion of numeracy elements when measuring literacy skills. Individuals can perform computations either by writing down numbers and operators, or without this aid. Are both of these types of computation part of the literacy realm? Should purely oral computations be considered in a literacy test? At the same time, just because someone can identify written numbers or draw their shapes, this does not mean that he or she has the quantitative skills he or she needs, which would probably entail: knowledge of certain mathematical concepts, the ability to apply computational algorithms and the capacity to solve problems by using numbers.

Another potential problem refers to the social contexts where literacy and numeracy skills are used. For instance, when a test intends to measure numeracy skills by using mostly market-related situations (as if a meaningful use of numbers would refer exclusively to prices, discounts, etc.), we would also be facing a validity issue, because this constrains the construct's real scope. This issue is even more serious in contexts where subsistence agriculture plays an important role, or where payment in kind is used often, as opposed to legal tender, in commercial transactions.

2. A proper definition of the target population

Together with the overall purpose, properly defining the target population is extremely important since it defines the scope of the study and some of the technical elements to be taken into account.

For instance, the OECD Programme for International Student Assessment (PISA) was originally conceived as an effort to measure the skills of the population of 15 years of age.[10] Since in most OECD countries this population is almost universally enrolled in secondary education programmes, a good strategy to reach the target population was to conduct a test in the secondary schools, even if this test was not intended to measure curriculum-based elements. Actually, this strategy led to excluding from the study's universe a small percentage of the population

[9] One such example is a recent literacy survey (Bangladesh Bureau of Statistics 2008) that reported results on one scale with four levels (non-literate, semi-literate, literate-initial and literate-advanced). The scale merged information from four domains: (i) reading (oral reading of five isolated words and a passage made up of short and simple sentences); (ii) writing (five exercises writing isolated words); (iii) numeracy (12 purely algorithmic exercises, one subtraction embedded in a passage and one simple series) and (iv) general knowledge (nine visual/oral exercises not requiring any ability to read, write or compute). Each section accounted for one quarter of the overall score; at the same time, each level represented one fourth in the range of possible scores. Thus, if for instance an individual got all the points in the "general knowledge" section and only one more in any of the others (a total score of 26/100), that individual would be classified as semi-literate, i.e. someone with the "ability to recognize and write some simple words, to count objects, and numbers at a very basic level" (op. cit., p. xiv).

[10] "PISA seeks to measure how well young adults, at age 15 and therefore approaching the end of compulsory schooling, are prepared to meet the challenges of today's knowledge" (OECD n.d.).

(approximately 3% in the OECD countries which conducted PISA's first round in 2000, except for Mexico) which is perfectly acceptable in any sample survey.

Nevertheless, in one OECD country (Mexico) and several non-OECD countries also participating in PISA, this strategy led to excluding a significant proportion of the 15-year-old population. In the case of Mexico, approximately 50% of the target population was excluded in this manner in 2000.

Given this situation, the original purpose (as still present in the previously quoted document) is usually rephrased as: "The Programme for International Student Assessment (PISA) is an internationally standardised assessment that was jointly developed by participating economies and administered to 15-year-olds *in schools*."[11]

In any case, a clear definition of the target population not only defines sampling strategies but, most importantly, establishes the limits of what can be inferred from a study. In the case above, the conclusions for Mexico refer to their 15-year-old population in secondary schools (half the total), while for most OECD countries this is almost identical to the overall 15-year-old population regardless of their schooling situation.

Another illustrative situation relates to the difference between assessing the literacy skills of the population and assessing the skills of those who are graduating from literacy programmes. While the former assessment usually intends to yield information on the national situation pertaining to literacy, the latter is (or should be) an integral part of efforts to appraise the effectiveness of specific educational programmes.

3. A sound implementation of sampling issues (design and actual implementation)

Since these assessments are usually based on a sample survey, the different elements involved in sampling (quality of the sampling frame, appropriateness of the sampling design – definition of stages, strata etc.; sound implementation etc.) should be treated with utmost care.

In order to make inferences about a population by using sample data, a probabilistic sample design is needed, and this is also a rather specialised matter which should be treated by taking into account the corresponding expert advice.

4. A sound test development process, including pre-testing

This is a critical issue, since tests are the actual tools that will generate the information on skills. Proper instrument development is required in order to ensure that reliable and specific information is obtained.

For instance, reading is a complex domain that comprises the ability to cope with texts of different natures (descriptive; narrative; explanatory; continuous, non-continuous; mixed texts) and degrees of complexity (more pieces of information, different grammatical structure, vocabulary, etc.) on which the reader can perform different operations (locate and retrieve information, integrate pieces of

[11] Quoted from the first paragraph of *What PISA is* in the OECD website for PISA, available at http://www.pisa.oecd.org/pages/0,3417,en_32252351_32235907_1_1_1_1_1,00.html. Italics added by the authors.

information, make inferences, etc.) In order to measure reading skills, a balanced combination of these different elements is required.

At the same time, the ability to read relies on some pre-reading abilities like linguistic competence or the ability to decode words.

Thus, measuring reading skills by using as the only evidence the capacity of the individual to read aloud a few isolated words would not work. In this way partial evidence on either decoding or word-sight recognition skills may be generated, but that says little about actual reading.

This topic also relates to the need for a minimum number of items in the test which would yield a reliable measure. For instance, if the ability to read is measured by using a single reading comprehension question based on a paragraph, and the respondent is relatively skilful but does not know the meaning of one of the critical words in the passage, the measurement will be affected and, lacking other questions, there would be no way to correct this problem. A minimum number of questions and stimuli is needed in order to obtain reliable answers and that is one reason why a "short test" comprising a single passage (let alone a single sentence) would not be enough. In addition, a single-passage test would be strongly affected by the type of text selected (descriptive, narrative, argumentative, etc.) and by the interference of the background knowledge some examinees will have on the specific subject matter.

Moreover, test development should consider not only a minimum number of questions to obtain reliable results, but also the way these questions are posed, including not only the techniques to be used (a paper and pencil test, using multiple choice questions, etc.) but also the context the questions refer to and the specificities of the languages and scripts.

What was already mentioned in relation to numeracy items and market-related situations equally applies to other domains like reading or writing: the familiarity of the context, the potential interference of prior knowledge, etc. are key elements to be taken into account for test design.

Nevertheless, even the most careful of designs cannot anticipate every potential problem and circumstance and for that reason, every sound survey design (and not only those including tests) requires a field test or trial intended to show potential problems with the tools (questionnaires, test) and procedures to be followed in the assessment. Even if every professional knows this perfectly, there are situations where some people suggest "trimming" out the assessment effort by suppressing a field trial; proceeding in this way presents major risks that can compromise the success of the whole endeavour.

5. An adequate verification of translations of the tests where applicable

While a test can be conducted in several languages and contexts, its development might have been done in relation to one specific language and context source. For that reason, source instruments might need to be translated and will always need to be revised in relation to the context, and adapted as needed.

This is particularly so when there is an interest in having results that are comparable across different language groups (within a country or across countries), but not only in these cases. A mechanical transposition of an instrument from one language or context to another might end up measuring something completely

different and, for that very reason, even if the instruments have proved to be reliable in other places, they have to be carefully adapted to ensure they measure what they are intended to measure in a meaningful way in the new context. This is another reason why a field test is always an imperative.

Once instruments have been translated and adapted, a process of careful verification needs to be conducted to ensure some basic properties are preserved; namely, the test items should measure the intended skill at the intended level of difficulty across all languages used.[12]

6. A reliable field operation

It goes without saying that all the good work conducted in the preparatory stages can be compromised if the field work is not conducted properly. Aspects like respondent selection or the careful administration of the instruments are of critical importance to ensure that the assessment effort will yield usable results. This is particularly so when conducting a test on a household survey platform. Experts in educational testing are used to conducting tests in a controlled environment (schools) with examinees who are fairly accustomed to taking tests (students). At the same time, household surveys often rely on practices (e.g. using one respondent to provide all answers on behalf of the household members) that might be at odds with the technical procedures required for an assessment.

Careful monitoring of key issues is essential for enabling those responsible for the assessment to take corrective actions if needed.

7. A proper handling of data: test scoring, sample weighting, data recording and cleaning

Once the field operations are completed, questionnaires and tests have to be captured and processed. In the case of tests, scoring is needed and there are a number of provisions that should be taken in order to ensure scoring does not become a source of bias or noise. The quality assurance mechanisms for scoring should include: (i) clear scorings rubrics that feature specific procedures for the most common situations and provide guidelines with general principles and criteria to handle unforeseen ones; (ii) the selective hiring and extensive training of scorers, as well as a close monitoring of their work; (iii) double scoring mechanisms implemented to ensure that the performance of each scorer can be assessed by comparing it to those of all the other scorers and (iv) reliability analyses (computation of inter-scorer agreement statistics such as exact agreement and Cohen's kappa[13]) for subsamples of cases in both the field test and the main assessment.

[12] See Dept et al. (2010).

[13] Inter-scorer agreement deals with the degree to which two scorers, working independently from each other, arrive at the same score for a given answer provided by a respondent. The exact agreement is simply the proportion of agreements between the two scorers, while Cohen's kappa (κ) coefficient is another, more sophisticated measure used for categorical items. It is generally thought to be a more robust measure than simple percent agreement calculation, since kappa takes into account the agreement that is expected to occur by chance.

Something similar applies to data capture: quality control mechanisms should be in place to identify and correct potential problems at this stage; double data capture is strongly recommended.

Since a sample is drawn out of a population, the data are subject to sampling error and should be treated as such. That is to say, sampling weights should be computed taking into consideration the original design, the actual sample obtained, non-response factors, etc. In addition, the sampling design will most likely not involve a simple random sample, but rather a more complex, multi-stage design (for instance, one that involves sampling districts, then city blocks, then households, and finally individuals) which combines stratification (e.g. rural and urban areas, high and low socio-economic status areas, etc., in order to ensure a representative sample and improve the precision of the estimates) and systematic sampling (e.g. every eighth household as the interviewer moves "clockwise" around a block; this facilitates the interviewers' and supervisors' work in the field). Therefore, most standard estimation techniques, developed for simple random sampling, will not be suitable. As a result, other methods would have to be applied, such as simulations and re-sampling (bootstrapping, jack-knifing, etc.).

8. A clear, appropriate and sound data analysis

Once the data are "clean" (i.e. free from inconsistencies and out-of-range values), the analyses can be conducted. There is a wide array of statistical and psychometric tools that can be used and their selection must be informed by both the expert advice and the reporting needs. For instance, test results can be presented simply as scores computed as the proportion of correct responses, or, preferably, by using more sophisticated techniques that take into account the fact that not all items are created equal.[14]

In the same fashion, since the goal is to characterise the performance of different subgroups, rather than report on individual scores, each examinee in the sample represents a given segment of the population in relation to both its typical behaviour and the heterogeneity of that group. Computing a unique precise score for each individual can help in representing the typical behaviour of the group but will not be a good representation of its diversity. For this reason, some techniques have been introduced which intend to yield better estimations of the variability of skills in the target population rather than of the exact skill level for each individual respondent.[15]

[14] Item Response Theory (IRT) models have been developed to compute scores by using a set of item characteristics such as the difficulty and discriminatory power of each individual item used in a test (in the two-parameter logistic model, or 2PL). In some cases, (especially multiple choice or true/false questions, as opposed to open-ended questions) an element of pseudo-guessing is also factored into the model (the three-parameter logistic model, or 3PL). Thus, each score is a mathematical function that combines the individual ability with the characteristics of the items included in a test on a specific scale.

[15] This is the major topic surrounding the discussion on the use of a set of "plausible values" for each individual respondent (see IERI 2009).

9. Proper reporting

Finally, the whole effort would not be justifiable if the results were not properly communicated in order to be used by those in charge of policy and programme design and implementation. This topic relates to both the technical soundness of the reported results and analyses as well as its intelligibility and capacity to address the relevant policy questions.

Measuring literacy skills

The realm of skills is diverse and comprises a complex array of phenomena. Therefore, measuring skills requires, as a first step, clearly defining those to be measured and taking into consideration their complex nature. For instance, even if there is no single definition of what literacy is about, there is a consensus that it comprises[16]: (i) the ability to read, (ii) the ability to write and (iii) the ability to perform computations (numeracy).[17] Therefore, literacy is not a one-dimensional concept but a complex construct which encompasses at least three different sets of abilities.

A second element to take into account is that skills usually exist over a continuum ranging from very low levels (when the individual might not be able to perform even the most simple tasks) to higher levels of complexity.

A third element to consider is that even if skills can be defined in abstract terms, they do not belong to a sort of abstract or immaterial world. Skills are instantiated in specific circumstances which may or may not be present in the everyday life of individuals, families and communities.

A fourth element refers to the previous point: skills are deployed in different circumstances like work, home, community life, etc. as well as in relation to people's intentional actions, and are therefore linked to their expectations, needs, objectives, etc. At the same time, this also entails that different contexts and different actions present different demands in terms of literacy skills. Thus, the interaction between individuals and their contexts will lead to strengthening or weakening previously developed literacy skills.

These crucial elements translate into some sine qua non requirements:

a. The measurement's scope should be clearly defined. For instance, it can attempt to appraise reading, writing, (written-based) numeracy skills, or a combination of these.
b. If the measurement comprises several different dimensions, these should be separately measured, analysed and reported.
c. The instruments should be capable of portraying each dimension in a continuous or at least ordinal scale.

[16] See for instance: Global Campaign for Education (2005) and UNESCO (2005).

[17] While by definition reading and writing always refer to written materials, this is not the case for numeracy. Computations can be performed in fully oral situations, or by simply relying on graphical resources. In that sense, it is possible to suggest that the only numeracy tasks included in this definition of literacy are those that require written responses and tend to provide written questions or stimuli, or both.

d. The instruments should be developed in relation to concrete circumstances that resemble real-life situations as much as possible. Measuring skills as abstract operations (as is the case in some tests used in schools) is usually problematic since memorisation and formalism affect the measurement effort: individuals may be able to solve an algorithmic problem like $2 + 2 = x$, but they are not necessarily able to realise the need to deploy that skill in a concrete situation.
e. Constructing measurement devices that consider "real-life" circumstances is particularly complex since there is no universal set of these which applies to every single individual. Utmost care should be given to this, otherwise the whole effort may be void, since test-development might become, even if unintentionally so, ethnocentric and therefore invalid.

If these topics are properly addressed, the measurement of literacy skills can provide very rich information overcoming at the very least these two major pitfalls embedded in (erroneously) taking literacy rates as a proxy measure for skills: (i) literacy rates are unidimensional measures while literacy is a multidimensional phenomenon, (ii) literacy rates are constructed as a dichotomy while literacy skills exist over different continua.

Political discourses on measuring skills

Having briefly touched upon the main elements that comprise a sound approach to measuring literacy skills, it is time to turn to some political debates about them, which affect the way literacy surveys are designed and conducted.

A first element we would like to stress refers to CONFINTEA VI's recommendation about treating literacy as a continuum in surveys and data generation. This call reflects an increasing interest in measuring literacy levels which, in turn, entails conducting specialised assessments.

A second element we will focus on refers to existing debates on the complexity entailed by the various elements mentioned above and how they might compromise the feasibility of conducting literacy assessments. In this context, it is usual to find discourses which basically advocate for "simpler, quicker and cheaper" approaches to literacy testing.[18]

This way of framing the debate does not pay attention to a basic principle in scientific enquiry: *parsimony*. In fact, one basic requirement for any approach to be sound is for it to be *fit for purpose,* which means it has to be as simple as possible, but not simpler than that, since it also has to be as complex as needed.[19] Thus, while this approach is not very common in academia, it does resonate in the world of

[18] The use of these three comparative adjectives in relation to testing literacy skills is present in one academic paper (Wagner 2003).

[19] This idea is present in the philosophy of science at least since William of Ockham (the expression *Occam's razor* exactly refers to the need to suppress unnecessary complexity). Albert Einstein (in his Herbert Spencer lecture at Oxford in 1933) felt it necessary to stress that simplification should not go so far as to compromise the whole effort (things should be simple but not simpler or oversimplified).

practice, where some pressures and constraints may call for something simpler, quicker and cheaper. However, for this kind of approach to be effective, it has to show that the others fail because of their unnecessary complexity. Thus, if we take into account the above-mentioned elements or standards, the question will have to be posed in the following terms: where can we apply Occam's razor? If the answer is that every single one of Postlethwaite's standards is needed, then there is no room for Occam's razor and a simpler option would not work.

This paper asserts that oversimplifying or simply overlooking any of the above elements will compromise the validity of a literacy assessment. Therefore, arguing that political demands are so urgent that some of these issues could be skipped in order to have a quick response can easily lead to wasting resources and producing misleading information. In summary, if a literacy assessment cannot be conducted in a sound manner, it would be better to do something else with the available resources.

At the same time, it is important to notice that educational testing is not only a scientific discipline but also a business. In that regard, it is important to take into consideration how the legitimate commercial interest of some consultants or companies can affect the design and characteristics of a given study. Potential concerns point to country ownership, the fostering of national capacities and sustainability. It is important to prevent situations where key elements of the technical process are not shared, thus creating a sort of impenetrable black box. In any case, these potential problems are easy to identify when directly approaching copyright issues and the level of participation of a country team in tasks pertaining to: (i) instrument development and (ii) analysis. While some countries might legitimately opt for having a contractor who will not share some critical information, this should be a fully conscious decision.

Another element is that educational testing is, as almost every single topic in education, an arena where competing political and ideological discourses are present. Ethnocentric approaches to measuring literacy skills might not be "mistakes" but actually the result of promoting a given ideological approach where, for instance, the mobilisation of literacy skills is something that only happens at individual level (home, work), but not at community or local level; or the prevalence of an official (usually colonial) language is undisputed.

Finally, there is another issue that relates to the previous discussion: what is the sort of information that is needed for literacy policy purposes?

Literacy is a complex array of phenomena, and measuring literacy skills, while being central to it, is just one piece in the puzzle. Measuring self-perceptions, social practices at local level, studying the relationship between the written usage of a given language and the power structures, and the literate environment, are other relevant pieces. If a given country is determined to face the literacy challenges but, at the same time, it lacks the resources to conduct a sound literacy assessment, there are many other things it can do to inform policy. While literacy skills assessment is extremely important, there is also a risk of its becoming a meaningless fad. Again, the first critical questions are about identifying what a country needs and how the data will be used. Then, the feasibility of conducting a given study is to be appraised and decisions made.

Lessons from the LAMP experience

In 2003, the UNESCO Institute for Statistics (UIS) launched its Literacy Assessment and Monitoring Programme (LAMP). For 7 years, the UIS team has worked with a group of countries from different regions of the world in order to validate LAMP and have a tool that countries can use to conduct sound literacy assessments.

LAMP is a programme aimed at: (i) developing a methodology to measure, in a cost-effective way, some key elements related to literacy, such as the literate environment, self-perceptions, socio-demographic characteristics of the respondents, reading and numeracy skills; (ii) work collaboratively with countries in order to ensure they develop both ownership of the approach and tools, as well as the technical capabilities to conduct the different stages of the process and (iii) generate sound evidence on literacy that can inform policy and programme design and implementation.

The LAMP design comprises the following set of tools: an enumeration area information sheet (which gathers locality level variables); a background questionnaire (for household and individual level variables); a set of reading and numeracy tests (with some 79 items in total)[20] and a set of Reading Components exercises (to measure pre-reading skills among low-skill respondents and to produce more detailed information useful for programme design).

As of mid-2010, LAMP tools have been pre-tested in eight countries.[21] These countries are currently developing the preparatory work to conduct the main assessment which started in October 2010. Five other countries have also initiated the implementation recently.[22]

Even if the number of countries may be regarded as small, it is important to note that this is the first international experience concerning youth and adult literacy comprising non-European languages.[23] Actually, LAMP has been tested in the following languages organised by language family:

- Afro-asiatic: Arabic, Hausa, Tamasheq
- Altaic: Mongolian
- Austro-Asiatic: Vietnamese
- Indo-European: Spanish, French
- Niger-Congo: Fulfulde
- Nilo-Saharan: Kanuri, Zarma

[20] These tests are distributed across different instruments, so every single respondent is exposed to a smaller number of items ranging from 35 to 49.

[21] El Salvador, Jordan, Niger, Morocco, Mongolia, Occupied Palestinian Territory, Paraguay and Vietnam.

[22] Anguilla, India, Jamaica, Laos and Namibia.

[23] While the OECD studies (International Adult Literacy Survey – IALS and Adult Literacy and Life Skills Survey – ALL) were conducted in more than 20 countries and 15 languages; these languages were all European (13 Indo-European and two Uralic) and all of them using the Roman alphabet and Western Arabic numerals.

This degree of linguistic diversity also entailed working in three scripts (Arabic, Cyrillic and Roman) and two numeral systems (Eastern and Western Arabic numerals).

If, in addition to this important degree of diversity, the differences across countries in relation to their institutional features, cultural characteristics and socio-economic development levels are taken into account, it is possible to realise the value that testing the approach, tools and procedures in these countries has had.

The past 7 years, therefore, represent a significant investment that is starting to bear fruit in relation to the objectives that explain why LAMP was created. Obviously, the process has not been free of mistakes as well as in-country issues that have created some delays, but most of these situations have provided opportunities for learning and improvement.

Actually, the importance of the above-mentioned elements following Postlethwaite has been clearly ratified by the LAMP experience. It is possible to highlight a few of those key elements that stem from both good practices and mistakes made while implementing LAMP:

- LAMP started out by relying on the previous OECD experience and some time was needed in order to show the importance of ensuring proper translation and adaptation guidelines. The fact that OECD studies can be considered the best starting point for LAMP did not mean that working with a completely different set of languages and cultures could be successfully done just by enforcing rigid guidelines more intended to preserve the original sources than to ensure the adapted tests were able to yield meaningful and valid measures.
- The previous OECD studies were also conducted in a given institutional context according to a given policy agenda. While this agenda has been criticised elsewhere,[24] it is evident that UNESCO has a different agenda, as do the countries where LAMP has been operating these years. A simplistic attempt to "just replicate" IALS/ALL by extrapolating it to a different context would not have succeeded, since it would have been both ethnocentric (and therefore scientifically and ethically invalid) and meaningless.
- The composition of the national teams is of extreme importance to ensure both the scientific integrity of the study as well as country ownership and relevance. Country teams need to combine the best existing expertise in the country (educational testing experts; adult education and literacy experts; experts in reading acquisition in the assessment languages; household survey experts) with the leading role of those who are interested in the information and are going to make use of it.
- Even if LAMP is led by a statistical body (the UIS), it is really valuable to be open to contributions from other forms of scientific enquiry. Thus the UIS has been encouraging countries to conduct ethnographic studies in parallel with LAMP implementation so as to better report on cultural specificities and attributes which are part of the literacy phenomenon and which would enrich the

[24] For instance Darville (1999), Hamilton (2001) and Hamilton and Barton (2000).

analysis and use of the statistical evidence. Two countries (El Salvador and Paraguay) conducted this sort of study in parallel with the field test, and a similar effort was conducted in Mongolia during the main assessment in October–December 2010.

- In order to provide the best possible support to countries and continuously improve LAMP conceptual and methodological elements, the UIS has been working with individuals and institutions from different continents, institutional affiliations and theoretical and professional perspectives. Even if the UIS team is responsible for ensuring that all these contributions are integrated in a coherent approach, it is important not to advocate or espouse a single perspective attached to only one partner or contractor.
- Therefore, the UIS team has had to be strengthened, not only to cope with the challenges that stem from the diversity of situations and the rich input mobilised, but also to ensure that each instance of LAMP implementation follows a minimum set of standards and, therefore, yields useful information.

The first round of LAMP field tests also showed the need to introduce several changes in the original approach: (i) improving translation and adaptation guidelines; (ii) shortening some sections of the assessment tools without compromising reliability; (iii) simplifying the administration by eliminating some instruments; (iv) streamlining the design of the background questionnaire; (v) modifying some sections in the assessment tools that clearly posed conceptual or operational problems; (vi) including specific tools to record information on the local environment; (vii) appraising different approaches to the analysis tasks, etc. While most of these changes have made LAMP "simpler", a few introduced additional complexity. In any case, those modifications have been introduced with utmost care to ensure parsimonious solutions in every circumstance.

Final remarks

Assessing literacy skills is, as any other area of scientific enquiry, something that requires complying with some basic standards that ensure a sound measurement endeavour that would be able to yield useful information.

Suggesting that these standards should be relaxed or simply overlooked compromises the validity of the endeavour and should therefore be avoided from the point of view of a serious commitment to the adult literacy agenda.

While simple solutions are always welcome, simplistic approaches are not. Identifying the difference between these two is also a matter of expert judgement that requires some degree of complexity that is unavoidable.

LAMP is an attempt to come up with a meaningful and sound approach that is *fit for purpose*, no more and no less than that. The UIS is responsible not only for designing LAMP in the best possible way, but also for continuously developing it in order to look for improvements that in some cases will make it simpler but in others will make it more demanding depending on what is needed. LAMP is developed according to basic standards in the field of educational testing, but also according to

basic principles of parsimony which are embedded in the best traditions of the philosophy of science.

CONFINTEA VI sends several important messages regarding literacy and literacy measurement and it is extremely important that practitioners around the world would welcome sound approaches to literacy measurement leaving aside discourses that show some disrespect for their subject matter when stating that "not so scientific" solutions are required. It also calls for a more comprehensive view of the complex nature of literacy that precludes oversimplifications like equating the literacy challenges to the improvement in *literacy rates* as if reaching a given and arbitrary threshold would mean that illiteracy is "eradicated". Measuring literacy skills not only promises yielding better evidence, but also reinforces a deeper understanding on what literacy is about.

References

Bangladesh Bureau of Statistics. (2008). *Literacy Assessment Survey 2008*. Dhaka: Bangladesh Bureau of Statistics.

Cooper, B., & Dunne, M. (1998). Anyone for tennis? Social class differences in children's responses to national curriculum mathematics testing. *The Sociological Review, 46*(1), 115–148.

Darville, R. (1999). Knowledges of adult literacy: Surveying for competitiveness. *International Journal of Educational Development, 19*(4–5), 273–285.

Dept, S., Ferrari, A., & Wäyrynen, L. (2010). Developments in translation verification procedures in three multilingual assessments: A plea for an integrated translation and adaptation monitoring tool. In J. Harkness, M. Braun, B. Edwards, T. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennell, & T. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts.* Wiley series in Survey Methodology. New Jersey: Wiley.

Global Campaign for Education. (2005). *Writing the wrongs: International benchmarks on adult literacy*. London, Johannesburg: Global Campaign for Education and ActionAid International.

Hamilton, M. (2001). Privileged literacies: Policy, institutional process and the life of the IALS. *Language and Education, 15*(2&3), 178–196.

Hamilton, M., & Barton, D. (2000). The International Adult Literacy Survey: What does it really measure? *International Review of Education, 46*(5), 377–389.

IERI (IEA-ETS Research Institute). (2009). *Issues and methodologies in large-scale assessments*. IERI Monograph Series Vol. 2. Hamburg: IEA-ETS Research Institute, IERI.

Mahony, P., & Hextall, I. (2000). *Reconstructing teaching: Standards, performance and accountability*. Falmer: Routledge.

OECD (n.d.). *PISA—The OECD programme for international student assessment*. PISA Brochure. Paris: OECD.

Postlethwaite, T. N. (2004). *Monitoring educational achievement*. Fundamentals of Educational Planning series. Paris: UNESCO International Institute for Educational Planning.

Power, M. (1997). *The audit society rituals of verification*. Oxford: Oxford University Press.

Revell, P. (2005). *The professionals: better teachers better schools*. Stoke on Trent: Trenthan Books.

UIL (UNESCO Institute for Lifelong Learning) (2010). *Belém Framework for Action. Harnessing the power and potential of adult learning and education for a viable future*. Hamburg: UNESCO Institute for Lifelong Learning.

UIS (UNESCO Institute for Statistics) (2009). *The next generation of literacy statistics: Implementing the Literacy Assessment and Monitoring Programme (LAMP)*. Montreal: UNESCO Institute for Statistics.

UIS (UNESCO Institute for Statistics) (2010). *Global Education Digest*. Montreal: UNESCO Institute for Statistics.

UNESCO (1958). Recommendation concerning the International Standardization of Educational Statistics. In *Records of the general conference*. Tenth Session. Paris: UNESCO.
UNESCO (1978). Revised Recommendation concerning the International Standardization of Educational Statistics. In *Records of the general conference*. Twentieth Session. Paris: UNESCO.
UNESCO (2004). *The plurality of literacy and its implications for policies and programmes*. Paris: UNESCO.
UNESCO (2005). *Aspects of literacy assessment. Topics and issues from the UNESCO expert meeting, Paris, 10–12 June 2003*. Paris: UNESCO.
UNSD (United Nations Statistical Division) (1997). *Principles and recommendations for population and housing censuses*. New York: United Nations Statistical Division.
UNSD (United Nations Statistical Division) (2008) *Principles and recommendations for population and housing censuses*. Revision 2. New York: United Nations Statistical Division.
Wagner, D. (2003). Smaller, quicker, cheaper: Alternative strategies for literacy assessment in the UN Literacy Decade. *International Journal of Educational Research, 39*, 293–309.

The authors

Cesar Guadalupe B.A. in Sociology (Pontificia Universidad Católica del Perú); M.A. Social and Political Thought (University of Sussex, UK); Ed.D. (University of Sussex, UK).

Manuel Cardoso B.A. in Sociology (Universidad de la República, Uruguay); Ed.M. (Harvard University, USA).