GAML5/REF/4.4.2-02

United Nations Educational, Scientific and Cultural Organization

UNESCO INSTITUTE FOR STATISTICS

GLOBAL ALLIANCE TO MONITOR LEARNING

SUSTAINABLE DEVELOPMENT GOALS

Recommendations on Assessment Tools for Monitoring Digital Literacy within UNESCO DLGF

Maart Laanpere, Tallinn University, Centre for Educational Technology

GAML Fifth Meeting
17-18 October 2018
Hamburg, Germany

Introduction

This report summarizes a desk research project that aims to advise UIS in designing an instrument for assessment of digital literacy skills in the context of collecting data on SDG indicator 4.4.2. The SDG Target 4.4 declares that "*By 2030, substantially increase the number of youth and adults who have relevant skills, including technical and vocational skills, for employment, decent jobs and entrepreneurship*" and it contains three indicators (UIS, 2017):

- *4.4.1 Proportion of youth and adults with information and communications technology (ICT) skills, by type of skill*
- *4.4.2 Percentage of youth/adults who have achieved at least a minimum level of proficiency in digital literacy skills*
- *4.4.3 Youth/adult educational attainment rates by age group, economic activity status, levels of education and programme orientation*

UNESCO Institute of Statistics (UIS) is responsible for development and validation of new methodologies for indicators under SDG Target 4.4. While indicators 4.4.1 and 4.4.3 have been already implemented in reporting for 2017, the status of indicator 4.4.2 is still "*under development*" (UIS, 2017). Although many countries have been collecting statistical data on digital skills or ICT literacy of their citizens for various purposes, there is no common agreements on what constitutes a "minimum or basic level" of proficiency in digital literacy that would allow aggregation of national data on the global level. As a result, there exists a serious knowledge gap about the global state of digital literacy skills of youth and adults, while these skills play increasingly important role in achieving SDG.

As a major milestone in the process of developing UIS's framework for digital literacy, UIS commissioned a report "A Global Framework of Reference on Digital Literacy Skills for Indicator 4.4.2" (Law et al, 2018). This recent report reviews digital literacy assessment frameworks used in 47 countries and summarizes consultations with a number of experts, resulting with suggestion to use European DigComp framework as the foundation for UIS Digital Literacy Global Framework (DLGF), while expanding it by five additional competences and adding two competence areas. The report raises three challenges. First, the need for mapping existing instruments for digital skills assessment to DLGF, pointing out that " *… there is not a one-size-fits-all assessment of digital competence that can serve all purposes and contexts*". Second, it also calls for cost-effective cross-national R&D programmes to develop and validate "*context-sensitive and fit-for-purpose digital literacy indicators and assessment instruments*". Third, the report points out the discrepancy between the proficiency levels and related measurement scales of SDG indicator versus DigComp. While SDG Indicator 4.4.2 focuses on "*minimum level of proficiency*", DigComp discriminates eight proficiency levels.

There exist some supra-national initiatives or services in the field of digital competence assessment, but those have focused on international research projects involving several countries (e.g. International Computer and Information Literacy Study, ICILS) or professional certification in specific occupational fields (e.g. International Computer Driving Licence, ICDL). Such supra-national initiatives could definitely inform UIS in designing a global instrument for collecting reliable and valid data on SDG digital literacy target, but none of these practices was specifically designed to inform SDG Indicator 4.4.2.

Another relevant category of digital literacy assessments are supra-national frameworks and policy indicators for measuring digital skills. The closest to DLGF is a new standard on digital competence framework for European citizens (DigComp), which has already been used for various purposes in several European countries (Carretero et al., 2017). European agencies DG Connect and EuroStat have already used DigComp to redesign their Digital Skills Indicator (DSI) in 2015. The DSI survey is carried out on a sample that is representative to general population of a country, asking respondents about digital activities carried out within the previous three months and assuming that "*persons having realised certain activities have the corresponding skills*" (European Commission, 2016). The DSI instrument defines three levels of proficiency: below basic, basic, and above basic level. However, there is no common European instrument for knowledge- or performance-based assessment of digital competence of citizens based on DigComp.

These three challenges raised by Law et al. are addressed by current desk research exercise that has three objectives:

- mapping existing digital literacy assessment to DLGF
- evaluating advantages and disadvantages of selected assessments that cover a large part of the DLGF, with emphasis on their cost-effectiveness for rollout on population scale
- recommending the next steps on developing an assessment tool suitable for SDG indicator 4.4.2.

Mapping existing National and Cross-National Assessments of ICT and Digital Literacy Skills against the Competencies of DLGF

Definitions of Construct

Digital literacy is a relatively new concept that entered to the semantic space that was already partly occupied by competing concepts such as ICT, media, information and computer literacy (or competence). Ferrari (2013) was among the first authors who tried to settle the relationship between these existing labels and newcomer (digital literacy/competence) in a similar manner with the definition suggested by Law et al (2018): "*Digital literacy is the ability to access, manage, understand, integrate, communicate, evaluate and create information safely and appropriately through digital technologies for employment, decent jobs and entrepreneurship. It includes competences that are variously referred to as computer literacy, ICT literacy, information literacy and media literacy*".

This definition builds on previous practices by incorporating vocabulary from predecessors (e.g. from information, media, ICT literacy frameworks), resulting with a list of 26 competences grouped into seven competence areas. As experience with European DigComp has demonstrated, such competence framework can be used for various pragmatic purposes: re-designing the outdated curricula and professional development programs, developing policy indicators, professional accreditation, recruitment and (to a lesser extent) research. Such "pragmatic" competence models are usually created by a panel of experts using top-down design approach, resulting with relatively large sets of competences and 5-8 proficiency levels. These models suit well for the purpose of curriculum design or developing diagnostic self-assessment tool. However, it is quite challenging to come up with

a scientifically reliable and valid performance-based test that measures 22 competences of DigComp on 8 proficiency levels (where each competence is a separate construct to be measured independently).

As an alternative to pragmatic approach, recent **psychometric** approaches to measuring digital literacy have been guided by Multidimensional Item Response Theory (MIRT) that understands Computer & Information Literacy (ICILS 2013) or Digital Information Literacy (Sparks et al, 2016) as a single latent trait that cannot be directly observed in test situation and, thus, should be inferred indirectly through the statistical analysis of the test results. Like any mathematical model, MIRT has some assumptions that need to be fulfilled in order to make valid inferences on the basis of test results. For instance, monotonicity assumption requires that the instrument does not make knowledgeable persons more likely to participate in the test (Chenery et al., 1988). Assumption of local independence means that performance in one item in a test does not influence performance in other items. While such assumptions are relatively easier to guarantee in case of knowledge-based multiple-choice tests, the same might be quite difficult in case of authentic performance-based assessment. Psychometric tests are often created in a bottom-up manner, by experimenting with a large set of test items and grouping these on the basis of factor analysis performed on pilot test results. Such assessment instruments (for instance, ICILS and iSkills) are usually focusing on a single construct, which might be multidimensional (having several aspects).

Two approaches to digital literacy assessment that were described above illustrate a struggle between internal and external validity in the context of educational assessment. Validity in general is understood as degree to which test results can be interpreted and used according to the stated purposes of the assessment (AERA et al., 2014). Internal validity refers to methodological correctness/coherence of a research instrument, while external validity can be interpreted as its re-usability through the relevance or usefulness for a wider audience. Pragmatic approach to defining and measuring digital literacy tends to result with poorer internal validity, but higher external validity of the assessment instrument, as it is better understood and accepted by various stakeholders (most of whom may not have background in mathematical statistics or psychometry). On the other hand, psychometric approach guarantees higher internal validity quite often in expense of reduced external validity.

Law et al (2018) recommend to use Pathway Mapping methodology for operationalising DLGF, focusing rather on users' perception of digital literacy in various contexts and concerning with external validity of assessment. Eventually, the digital literacy assessment based on DLGF will have to address the challenge of balancing internal and external validity, both through the methodological considerations and design of the digital literacy assessment instrument.

Framework for Analysing existing Digital Literacy Assessments

In order to decide about the methodology for selecting existing DL assessments and mapping these to DLGF, I reviewed some previous studies that had similar goals. Carretero, Vuorikari and Punie (2017) have reviewed 22 existing instruments that are used for assessment of digital competence in line with DigComp framework in various European countries. They grouped these instruments into three major categories based on the data collection approach:

- **performance assessment**, where individuals are monitored by human observer or software while being engaged in solving authentic, real-life problems by using common software tools (e.g. browser, word processor, spreadsheet)
- **knowledge-based assessment**, where individuals are responding to carefully designed test items that measure both declarative and procedural knowledge
- **self-assessment**, where individuals are asked to evaluate their knowledge and skills with questionnaires that might range from structured scales to free-form reflection.

These approaches can be strengthened by combining them and also by conducting secondary data-gathering (e.g. by providing an e-portfolio that contains creative works, certificates and other documentary evidences). However, it would definitely increase the cost and duration of assessment, while decreasing its scalability. Carretero et al did not exactly map the selected instruments to DigComp model on the detailed granularity level (to competences and proficiency levels), but their work guided me to some relevant assessments that have not yet been published.

An alternative set of digital literacy assessments and their analysis framework is suggested by Sparks et al (2016), who consider both the focus of assessment and item design:

1. Assessments having **information literacy** focus with multiple-choice and constructed response items (such as ILT, SAILS, RSA, ISS, ILAAP, ICILS)
2. Assessments having **technology literacy** focus with multiple-choice and constructed response items (such as IC3, ICDL/ECDL)
3. Assessments having **digital information literacy** focus with performance-based tasks (such as CLA+, PIAAC PS-TRE, iSkills)

Sparks et al (2016) are discussing also variations in the design of digital literacy assessment instruments caused by different **purpose** of the test. Some assessments are designed for research purposes, while the others support accreditation or institutional quality assurance and curriculum design. Yet another purpose is self-assessment to support professional development, where the instrument has been designed to meet the needs of test-taker as the primary user.

Yet another alternative typology of digital literacy assessments is presented by van Deursen et al (2017), who are mainly concerned with differences in item design:

Type 1: Self-reporting surveys with questions that ask for the frequency of technology use, which are assumed to deliver indirect evidence for the command of skills. When an individual uses an application that is considered to be difficult to use (or uses a large variety of applications), this is held to be an indication of a high level of skills. An example of this kind of assessment is EuroStat's Digital Skills Index.

Type 2: Surveys with questions that request self-assessments of skills on a predefined scale. This is the most commonly used method, some examples are: Digital Skills Wheel by DigitalDannelse, MDS questionnaire by van Deursen.

Type 3: Performance tests in a laboratory or other controlled environments that provide subjects with particular assignments to observe their command of Internet skills, for instance: PIAAC PS-TRE test.

The most thorough and relevant approach for analysing digital literacy assessment tools has been applied by Siddiq et al (2016), they have addressed the differences between ICT literacy assessments regarding reliability, validity, match to DigComp (on the level of single competence) and interaction types used in the item design. The analysis criteria used by Siddiq were:

- Country of implementation
- Year of implementation
- Sample size in empirical study
- Target group: primary, secondary or high school (no interest in adults)
- Item interaction type: multiple choice, interactive, authentic
- Framework used for competence modelling
- Availability of test items or their descriptions
- Match with adapted DigComp model
- Data: qualitative, quantitative or mixed
- Duration of the test (in minutes)

The current study addresses the three categories of instruments for digital literacy assessment, as described by Carretero et al (2016) to identify the existing practices and evaluate their applicability in the context of data collection for SDG indicator 4.4.2. The applicability analysis is mainly inspired by Siddiq et al (2016) and is focusing on relevance (purpose, match with DLGF, cost-effectiveness) of the given instrument, but also considers its reliability and validity, following the discussion above. The existing digital literacy assessment practices and instruments were searched from three types of sources:

1. scientific research publications,
2. policy documents in education and employment domains,
3. professional certification frameworks and related technical documents.

As a result of this search, 44 assessments were selected for closer review: 30 from Siddiq's sample (leaving out the assessments that were older than 10 years) and 14 additional ones, see the Annex 1.

In addition to indicators used by Siddiq (see the list above), two additional criteria were applied in analysis (inspired by Sparks et al):

- the purpose of assessment:
    - **research**, mostly international comparative studies
    - **credentialing** the digital skills, mostly for employment and qualification

    - educational **statistics**, mostly for policy planning/evaluation or quality assurance of educational system
    - **diagnostic** self-assessment, mainly for personal use.
- the focus of assessment:
    - **technical skills**: manipulation of common software applications, e.g. word processor, spreadsheets, presentations, internet browser, e-mail client
    - **information literacy**: a single construct that has several facets or dimensions related to skills to find, evaluate, make use of information
    - **digital competence**: a set of multiple (20+) competencies that represent various capabilities of a person to solve problems in authentic context using digital technologies.

Overview of existing assessments and their mapping to DGLF

The current analysis involved 44 assessments of DL, 7 of those being designed for and implemented on a global or multinational level. The oldest instruments date back to 2006, while the most recent ones have been developed in 2017. The main body of cases (30) was taken from Siddiq's paper that focused on digital literacy testing only in primary and secondary schools. Additional 14 cases represent mostly the DL tests that target wider population: two instruments target specifically university students, five are designed for adult employees, two have been implemented both in schools and with adult respondents. While a few of the assessments in our selection were implemented on relatively small samples (50-60 respondents), majority of the instruments have been tested with thousands of users. My selection of DL assessments is composed so that it includes various alternative designs due to different purpose (R - research, S - statistics, C - credentialing, D - diagnostic self-assessment) or focus (T - technical skills, I - information literacy, D - digital competence). There are also variations in item types (MC - multiple choice items, INT - interactive/dynamic items, AUHT - authentic performance) and duration of the test, ranging from 20 to 180 minutes. Most of the instruments are designed in accordance with a specific digital/information literacy framework or national curriculum (in case of school-oriented tests). In 18 cases out of 44, the set of test items have been explicitly available for analysis and mapping with DLGF framework, while in the remaining cases I had to rely on information about the specific competence within a framework/curriculum and provided examples of tasks.

The method of mapping the specific assessment to DLGF involved comparing the test items (or underlying assessment framework) with DLGF competencies and judging about potential match. Proficiency levels were not included in analysis, only content of the description of a competency. As a result (see Annex 2), I can conclude that the best match to DLGF was found in these existing assessments:

- DigComp level test for 9th and 12th grade in Estonian schools: it covers the whole DigComp framework, has both self-reporting and knowledge based test component, but reliability of items related to competence areas 3-5 are weak, does not cover DLGF domains 0 and 6.

- PIX test in France: extremely promising solution (both instrument and open-source software), advanced platform and item design (incl. adaptive testing), does not cover competence areas 5-6 in DLGF
- Digital Competence Wheel in Denmark: a good example of effective and user-friendly instrument, covers most of the DLGF competence areas, attractive visual feedback; disadvantage: no knowledge-based items, only self-reporting.
- MDS by van Deursen et al (NL and UK): reliable, valid and effective instrument that covers most of the DLGF competencies; disadvantage: no knowledge-based items, only self-reporting.

Country	Year	Sample	Target	Item type	Duration	Focus	Purpose
Belgium	2012	650	school	INT	100	D	R
Australia	2008	10926	school	MC/INT	120	D	C
Australia	2011	11023	school	MC/INT	120	D	C
Korea		17547	school	MC/INT	50	D	R
Korea		143	school	MC/INT	50	T	R
UK	2006	172225	school	MC/INT	100	D	C
Italy	2010	1056	school	MC	60	D	R
Lithuania	2017	35600	school	MC/INT	60	D	C
Korea		40072	school		40	I	C
Hong Kong		199	school	MC		D	R
Chile	2009	1185	school	MC/INT		D	R
Israel	2007	58	school	AUTH		I	R
Germany	2009	315	school	MC/INT	120	I	R
Norway	2012	3335	school	MC		D	R
Norway	2013	4216	school	MC		I	R
Norway		1793	school	MC		I	R
Norway	2010	4087	school	MC/INT	70	I	R
USA	2010	5884	school	MC/INT	30	D	R
USA	2005	1016	school	INT	75	I	C
Korea	2011	15558	school	MC/INT	45	T	C
Korea	2011	11767	school			T	C
Hungary		1470	school	MC	60	T	R
Hong Kong		1622	school	INT	90	I	R
Germany	2010	855	school	MC		I	R
USA	2008	350	school	MC/INT		I	R
Hungary	2013	60	school	INT	45	I	R
Netherl.	2011	54	school	AUTH		I	R
Russia	2011	398100	school	INT	120	I	R
global	2013	60000	school	INT	60	D	R
USA	2011	100	school	INT	45	D	R
NL & UK	2013	1017	any	MC	25	D	R
Finland	2014	3159	school	MC	60	T	C/D
Estonia	2017	1450	school	MC/INT	45	D	C
EU - DSI	2015		adults	MC	20	D	S
ECDL - global			adults	MC		T	C
IC3 - global			adults	MC/INT	45	T	C
USA - ILT	2014		adults	MC	60	I	C
USA - SAILS			adults	MC		I	C
USA - RRSA			university	MC		I	R
Canada	2016		university	MC		I	C
USA - CLA+			any	AUTH		D	C
PIAAC	2013		adults	INT		D	R
France - PIX	2017		any	MC/INT		D	D
Denmark	2017		any	MC		D	D

Analysis of Advantage and Disadvantage of Different Approaches to Assessment

It is likely that performance assessment and additional analysis of secondary (qualitative) data are not cost-effective, neither scalable approaches to digital skills assessment in the context of global assessment of digital literacy in the context of SDG. Self-assessment would be the easiest and most cost-effective to implement, but will likely suffer from low reliability and validity. However, it should be possible to combine self-assessment with knowledge-based or performance assessment. For instance, Põldoja et al. (2014) have designed and validated an instrument called DigiMina that combined self-assessment of teachers' digital competence with peer-assessment, knowledge-based tests and e-portfolio containing teacher's reflections and creative works. Within the DigCompEdu project, JRC tried to balance internal and external validity of assessing school's digital capability with the design of the SELFIE tool, so that schools were allowed to expand the scientifically validated core instrument with additional items from pre-designed, publicly available pool or even design their own additional items that seemed relevant to them (JRC, 2018). The future instrument that will be designed by UIS for digital literacy assessment might also benefit from similar balancing of needs for global standardisation (contributing to internal validity) and local context (contributing to external validity).

The purpose of DL assessment has a major impact on design of the instrument, its items and testing procedures. For instance, the instruments that are designed for research purposes apply mostly psychometric approach described in 1.1. It results with large efforts invested in item design and validation, sampling and data collection procedures, and analysis of results. The test items are mostly knowledge-based, combining multiple choice and interactive tasks, in a few occasions also simulations (PIAAC PS-TRE, OECD PISA). The data is collected in a short time frame, in controlled environment (group wise in computer labs). This type of DL assessments is costly, often funded directly by governments and repeated infrequently (once every 3-5 years). The DL assessments designed for statistical purposes (e.g. EuroStat DSI) are also applied on carefully composed sample, but the design of the instrument and items (mostly multiple-choice items) is significantly cheaper and allow frequent use, expecting low effort from respondents and analysts. Data collection for statistical purpose takes usually place in relatively wide and flexible time frame in loosely controlled settings. Compared with research-focused DL assessments, statistical data gathering is clearly less reliable and weaker regarding the validity (both construct validity and internal validity). DL assessments designed for credentialing purposes (e.g. ICDL/ECDL) address mainly the needs of employers or academic institutions who are increasingly more concerned about comparability of digital skills of their employees or students. This is why credentialing-focused assessments are designed for frequent and scalable use in highly controlled environment. High level of standardization and focus on universal, practical, technical skills (e.g. word processing, spreadsheets) increases reusability of items and instruments over time and across nations, thus reducing the cost that is left to be covered mainly by test-takers. Although the credentialing-type of DL assessment are usually highly reliable, their construct validity and internal validity is likely weaker than those of research-focused instruments. Diagnostic assessments of DL do not target high reliability and internal validity, as they prioritize external validity and perceived usefulness of assessment to the end users (test-takers). Diagnostic self-assessments (e.g. DigiMina, DigitalCompetenceWheel, MDS) are likely to include mainly self-reporting scales and sometimes also authentic tasks. Authentic assessment reduces the possibilities

for automatic scoring and requires involvement of human reviewers. This is why such assessments are costly and do not scale up, are mostly used on the level of a single organization. However, diagnostic self-assessment that uses only multiple-choice items (e.g. DigitalCompetenceWheel, MDS) is scalable and low-cost, suffering mainly for low reliability and biased sample, as persons with low self-esteem and poor digital skills are unlikely to take the test.

Looking from the perspective of UNESCO DLGF, it is possible to combine two or three types of DL assessment. As the aim of evaluating SDG Target 4.4.2 is to establish "*the minimum or basic level of proficiency in digital literacy*" among the population of Earth, the statistical purpose seems to be the main priority. Yet, the uptake, reliability and external validity of the future DLGF-based assessment can be increased by adding some elements of credentialing-focused approach or diagnostic self-assessment approach.

Recommendation on Assessment Tools for Monitoring Indicator 4.4.2

Based on the analysis above, five general recommendations are provided regarding the design and delivery of DLGF based Digital Literacy assessment:

- the core test of DLGF should be based on self-reporting, similar to Danish Digital Competence Wheel and Dutch MDS, responses on 3-5 point scale - none, basic, advanced; test duration should not exceed 15-20 min, automatic assessment for all items;

- test should be piloted with at least 1000 respondents in 3 different languages, different countries, to validate the items (CFA); an international steering group of researchers should be formed to evaluate and improve quality of items and procedures, also reliability and validity of assessment.

- to increase the validity and reliability, enhance the self-reporting instrument with additional knowledge-based test (similar to PIX, MC/INT items, adaptive); there's no need to cover the whole DLGF model with this additional test - only the competencies that were ranked higher by respondent; knowledge-based test items should be composed in line with Evidence Centered Assessment Design (ECD) approach, ready for MIRT analysis.

- software architecture for testing tool should be similar to PIX, with built-in data upload to UIS database (in processed, anonymized form); software and test items should follow annual versioning in Github; user interface has to be responsive, test runs also on smartphones and tablets; items and tests should be available for export in IMS QTI 2.2 format; special attention to meeting high national (and EU) standards for privacy and data protection: data is anonymized in early stage, no private information of respondents stored.

- software architecture (API, interfaces, protocols) should enable potential extensions for e-portfolios (for diagnostic self-assessment) and microcredentials (e.g. Open Badges).