ANALYSIS OF NATIONAL LARGE-SCALE LEARNING ASSESSMENT SYSTEM IN ARMENIA

Executive Summary

Yerevan, 2022

ANALYSIS OF NATIONAL LARGE-SCALE LEARNING ASSESSMENT SYSTEM IN ARMENIA

Executive Summary

The Analysis of the National Large-Scale Learning Assessment System in Armenia was commissioned by UNICEF within the framework of the Education Sector Plan Development Grant (ESPDG) to the Republic of Armenia from the Global Partnership for Education (GPE). The study was developed through close consultation with the Ministry of Education, Science, Culture, and Sports (MoESCS). The study was conducted by an expert team, peer reviewed by independent reviewers, and reviewed by staff/consultants from UNICEF and the Asian Development Bank (ADB), respectively the Grant Agent and the Coordinating Agency, selected by the Local Education Group (LEG), which was established by the MoESCS within the framework of partnership with the GPE.

Acknowledgements

First and foremost, the United Nations Children's Fund (UNICEF) would like to recognise the invaluable support of the Ministry of Education, Science, Culture, and Sports (MoESCS) in producing this report.

This report was prepared by a core team consisting of the following consultants (see Appendix 2): Hayk Daveyan (Team Leader), Ella Karagulyan (Senior Expert), Shushanna Manjikyan (Junior Expert), and Ani Asryan (Junior Expert). The work was carried out under the overall guidance of Alvard Poghosyan (Education Specialist, UNICEF) and Arman Gasparyan (Education Officer, UNICEF). Ramya Vivekanandan (Senior Education Specialist, Global Partnership for Education (GPE)) and Medjy Pierre-Louis (Learning Assessments Consultant, GPE) conducted the peer review of the report and carried out an orientation session to guide the team on application of the Analysis of National Learning Assessment Systems (ANLAS), which was initiated with the aim to support education sector planning of GPE partner countries and piloted in three GPE partner countries. Each of the three piloting countries contributed significantly to the consultative development process of ANLAS, allowing the testing of both the content of the ANLAS toolkit and the process of using these resources to analyse the national learning assessment system and to make recommendations for improvement to inform strategies within education sector plans.

The team is extremely grateful to Adrine Babloyan (Research and Evaluation Officer, UNICEF), Astghik Martirosyan (Child Rights Monitoring and Evaluation Specialist, UNICEF) for serving as peer reviewers and to Zoya Andekyan (Programme Assistant, UNICEF) for her support with administrative issues, as well as to Christian Garbis Adanalian, for proofreading the English version of the final report. The team would also like to thank Narek Manukyan (Consultant for Supporting Asian Development Bank (ADB) in its Role as Coordinating Agency in the Global Partnership for Education (GPE) Processes) for his comments and suggestions. The team also extends its gratitude to International Consultant Dr. David Rutkowski, for contributing to the study with his expertise in the field of educational assessment programs.

Finally, we would like to thank all ANLAS participants drawn from teachers, senior experts in education, teacher training providers, development partners, and the staff of the Assessment and Testing Center for providing invaluable information and perspective.

We apologise for any omissions and express our sincerest thanks to everyone, whether named here or not, who graciously gave their time and expertise.

Executive Summary

Purpose and Methodology of the Study

The study was conducted for the Analysis of Armenia's Large-Scale National Learning Assessment System (ANLAS). The Study was commissioned by UNICEF within the framework of the Education Sector Plan Development Grant (ESPDG) to the Republic of Armenia (RA) from the Global Partnership for Education (GPE). The Analysis of Armenia's Large-Scale National Learning Assessment System was developed through close consultation with the Ministry of Education, Science, Culture, and Sports (MoESCS).

The study assessed the current state of large-scale national educational assessment programs in Armenia and the lessons learned from the implementation of those programs. The **primary purpose** of the study in Armenia was to support the improvement of Armenia's national learning assessment system through comprehensive analysis and policy recommendations that are systemic and would support the implementation of fair and equitable assessment of all students. By aiming to contribute to the improvement of Armenia's national learning assessment system the study utilized qualitative analysis with a focus on three questions based on ANLAS three dimensions of quality assessment systems.

To achieve the purposes of this study and answer the study questions the ANLAS Manual was used. The ANLAS approach includes the following key dimensions:

1. Context of the assessment system (CN),
2. Quality of large-scale assessment and examination (QLE),
3. Coherence of the assessment system (CH).

Each dimension was evaluated based on dimensional key areas and corresponding quality objectives. The **first dimension** refers to the broader context in which the national learning assessment system is located and the extent to which this context is supportive of the assessment system. The key areas include: (i) legislation or policy, (ii) institutional arrangements and governance structures, (iii) funding, (iv) leadership[1]. The **second dimension** refers to the overarching key quality concepts for learning assessments: clarity and

consistency of purpose, fitness for purpose, technical rigor, objectivity and independence, transparency and accountability and ethicality and fairness. To provide meaningful data for education policy and practice, large-scale assessments and examinations must use well-founded methods. As such, the quality objectives derived for ANLAS focus on the aspects of each key area that are most essential for achieving technical rigor[2]. The **third dimension** is examining the extent to which the assessment system is consistent with important aspects of the broader education system and other aspects within the assessment system. The key areas include: (i) learning standards and curriculum, (ii) education system structure, (iii) education priorities, (iv) system-level data use[3].

Below is the list of assessment questions corresponding to each key area defined by ANLAS Manual.

1. To what extent is the context of the national learning assessment system supportive of large-scale assessment programs?
    - **a.** How is the assessment system guided by legislation or policy?
    - **b.** What institutional arrangements and governance structures exist for learning assessments?
    - **c.** Does the government provide sufficient and stable funding for the assessment system?
    - **d.** Does the government demonstrate leadership and political will in support of the assessment system?
2. To what extent does the quality of large-scale assessment programs implemented in Armenia fit the key quality concepts for learning assessments?
    - **a.** Are there quality assurance mechanisms in place to ensure that field operations are standardized, monitored and documented, so that the data are collected under the same conditions, independent of the administration context?
    - **b.** How does the implementing agency ensure that the final database is free from discrepancies and errors, appropriately structured and documented?
    - **c.** Are there quality assurance mechanisms in place to ensure the assessment instruments are reliable, valid and fair?
    - **d.** Does the implementing agency use scientific sampling methods, which assure appropriate and measurable levels of statistical

precision and validity in the interpretation of assessment results?

e Are there technically sound and appropriate data analysis techniques to provide analytical results that permit valid and useful inferences?

f. Do the responsible institutions develop and use appropriate products to reporting and dissemination tailored to the different stakeholders groups and promote appropriate and effective use of assessment data and results by those groups?

3. To what extent is the assessment system consistent with important aspects of the broader education system and other aspects within the assessment system?

a. Does the assessment system provide relevant data on students' knowledge and skills focus on the application of knowledge and demonstration of skills aligned with official learning standards and curriculum?

b. Does the assessment system provide relevant data on the state and progress of students' learning at key stages of primary and secondary school education, and for relevant levels of the education system?

c. Does the assessment system provide robust evidence on students' learning and the contexts in which learning takes place, to inform priorities in education policy and practice?

ANLAS in Armenia was conceptualized as a country-led, participative process implemented by a national team, led by a Team Leader, guided by a Steering Committee,[4] and followed a comprehensive two-stage approach. The two main methods for completing the qualitative analysis for ANLAS were document review and consultations with key stakeholders in the education system and in the assessment system.

The national team conducting the analysis consisted of four experts recruited through a competitive selection process by UNICEF to ensure their impartiality, avoid any conflict of interests. The team received training to adhere to the ethical and methodological standards of UNICEF in research, studies and evaluations (June 10, 2021) by Child Rights Monitoring and Evaluation unit in UNICEF Armenia. The team was also trained on the ANLAS Methodology prior to developing the implementation plan. The three-day workshop (June 16-18, 2021) covered ANLAS model, processes

and tools conducted by the study Team Leader.

The study included a document review of a list of documents identified through a mapping exercise and conducting a consequent document analysis against ANLAS quality criteria. A review of the documents was conducted in order to assess each key area against the defined quality objective. For each of the ANLAS dimensions a list of legal and policy frameworks related to overall national assessment programs, project-related documents, including the legal package of the establishment of particular assessment programs and their implementation was created.

The primary data for the analysis was collected through consultations with the stakeholders, identified during a mapping exercise, with the purpose to discuss the initial description of each key area in order to evaluate those against the defined quality objective. The stakeholder consultations were organized through Key Informant Interviews (KIIs), (ii) Expert Interviews (EIs) and (iii) Focus Group Discussions (FGDs). In total 29 stakeholders, including senior representatives from national and sub-national government units, senior representatives from external assessment agencies, subject-matter experts, representatives of higher education institutions (HEIs) and development partners were interviewed. Total of 12 Focus group discussions were conducted with 75 stakeholders including 59 teachers and 16 university applicants (students admitted to university) assuring diversity across regions and settlement type.

Analytical tables and instructions for completing them were used to document and guide the analysis of the three ANLAS focus areas. Within each focus area, several key areas were analyzed by using sets of guiding questions. For each key area, a quality objective was defined, against which the key area was described and evaluated. These quality objectives were indicated as part of the ANLAS findings. The analysis of the national learning assessment system consisted of three major steps:

1. Describing the key areas for each ANLAS focus area
2. Evaluating the key areas for each ANLAS focus area, identifying aspects for improvement and making recommendations
3. Synthesis of the findings

With the purpose to discuss the initial description of each key area

in order to assess those against the defined quality objective the secondary and primary data was utilized. Recommendations for improvement were identified. These recommendations aimed to inform the development and implementation of improvement strategies as part of the wider education sector planning process. It is expected that the results of the study will be primarily used by the Government of the RA, MoESCS and Assessment and Testing Center.

All evidence generating activities ensured respect for all persons involved in the study in accordance with UNICEF Procedure for Quality Assurance in Research, Policy on Personal Data Protection, UNICEF Procedure for Ethical Standards in Research, Evaluation, Data Collection and Analysis.

Internal quality assurance procedures were implemented to assure the compliance with research ethics and the principles outlined in the above mentioned procedures.

Findings and recommendations

The evaluation of the context of the assessment system (Focus Area 1) was guided by four key assessment questions which match respective key areas of ANLAS. Three of them were assessed as “partly achieved”, one as “achieved”.

© UNICEF Armenia/2021/Maggarian

With regard to legislation the data obtained from the analysis of stakeholder consultations and the document review revealed that the country has clear guidelines on organizing and implementing public examinations of students in compliance with the requirements of the state standard. However, there is lack of policies guiding the educational system in general and student assessment or teaching standards on the quality of instructional practices in particular. Thus, it was **recommended to stress the guidance of a large-scale learning assessment system in the Education Development State Program 2030**, which is

currently open for public debates and discussions. The guidance should address issues related to the mission, vision, functions and purposes of the assessment system and highlight the importance of the assessment system's coherence with national priorities and standards.

For the implementation of large-scale learning assessment programs a separate body, the Assessment and Testing Center (ATC) was established. ATC is responsible for the quality implementation of large-scale learning assessment programs from development of standardized tools (in case of national assessment programs) and adoption of tools of international studies (in case of TIMSS) to fieldwork and reporting. ATC is also responsible for the organization of centralized university admission procedures (including the competition) in Armenia. In general terms, issues which could essentially risk the functions of the assessment system are not observed, however, minor recommendations were suggested.

The Government provides stable funding for the implementation of both program and statutory obligations of ATC. According to the grant agreements and financial and programmatic annual reports of ATC, 100 percent of this funding is provided by the Government. Only centralized examinations are co-funded by the participants. However, there were some issues that were revealed during the consultations, such as, the amount of salaries allocated for technological professionals, shortage in the allocated budget for the implementation of TIMSS 2019 and minor financial restrictions which could affect the quality of assessment programs. Hereby, it could be concluded that stability of core funding is ensured by the Government, and ATC has no risks of failing its operations due to lack of funding, however, for **hiring and retention of more qualified staff, quality implementation of projects and full participation in international studies, additional funding or reallocation of the budget is required**.

According to the information obtained mainly from the interviews with senior representatives of the national government there was encouraging political commitment by leadership on promoting and supporting the assessment system. The importance of large-scale learning assessment programs was highlighted generally by all interviewed experts and officials. In spite of this, a very low level of trust towards the results of national, international learning assessment programs and school

graduation examinations is observed. In addition, the representatives of institutions and experts, which are supposed to be the primary consumers of the national and international assessment program results, are either not informed about the programs or do not have enough capacity to use their results in decision-making. Hereby, on the one hand there's a situation where the leadership is highlighting the importance of assessment programs, and on the other hand, the message is not constantly communicated with the main stakeholders (e.g. self-government bodies, research community etc.). The study recommended **considering making changes in the communication strategies, mutually voicing concerns and suggestions,** particularly setting-up recurring meetings with clear agenda related to the discussion of quality issues of the large-scale learning assessment programs implemented by ATC.

According to ANLAS guidelines adopted for Armenia, the quality of large-scale assessment (Focus Area 2) has six key areas, all of which were rated as partly achieved.

The results regarding the quality of large-scale assessment programs and examinations are significantly different from program to program.

In case of TIMSS, the implementing bodies needs to make minor improvements regarding the validity and compliance to state standards. Based on interview results it is reported that an essential part of the STEM subject domain starts not from 4th year, but from 5th year of studies. From the subject-matter expert's point of view, the program of 5th graders is more compliant with the TIMSS assessment framework, but due to the sampling approach, Armenia is not allowed to include 5th grade students in the study sample. Considering the current issue, it was recommended **developing a holistic report** (other than the matching table of curriculum content included in TIMSS documents) on the **compliance of country's education program with the TIMSS assessment framework** and only after making respective conclusions.

TIMSS assessments employ a two-stage random sample design, with a sample of schools drawn in the first stage and one or more intact classes of students selected from each of the sampled schools in the second stage. To ensure the consistency and uniformity of approach necessary for high-quality, internationally comparable data, TIMSS study in Armenia follows a set of standardized operations and procedures. The major steps of the

operations and procedures are similar from one assessment cycle to the next. Thus, Armenia is fulfilling all the requirements of TIMSS regarding the sampling approach and methodology. Once Armenia creates data files and submits them to the International Association for the Evaluation of Educational Achievement (IEA), a process of checking and editing known as “data cleaning” begins. Data cleaning is the process of checking data for inconsistencies and formatting the data to create a standardized output. TIMSS reports on the results are improving from year to year. The reports of 2019 are more user-friendly and present the information accompanied with charts and other visuals. This facilitates the consumption of information not only by professionals, but also by the broad public. Interactive released items embedded right on the website are making the essence of TIMSS more understandable. As a conclusion, it was clear from the study that there is a minor, non- essential need for improvement of TIMSS implementation in Armenia.

On the national level Armenia is locally developing and implementing large-scale learning assessment programs including (i) Armenian language, literature, and history (HAAS, started in 2010), (ii) physics and chemistry (BAAS, started in 2011), (iii) foreign languages (OLAS, started in 2013).

Validity of the tests of nationally developed large-scale learning assessment programs is assured based on expert opinion of item developers. According to the interviews and document review results there is no complete description of the assessment framework, which is a crucial reference point for validity evaluation. It poses risks for validity of tests, since item developers have no criteria to make sure the test represents the entire range of possible items it should cover. The study revealed that the decision about content validity of the test is based on (i) item developers’ opinion on assessment construct, (ii) research of the international practices in the field (mainly TIMSS and PISA), (iii) schoolbooks, (iv) items bank developed during the previous years. The reliability measures are included in internal reports on item quality. These measures are mainly examining internal consistency of different items within the same test. The data from these internal reports is actively used for making decisions regarding item quality. The study recommends **adopting practices of examining reliability for parallel or alternate forms and test-retest reliability** as the students included in

© UNICEF Armenia/2021/Galstryan

the assessment programs are being assessed by parallel test versions.

It was also clear from interviews, that the sample is being adjusted based on the resources available for the studies. In addition, the representativeness of the sample is also affected: the team tends to exclude schools with a small number of students. Thereby, given the current circumstances the sampling approach is not clear and implemented according to the scientific approach. It was recommended to **(i) review and finalize sampling methodology, (ii) develop a clear description of the sampling implementation, including general approach, target groups, cases of exclusion and participation rates, (iii) initiate budget adjustment accordingly**.

Data collection and the fieldwork are organized and implemented by the Examination Organization

Department of ATC. The department is responsible for overall organization of the fieldwork of all studies and examinations. To ensure the quality of the field operations the department is organizing hiring and training of administrators. In case of national large-scale learning assessment programs the department is utilizing the internal capacity of ATC. They receive guidelines for administering assessment tools. The guidelines are aimed to assure standardized conditions for collecting assessment data. As for the data safety, appropriate data storage procedures are in place to assure security and confidentiality of the test results, however, considering the technical staff turnover, it is recommend to **pay extra attention to developing detailed manuals for data safety and protection, management and processing**.

After each wave of national learning assessment programs a complete report of results is published including a secondary analysis of data. The analysis includes (i) descriptive statistics, (ii) correlations and (iii) regression models. Considering the fact that this study is implemented repeatedly it should also inform about the trends observed from year to year. However, there are two obstacles to this. First, there is no rigorous evidence about the validity of knowledge assessment tests and contextual surveys. Moreover, the assessment framework still needs to be thoroughly developed and described. Second, every year ATC has an obligation to implement one of the above-mentioned national assessment programs, but the principles of selecting a program are not clear. Thus, the databases from different cohorts are incompatible in terms of timing, tools and assessed constructs.

The study recommends **taking advantage of repetitive studies and getting back to trend analysis after building-up solid foundations for it**.

For national large-scale learning assessment programs in Armenia the only product is the final report of results accompanied with information on the ATC webpage about these programs separately. The results are being disseminated through the official channels of ATC (mainly webpage) and during the various meetings with stakeholders. In addition, all general education departments of regions are receiving the copy of the report. During the interviews it was revealed that databases of assessment program results were not available for users to download. The study **recommends enriching dissemination products**

by having the database of assessment results in an open access. It will both promote the dissemination of information about the programs and make the results accessible for the research community and decision-makers.

In case of examinations, content validity is defined as a representation of the entire range of possible items the test should cover. Test development is conducted under the supervision of an ATC subject-specific professional who supervises the overall process of test development while working with external experts. ATC is both coordinating and in some cases takes part in development based on the Director's decision. Item developers are using both the guidelines and item banks for constructing the tests. Items included in the bank are the ones included in the tests of previous years. They have respective quality indicators included in ATC internal technical documents such as test reliability, item difficulty and discrimination indicators. Item developers are allowed to make respective changes aiming to make them more relevant to the content domains. Thus, taking into consideration the commonly agreed knowledge domain assessment purpose of both examinations and external assessment programs, it could be concluded that the validity and reliability are mainly considered. Also the study team **recommends considering piloting opportunities before main examinations by taking into consideration confidentiality measures whenever necessary**.

Examination programs are census-based, which means they are administered to all students fitting the test criteria. For the grade and state graduation examinations 9th and 12th year students are obligated to take examinations. Stratified sampling approach is used by ATC for external learning assessment programs. First stratum of the sample is based on the administrative division of the country, the second is based on the type of settlement (capital, urban, rural). Issues related to the sampling approach are common for all nationally developed and implemented assessment programs. Thereby, given the current circumstances the sampling approach is not clear and implemented according to a scientific approach. It was recommended to **(i) review and finalize sampling methodology, (ii) develop a clear description of the sampling including general approach, target groups, cases of exclusion and participation rates, (iii) initiate budget adjustment accordingly**.

© UNICEF Armenia/2020/Grigoryan

Field operations of both examinations and external assessments are supervised and led by the Examination Organization Department of ATC. However, in some cases public schools are responsible for the administration. The main issue revealed during the interviews is the reliability of administration when the schools are responsible for it. Schools are inclined to help students during the test administration, trying to make them achieve scores. ATC does not have reliable mechanisms for controlling these actions.

The results of examinations are being reported during the 24 hours after the examination is finished. Those results are reported to individual students

after verification. ATC is generating descriptive analytical reports. After each wave of external assessment programs a complete report of results is published including a secondary analysis of data. The analysis includes (i) descriptive statistics, (ii) correlations and (iii) regression models, (iv) analysis of contextual data.

Annual dynamics analysis of the external assessment programs have issues identical to that of national assessment programs. First, there is no rigorous evidence about the compliance of the tools with each other. It means that tests used during different years are not compatible. The part of the assessment framework which presents learning context still needs to be thoroughly developed and described. Second, every year ATC has an obligation to implement one of the external assessment programs, but the selection principles are not clear. This makes databases from different cohorts incompatible in terms of timing and tools. The study recommends **taking advantage of repetitive studies and getting back to trend analysis after building-up solid foundations for it**.

The results of examinations are presented on the ATC webpage in the form of reports described above. For external assessment programs the only product is the final report of results accompanied with information on the ATC webpage. The results are being disseminated through the official channels of ATC (mainly webpage) and during various meetings with stakeholders. In addition all general education departments of regions are receiving the copy of the report. However, there is no holistic strategy for the results dissemination tailored to the target groups, specific channels and dissemination products properly designed based on the needs of those groups. Therefore, it is recommended to **develop an overarching dissemination strategy with its implementation plan**. With a more consistent and preliminary planned approach ATC will have an opportunity to receive feedback and increase the importance of its programs.

Alignment with official learning standards and curriculum, progress of students' learning at key stages of school education and evidence on learning contexts are the 3 key areas of coherence of large-scale assessment system.

The main area for improvement is the coherence of the assessment program with the learning standards and curriculum. To make this happen

clearly defined and commonly agreed assessment frameworks are required for each assessment program. In case of the international assessment program (TIMSS) compatibility to a country's national standards should be thoroughly investigated. In fact, there is some common understanding among the item developers and implementers of assessment programs regarding the assessment framework for each program. However, the common understanding on the expert level is not enough, it needs more evidence and justification. The shift from expert level to more evidence-based and justified understanding of assessment frameworks are of primary importance. The reason is the recent change in general education state standards. Compared to the current standard, the new standard adopts a more systemized approach to learning outcomes in the form of eight competencies.

From the data collected during the document review and stakeholder consultations it was concluded that the only reliable and technically sound data that can help make inferences regarding the progress of students comes from TIMSS. The reason for this is the absence of clearly defined, communicated and operationalized assessment frameworks and procedures. TIMSS has been implemented in Armenia starting from 2003. There is data available for 5 waves based on the study conducted both at the end of primary and secondary schools. Based on current study results, it is recommended to **adhere to TIMSS and continue conducting the study in Armenia including the population of 8th grade students**.

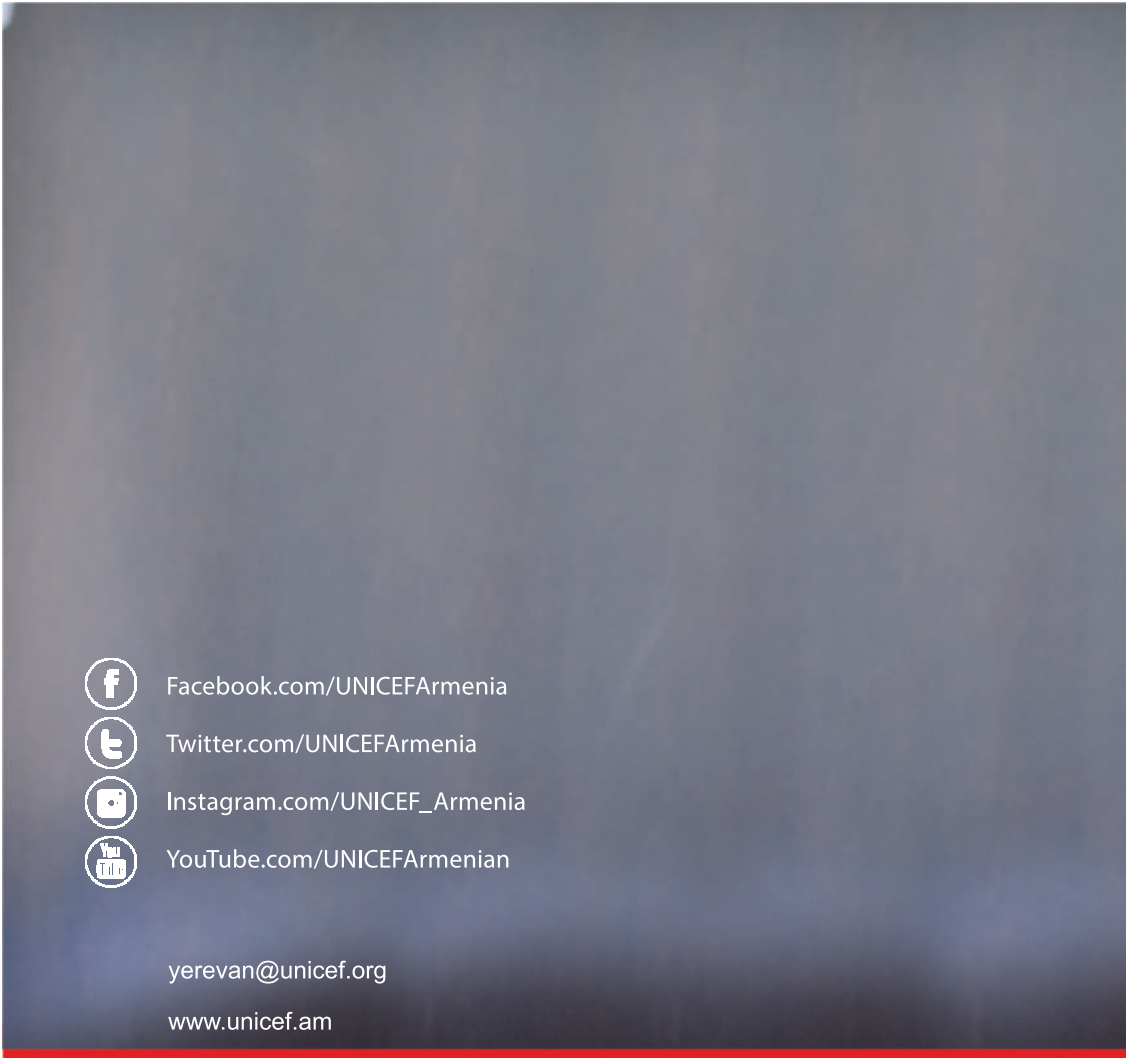
In national large-scale learning assessment programs the evidence on learning context traditionally remains within the framework of school location, gender of students and achievements by subjects. Nevertheless, it should be mentioned that in recent studies students' attitude, expectations and parental support were covered by surveys. Though national assessment programs are gradually adopting surveys related to the broader learning context due to absence of a clear assessment framework the data from national assessment programs is not reliable. In addition, other stakeholders (e.g. parents, teachers and school principals) are represented in the surveys.

Armenia has built up a large-scale learning assessment programs throughout the last 20 years. It has succeeded in building rigorous organizational structures and institutions, capable of handling

studies with 5000 and more participants. The assessment system has experience in working both with international and local partners. There is a system of database development and management in place. This ensures strong foundations for coherent implementation of programs.

However, the system needs to be boosted by stressing the quality of assessment programs, particularly the ones which are developed and implemented at the national level.

These programs should improve their quality by thoroughly following best scientific practices in validity and reliability evaluation, quality of field work and culture of reporting and dissemination. All these components are necessary to make sure that the programs have high level of trustworthiness due to their proved and clearly communicated quality. This will help increase the interest towards the results of the programs and make them a part of education policy.



Facebook.com/UNICEFArmenia

Twitter.com/UNICEFArmenia

Instagram.com/UNICEF_Armenia

YouTube.com/UNICEFArmenian

yerevan@unicef.org

www.unicef.am

unicef

for every child