United Nations
Educational, Scientific and
Cultural Organization

UNESCO
INSTITUTE
FOR
STATISTICS

SUSTAINABLE
DEVELOPMENT
GOALS

SDG4 Item Bank and Exchange Platform

Concept Note

May2019

GAML6/REF/6

1. BACKGROUND

The Sustainable Development Goals (SDGs) set out new ambitions for education, with SDG 4 requiring a quality education from pre-primary to upper secondary level of education for every child by 2030. The focus on quality education is a deeper and more demanding focus than the Millennium Development Goals (MDGs), since it puts more emphasis on learning outcomes.

SDG target 4.1 calls on all UN Member States to 'ensure that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes' by 2030. Specifically, indicator 4.1.1 measures 'Proportion of children and young people: (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex'.

Around the world, there are dozens of countries for which we do not know the levels of learning and the costs of assessing them are very high. There is also a need for better technical documentation to guide countries in producing proficiency statistics. The challenge is to find the most cost-efficient, fit-for-purpose way of doing this. A practical low-cost solution is an item-bank and test-scoring Platform that utilizes machine learning algorithms to lower costs and increase efficiency of large-scale assessments. Specifically, using algorithms developed for high-dimensional analysis of natural language, familiarly used to generate web search results, machine translations, and shopping recommendations, allows the Platform to automate many frequent tasks that would otherwise require manual input. Manual activity will be required where it is most useful, for one-time activities related to initial item classification and definition of performance level descriptors, while routine tasks, such as item response scoring, test form development, item response analysis, test score interpretation, standard setting, and test scale linking, may be largely automated or heavily machine-assisted. Recent evidence from machine learning research indicates that trained, cross-validated neural networks are comparable or superior to human judgment and traditional statistical models for both prediction and classification. The item bank focuses on the performance level descriptors and the exact cognitive requirements of each item, which, as the basis for scoring and data analysis, would facilitate expert-level automation.

The recent finalization of the Global Content Framework (GCF) for Reading and Mathematics provides a solution for indexing current and future assessments. The common framework simplifies the use of dimensional mapping algorithms to link assessment content, performance level descriptors, and student assessment results.

2. PROJECT DESCRIPTION

General objective

The purpose of the Item-bank Project is to create an objective and highly reliable reading and mathematics assessment scoring and reporting tool, is free and available to children all over the world.

The item bank and test platform would allow assembly and scoring of fit-for-purpose assessments for countries' own tracking and for reporting on SDG4 4.1.1 grades 2/3, the end of primary and end of lower secondary. The statistical focus of the UIS item bank is on the performance level descriptors

(PLDs) and the exact cognitive requirements of each item as the basis for scoring and data analysis (in contrast to the most common current practice, which uses the cognitive definitions to specify test design, but largely ignores them in subsequent data analysis and scoring).

Specific Objectives

The specific objectives are to build a Platform that facilitates three functions

1) ***Item Banking***
   a) Aligned with the Global Content Framework (GCF) already developed under UIS auspices
   b) Classification of items according to reporting on the Minimum Proficiency Level as required by SDG4.
   c) Defined mechanisms and protocols to share items and test data for academia and research centers.
   d) Plan and set up an exchange system whereby countries request items from UIS and add to the Bank of Items.

2) ***Test assembly, scoring and analysis with the rules that allow reporting on SDG4.1.1 in reading and math.***
   a) The test builder tool will assist designers and draw on items in the Platform using trained machine learning algorithms.
   b) Some prototypes of assessments, based on the actual process to be used for designing other assessments, would be available.

3) ***Linking country-specific test results to international PLDs and the GCF.***
   a) Indexing assessment content and PLDs against the common GCF allows the linear scales of independent assessments to be mapped to a common manifold.
   b) Results for specific assessments may be reported against expected placement and comparative performance on other large-scale assessments linked to the GCF.

Advantages and intended universe

The Platform will serve as a practical and low-cost solution to low- and lower-middle-income countries who are unable to participate in financially demanding international programmes yet want to implement their own national assessment and report data against the indicator. This will allow for country ownership and will democratize testing, since countries will make their own decisions, receiving guidance rather than instructions from the UIS, which will be acting in its capacity as a facilitator.

By targeting low- and lower-middle-income countries, such an approach would foster inclusive education systems by accounting for those large groups whose levels of learning are unknown. Even for countries that already have assessments, the availability of the tools would enable them to improve the quality of their assessments, by allowing them to benchmark their own assessment against one that embodies the best items and psychometrics available. Access to an item bank that

coordinates data collection and analysis will increase the sustainability of national assessments by lowering vulnerability to staffing turnover and reducing operational costs. It would also benefit the global community by making it easier to create comparisons among different assessments and render them onto a common scale, without necessarily having to have the same assessment. In addition, it would become an internal resource to ease the work of UIS on SDG4 reporting.

The Platform may also support implementation or development of international and regional assessments. Where countries have existing commitments to participation in international assessments, and where the implementing agencies allow, countries may use the data collection capabilities of the Platform to administer tests and/or questionnaires and process data for submission to the international implementing agency. Where countries have similar measurement goals, they may share assessment instruments that share constraints on administration conditions, composition, and minimum accuracy; sharing common constraints ensures equity and facilitates joint reporting.

In all scenarios using items that are indexed against the Global Content Framework, collected item response data may be used to inform progress on SDG 4.

3. THE SDG4 TEST PLATFORM

The SDG4 item and test Platform is the main deliverable based on the work carried out by the UIS regarding the item bank Platform design and in close collaboration with countries and other partners as part of the broader objective of creating cost effective ways of testing and reporting for SDG4. The final output is a single-server application specifically to facilitate the assessment needs of countries through their in-charge public agencies (such as Ministries of Education).

Concretely, these are the features the UIS requires for the Item Bank and Test Platform (the Platform)

Country Engagement and Organization Terms of Reference

Organizations representing public agencies responsible for assessment needs of countries will be granted access to the Platform by UIS administrators. An organization may be a government or non-government entity, but it must be a formal legal entity with an executive. Organizations will be bound by Terms of Reference that encapsulate the following principles:

i. Recognition of the primary goal of supporting monitoring requirements of SDG 4. UIS will provide learning resources to assist organizations to understand SDG4 as it relates to their organizational interests. Organizations will be responsible for ensuring their managers are familiar with SDG 4 as it relates to their responsibilities.

ii. Capacity-building with respect to the GCF: developing assessment content and indexing items using the cognitive skills in the GCF. UIS will provide learning resources about the structure and use of the GCF for both content development and item review. If organizations develop items that are relevant to populations that fall within the scope of the GCF, they are encouraged, but not required to index these items using the GCF.

iii. Contribution of a minimum number (TBD) of test items per year that are released to UIS for either public use or development of ILSAs. All items that an organization releases to UIS must be indexed against the GCF. When an organization releases an item to UIS, it grants to UIS all rights

to use the item for non-commercial purposes and allows, but not requires, UIS to extend this right to other organizations. The terms should correspond to an existing licensing framework, such as the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/). Items released to ILSAs are not necessarily available to other organizations; UIS may determine which items are to be made publicly available. To facilitate preservation of organization-specific workflows that depend on common items, designation of items as publicly available is irrevocable.

iv. An organization must define a data scheme for any background data it may upload that links to individual students. This data scheme must identify variables that represent personal data (https://gdpr-info.eu/issues/personal-data/).

v. UIS may review and index an organization's test items against the GCF, even if they are not released to UIS.

vi. UIS may perform non-identifying analysis using item responses produced by test takers registered by the organization. These analyses will be used for one of two purposes:

   a. To produce aggregate statistics at national and international levels. Statistics may be calculated separately for males and females and for different age groupings, under the condition that no statistics will be calculated for groups where the sample size is smaller than 100, and all statistics reported by UIS will be accompanied by appropriate measures of uncertainty (e.g., standard errors, plausible values).

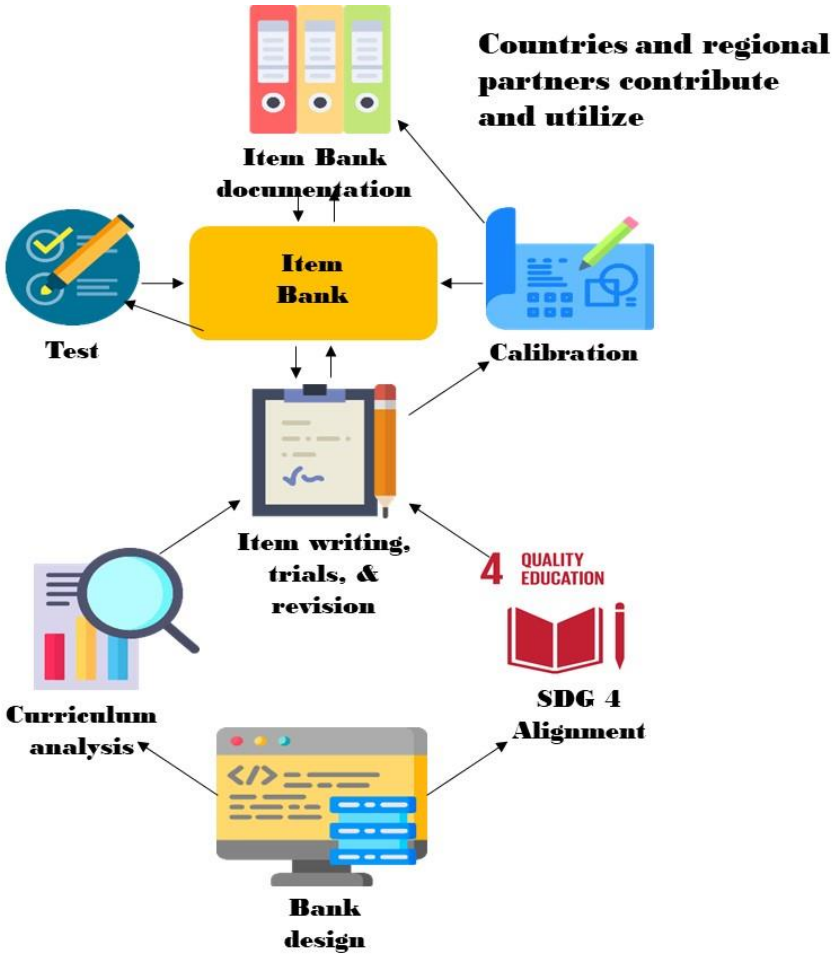
   b. To estimate the statistical properties that map items to the GCF.

vii. Organizations are free to use the Platform to perform tasks that are not specific to the monitoring goals of SDG 4. Any use of the system must be consistent with the AERA/APA/NCME standards of educational and psychological measurement (https://www.aera.net/Publications/Books/Standards-for-Educational-Psychological-Testing-2014-Edition)

viii. UIS may lock an organization's access, which includes access of any users who have been granted access under the authority of the organization, at its discretion if the organization fails to fulfill its responsibilities under the Terms of Reference or if individual users accessing the Platform under the authority of the organization violate the individual Terms of Use.

ix. Organizations are prohibited from using the Platform for commercial purposes or charging fees for access to the Platform. This prohibition does not preclude use of the Platform by private corporations operating for profit, insomuch as their use of the Platform is consistent with these Organization Terms of Reference.

Platform workflow and Country's engagement



International Law and Security

i. The Platform must be compliant with the EU General Data Protection Regulation (GDPR). The Platform provides a mechanism for organizations to classify student or test taker-specific data as personal data and allows users to delete personal data.

ii. No core functionality of the Platform depends on personal data.

iii. All access to data related to items, tests, users, students, and test results stored on the Platform must be protected by multi-factor authentication.

iv. User role-based security defines role-specific functionality and prevents users from accessing data that are not essential to their roles' responsibilities. At a minimum, the Platform distinguishes between UIS system administrators, organization managers, content developers, content reviewers, permanent student identities, and test takers (i.e., candidates for a specific test).

v. System administrators and managers can organize content developers and content reviewers into groups and control access of different groups to different subsets of content.

Globalization / Localisation

i. Interfaces may use non-language-specific symbols to communicate functionality, but all text on functional elements (e.g., buttons, menus) must be localizable. Users can select the interface language that best facilitates their work.

ii. The Platform provides a user-accessible mechanism to translate all assessment content and control assessment delivery in multiple languages.

iii. Where tests are available in multiple languages, the Platform provides a mechanism for test-takers or test administrators to select or switch between the languages of assessment.

Scalability

i. The Platform must use a server and database architecture that allows the application's common item and test database to serve all global regions.

ii. Server resources may be horizontally scaled by incrementing server resources to accommodate increased volumes of students, testing sessions, and reports.

iii. The Platform does not require scale-dependent licenses for databases or server.

Item Banking

i. Items may be created online on or uploaded to the Platform. Uploaded item definitions must be loaded in a standard schema and data format recognized by the Platform (e.g., QTI objects in JSON or XML format). NOTE: strict QTI interoperability is not essential, as many target countries may not be using QTI-compliant item banks; minimum required item interoperability must be able to import multiple choice items with a variable number of response options using a combination of formatted or unformatted text and one or more images from character-delimited or JSON formats.

ii. Items formats must support different delivery modes: paper and pencil, online, offline, and interview. Online item format support multi-state items that modify the information presented to and collected from test takers in response to timed events or test taker interactions. All online item formats are available in offline delivery using the offline test delivery mode of the Platform.

iii. Items must allow a variety of data capture and scoring methods. Minimum required response formats include: single and multiple choice, text and numeric short answer, speech, drawing, extended text response, and document highlighting.

iv. Items that are equivalent across languages, cultures, delivery modes, and response mode may be linked to facilitate common item parameter estimation, with mechanisms to unlink specific instances.

v. Item definitions for all types of items must allow online editing. Item definitions must allow specification of scoring guides for automated machine-scoring and manual scoring for both finite-response and open-ended response. Item definitions must separately define each distinct response category.

vi. Each instance of an item defines the references, tools and accommodations that are required and allowed to be used (e.g., calculator, ruler, glossary, text-to-speech). For instances that are delivered by the Platform, these resources are embedded in the item definition and rendered onscreen.

vii. Most the skills and tasks in the GCF reflect inductive and generative processes, for which multiple choice items may provide biased data. The Platform allows the definition of automated scoring methods for open-ended tasks, including text and numeric short answer, drawing, extended response, and sequential processes.

viii. The Platform provides tools for classifying items online according to their cognitive requirements with respect to the Global Content Framework. The Platform allows users to define distinct cognitive classifications for each response category. For response categories that may require or elicit different cognitive skills (e.g., individual GCF sub-construct "descriptors"), the Platform allows classification of response categories using both compensatory and conjunctive sets of skills.

ix. Organizations may use the item classification interfaces to classify items against organization-specific curricula.

PLD Definition

i. The Platform stores definitions for all PLDs and MPLs.

ii. The Platform provides tools for statistical and semantic analysis of PLDs that facilitate mapping PLDs onto both the high-dimensional space of the GCF and the lower-dimensional nonlinear manifold used by the Platform to link items and test results.

iii. Organizations may use the PLD definition interfaces to define organization-specific proficiency levels that map to organization-specific curriculum.

iv. On request and where sufficient data exist, the Platform will perform semantic and statistical analyses that map organization-specific proficiency levels to the GCF-mapped PLDs to determine how they compare to international interpretations.

Test Development

i. Test design tools allow users to target one or more PLD from both reading and mathematics and specify other constraints, which may include number of items, testing language, and delivery mode.

ii. The Platform provides test design templates that correspond to common testing situations (e.g., single-form, multi-stage adaptive, instruction-drill-and-test learning modules). The Platform

distinguishes between logical elements of the test design (i.e., testlets or modules) and the items that populate these elements. Users can modify and extend the templates to suit their assessment needs. Test developers populate the test design elements with items drawn from the item bank using manual or automated item selection and ordering.

iii. Automated test assembly functions select items and create a test form (or test schema, for tests delivered on the Platform) that can be downloaded or accessed online. Automated item selection will optimize selection against existing item family and item enemy definitions. Manual item selection uses search fields and test design constraints to find and filter items by item metadata and item content.

iv. Test definitions may be copied and modified to create new tests.

v. All test definitions include Terms of Use that describe the intended use of the test, inferences that are supported by the test data, and administration conditions that are required to support valid inferences.

vi. Users explicitly define testing constraints for Platform-delivered testing such as testing time, length of testing window, multiple sessions, secured testing, navigation between items, visibility of test progress, language switching, and review page accessibility.

Test Taker Management

i. The Platform's primary purpose is to monitor country-level progress with respect to SDG4. Accordingly, the Platform allows organizations to upload sample frames that describe primary sampling units with stratum identification information and population size. Using user-defined constraints on per-stratum sample allocation, PSU sample size, and within-PSU sample size, the Platform will design a scientific sample with predefined test-taker identifiers, which may be used for online, offline or mixed data collection. Online tests may be accessed by test takers using a modern web browser through their persistent student account, a unique URL, or by entering a test-specific authorization code. Authentication may be user-specific or test-release specific.

ii. The Platform supports offline delivery of tests using a secured test delivery application installed on a mobile device. The mobile application must be integrated with the Platform for synchronization of data, functionality, and security. The mobile application is compatible with current Windows mobile devices.

iii. Online and offline test administration save response data in real time, allowing test takers to recover interrupted testing sessions with their existing response data and progress.

iv. Where required by the test definition, Platform-based testing (online and offline) is secured by testing-site-specific or test-administrator-specific authentication, which is authorized on a per-test basis by the organization.

v. Any test may be administered using non-system modes (e.g., paper-and-pencil), and the response data may be uploaded to the Platform for scoring and analysis. Uploaded test response data must include a PSU identifier that has been provided by the Platform in a sample definition that pre-exists in the Platform.

Test Response Data Analysis

i. The Platform uses sampled data and the sample design to calculate sample-consistent weights for data analysis. Users may choose to perform analysis with or without the generated weights. Where weights are used, sampling errors are estimated using a design-consistent methodology (e.g., Jackknife).

ii. To facilitate the draft-pilot-main cycle of large-scale assessments, the Platform provides the following integrated psychometric analysis functions that provide validity evidence for test scores: classical item statistics, classical test reliability (Cronbach's alpha, overall and item-removed), exploratory and confirmatory factor analysis, differential item functioning, item response theory (IRT) item calibration, and IRT model-data fit. Users may define the specific sample of test taker response data used for item and test analysis, filtering by test release, date, specific item responses, and test taker metadata.

iii. IRT calibration should support dichotomous and polytomous scoring for single and multidimensional items. The results from IRT calibrations are available for use in the definition of reports.

iv. The Platform provides functionality to manage and collect data from standard setting panels using item-based methods (latent class, Angoff-style and Bookmark) and population-based methods (contrasting groups). Standard setting output includes numeric thresholds and sets of classified items for use in defining described scales. The numeric thresholds estimated from standard setting procedures are available for use in the definition of reports.

v. The Platform provides methods for linking items and scales of different test analyses. Required methods include common-item linking using the Stocking-Lord and Mean-Mean IRT methods, test-score linking using equipercentile and nonlinear regression methods.

Reporting

i. The Platform estimates test taker scores for each testing domain using user-supplied specifications. The Platform calculates the following types of score estimates: linear equation scores (sum, average, factor), weighted maximum likelihood (IRT), expected a posteriori (IRT), plausible values (IRT), manifold (GCF mapping). For IRT-based methods, users may specify the use of conditioning variables, which may include other testing domains, for the estimation of each scoring domain.

ii. The Platform estimates aggregate statistics for groups defined by sample strata or PSUs, consistent with the sample design of a test release.

iii. Individual and aggregate results may be downloaded from the Platform in the form of a report. Users develop reporting templates specific for each test release. Reporting templates may be copied and modified for reuse.

iv. The Platform supports definition of data-only reports, which may include data for estimated scores and individual items. Data-only reports are produced in delimited data tables, where each column corresponds to a single variable, and each row corresponds to a single observation, test

taker, or aggregation unit. Data-only reports may include sample-design data to facilitate in-depth secondary analysis in third party software.

v. The Platform supports formatted reports which can include a combination of text, graphics and tables that incorporate statistical estimates and their standard errors. Report content may be static, appearing the same for all versions of a report, or dynamic, changing according to the specific values of statistical data used to generate the report.

vi. Access to reports is restricted by role-based security by each organization.

vii. An organization's EMIS or LMS may access results directly from the Platform via the Platform's API. The API generates data objects that correspond to organization-specific API data schema.

API

The Platform provides an Application Programming Interface (API) that provides data interoperability with organizations IT systems. The API is secured with the authentication credentials, passed over HTTPS/TSL, that are associated with a manager role for an organization.

The API provides the following end points:

i. Create-Update-Delete (CRUD) operations on operational entities (managers, content developers, work panels, testing sites).

ii. CRUD operations on assessment content (items, item classification, PLDs, cognitive frameworks, tests, test releases, report definitions).

iii. CRUD operations on students.

iv. CRUD operations on test takers.

v. Read-only access to individual and group assessment reports.

vi. Read-only access to statistical results of saved test analyses.

4. Licensing Requirements

UIS has no commercial interest in the Platform and requires that it is freely accessible to all stakeholders. Accordingly, the terms under which the Platform and its components are licensed to UIS must ensure that there are no residual licensing costs for future use of the Platform and that there are no time limits or geographical restrictions on UIS' use of the Platform according to these Terms of Reference. Within these requirements, the Platform may use third-party licensed components, including open-source and commercial licenses.

The Platform does not contain any "premium" functionality that users or organizations must unlock through additional fees or subscriptions. This limitation does not prevent organizations from integrating their own applications with the Platform's API.

5. INSTALLATION

The Platform is hosted on a Microsoft Azure account. The Platform, including server applications, databases, file storage, mobile device client applications, and all other required services, must be installed and configurable within UIS' Microsoft Azure account, without dependencies on third-party services (e.g., web services, content delivery networks) for data, SaaS processing, or client-side JavaScript and CSS libraries.

6. PROJECT TIMEFRAME

The implementation of the Platform coincides with ongoing activities related to SDG4. These activities include the collection of test items from primary sources, indexing of test items against the GCF, and data collection activities of stakeholders. The implementation schedule should facilitate the following activities:

1. Entry of test items – August, 2019
2. Entry of PLDs – August, 2019
3. Indexing of test items – August, 2019
4. Online/Offline Platform-based test delivery – August, 2019
5. Data analysis – August, 2019
6. Data reports – September, 2019

The complete functionality must be implemented and available by December 31, 2019.

7. DOCUMENTATION

The Platform includes the following documentation:

1. A technical paper describing the system rationale and architecture.
2. A set of user walk-throughs describing the following workflows:
    a. Creating and modifying users
    b. Controlling access to assessment content
    c. Creating test items
    d. Translating/adapting test items
    e. Updating the GCF
    f. Classifying items using the GCF
    g. Creating a test (single for and adaptive)
    h. Defining a sample frame and testing sample
    i. Registering students and test takers
    j. Administering online test
    k. Administering an offline test

l. Administering a paper and pencil test and loading tabular response data
m. Analysing response data
n. Creating an individual report that integrates item parameters from response data analysis
o. Creating a group report that integrates item parameter from response data analysis
p. Exporting response data

3. Sample calls for each API method

4. Complete source code (for non-licensed and open-source components) and binary objects (for licensed components) required to compile and install the application(s) comprising the Platform.

Licenses, compiled libraries, version numbers and/or installation instructions for licensed components. Relevant licensing information may be included in source code materials.