

WG/GAML/11/4.2

SCOPING STUDY ON A VETTING FUNCTION AND VIRTUAL FUND FOR LEARNING ASSESSMENTS



DRAFT
February 2025



1. Background ¹

The UNESCO Institute for Statistics (UIS) has among its responsibilities the generation of evidence on the advances towards the Sustainable Development Goals (SDG). This includes a variety of challenges, including lack of data for some countries or when available, lack of criteria and procedures to compare those that exist across countries. This is the case for SDG 4.1.1: “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.”² For this indicator, UIS has made great advances, including the development of the Assessment for Minimum Proficiency Levels (AMPL) tool, which guides what should be included in evaluations of reading and mathematics. Furthermore, UIS has set up a Technical Advisory Group (TAG) to discuss how to carry out procedures that will allow to evaluate whether any specific evaluation has fulfilled all requirements to be considered valid and setting standards for comparing performances across countries. While great advances have been made on the technical side of evaluations in many countries and regions, there is a need to develop a vetting mechanism to advance on the above issues internationally. This is the first objective of this consultancy, to propose a series of procedures to implement the recommendations set by the TAG. The second goal is to propose a virtual fund to promote that more standardized evaluations are carried out and used for the improvement of education quality, particularly in Low- and Middle-Income Countries (LMICs) that have never carried out such evaluations. Given that the implementation of these two objectives would require an institutional setup, which would not be housed in UIS for a variety of reasons but would feed it with quality information and work in close coordination with its officers, we also suggest a profile of a host institution that could implement the vetting mechanism and virtual fund, aligned with SDG indicator 4.1.1. Ho (2022) provides a general framework for this type of evaluation, with low stakes, and for monitoring purposes, although below we suggest that the results of these evaluations should also be used for improving the skills of students, particularly those with lower results.

2. Goal

According to the above and the terms of reference provided by UIS, this consultancy will produce a report that will propose “an institution design for a function, ideally within an existing institution, to deal with two issues: vetting of countries’ or agencies’ submissions of assessments to be considered as reportable for SDG 4.1.1. and funding and funding coordination” (slightly amended).

¹ Authored by Santiago Cueto, Rebeca Costa and Juan Leon under the guidance of Silvia Montoya and Luis Crouch.

² Taken from <https://sdg.data.gov/4-1-1/>.




3. Methods

For this consultancy, we presented preliminary ideas in an international virtual workshop organized by UIS on September 10, 2024. We also reviewed documents and studies, as mentioned throughout this report. Finally, we interviewed key stakeholders with knowledge or experience in one or both goals. We employed thematic analysis (Braun & Clarke, 2006) to synthesize all data and interpret the patterns that emerged, identifying key characteristics for an effective vetting mechanism. We had a representation of interviewees that included experts from academia, testing agencies, networks of testing, governments and multilateral and funding agencies, among others. We balanced the interviews to include representation from all developing regions, making sure to include a significant representation from Low- and Middle-Income Countries (LMICs). Appendix 1 has a list of all interviewees and their institutional affiliations.

4. Preliminary Issues on testing validity and benchmarking

While the TAG that has been working on mechanisms to assess the reliability and validity of standardized testing programs, as well as establishing comparable benchmarks for achievement, that would allow reporting on SDG 4.1.1, inevitably the first reflection and issue that arose in many of our interviews was: *while assessing the quality of a national, regional or international testing program seems more or less straightforward using currently acceptable criteria for such practices, how will you compare performance standards across testing programs in a valid way?*

In other words, assessing the validity of testing procedures is much easier than assessing the validity of comparable benchmarks (or what the literature calls standard setting). For the first task, experts in assessment could define, if provided with relevant information, if a test in mathematics or reading covers the range of relevant contents and abilities (as specified in the Minimum Proficiency Levels or MPL, see below), has test specifications that show the main characteristics and design of the testing instruments, has a representative sample, implemented standardized procedures, analyzed the data using rigorous methods to generate results, no systematic bias occurred and overall that there is enough evidence that the data is acceptably valid to represent the skills of students at a national or regional level. However, each national, regional and international test typically will have at least one benchmark or threshold (standard) for achievement, above which students would be performing at an acceptable level, according to whatever framework has been adopted (e.g. a national curriculum). Establishing the percent of students in a single country reaching a satisfactory standard and then comparing this with other countries in reading or mathematics at any specific grade is the main piece of information that many of our interviewees would look for (see Cizek & Bunch, 2007 for a manual often used for standard-setting in education). In this way, the expectation is similar to establishing comparable international indicators for monetary poverty, chronic malnutrition, and others.



Thus, a key question is how can one compare performance standards across countries? In the extremes, among our interviewees there were two groups: the *rigorous stakeholders* said that the only way to compare test results and establish benchmarks across countries is to have a set of common items administered to different students or have a group of students take the tests that will be compared. In either of these cases, equating procedures could be performed using IRT methods. For the rigorous stakeholders, without equating, any comparison across countries will be invalid and potentially harmful as it would generate erroneous information. [Policy-linking procedures](#) that do not use equating are not considered valid by the rigorous stakeholders. In the other extreme, the *pragmatic stakeholders* will allow some flexibility in establishing comparable benchmarks (for example, analysis of the tests by experts to assess the level of difficulty and thus compare across evaluations).³

The rigorous stakeholders come all from academia and testing institutions and thus are experts in this field. It seems to us that they will object and publicly criticize the procedures of comparing benchmarks that do not include empirical equating. The pragmatic group seems to be not aware of the technical intricacies in comparing benchmarks across instruments or choose to ignore them because they favor the generation of comparable data.⁴

One way forward is to announce that the institution conducting the vetting mechanism is aware of the difficulty in establishing comparable, valid benchmarks without equating, but that when a benchmark is set for a country, a second reporting of this instrument in the future will only be allowed if there are equating procedures over time (typically, this takes the form of keeping confidential some items from one round of testing to be administered in the next round, and thus be able to equate over time). We think this is an important issue for the virtual fund, which should be acknowledged and planned to address.

Given that Minimum Proficiency Levels, as established by UNESCO, are critical for both the vetting mechanisms and virtual funds, we provide a brief description below.

³ Relevant efforts have been made to equate scores among regional evaluation programs: SACMEQ – Southern and Eastern Consortium for Monitoring Educational Quality; PASEC – Programme for the Analysis of Educational Systems; LLECE – Latin American Laboratory for the Assessment for the Quality of Education; SEA-PLM – Southeast Asia Primary Learning Metrics; and PILNA – Pacific Island Literacy and Numeracy Assessment, and the international evaluation programs, TIMSS and PIRLS. This was called the Rosetta Stone project, a collaboration between UIS, IEA and other stakeholders (https://uis.unesco.org/sites/default/files/documents/Draft_proposal_for_linking_regional_assessments_to_TIMSS_and_PIRLS.pdf).

⁴ UIS has hired two consultants who are working on methods for valid comparison of benchmarks across testing programs, whose report should be ready by the end of February 2025; the recommendations of this report will be very relevant for the current report.



5. Minimum Proficiency Levels (MPL)⁵

Established in 2018 and updated in 2020, the Minimum Proficiency Levels (MPL) serve as a standard for basic knowledge in different areas, highlighting reading and mathematics for SDG 4.1.1. It specifies the skills and capabilities that students must exhibit at a particular grade level in literacy and numeracy. Thus, MPL is a tool for countries to monitor progress towards learning goals and identify areas where support is most needed.

MPL definitions aim to ensure comparability across learning assessments. These definitions were derived through an analysis of performance level descriptors (PLDs) from international, regional, and community-based assessments in reading and mathematics (Ovsyannikova, 2019). To establish an operational definition of MPLs, cross-national assessments (CNAs) were used through a PLD analysis with international tests. This thus became a tool to comparability across learning assessments. UIS published in 2017 [a guide to implementing a national assessment](#). The UIS has also made advances in [instruments to measure MPL](#), including a list of countries where they had been used up to 2023 which are relevant particularly for the virtual fund mentioned below.

Before we turn to the recommended vetting mechanisms and virtual fund, we briefly discuss below a principle for all education practices, including testing, which has to do with the right for all children to not only access formal schooling, but also acquire skills that will allow them to be active and productive citizens.


6. Overarching principle: education as a human right

Education has been considered a human right since the Universal Declaration of Human Rights in 1948⁶. Since then, major gains have been made in terms of access to formal education around the world, particularly primary education. However, currently it is widely accepted that access to school is not enough, but children need to learn skills, particularly what have been called foundational skills, related to reading and mathematics⁷. The idea behind foundational skills is that they are both desirable by themselves as key for the development of individuals in areas such as health, work, communications and self-fulfillment, but also for learning of other subject areas at school and learning throughout

⁵ For more information about MPL and SDG 4.1.1., see <https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2020/09/Metadata-4.1.1.pdf> and <https://ampl.uis.unesco.org/minimum-proficiency-levels-mpl/>.

⁶ UN General Assembly, Resolution 217A (III), Universal Declaration of Human Rights, A/RES/217(III) (December 10, 1948), <https://www.un.org/en/about-us/universal-declaration-of-human-rights>.

⁷ For example, the [Coalition for Foundational Learning](#), integrated by FCDO, UNICEF; UNESCO, USAID, the World Bank and the Bill & Melinda Gates Foundation was established in 2022 to promote learning, particularly in low and middle income countries (LMICs).



life (Belafi, 2020). In other words, the right to education in the contemporary world de facto includes the right to learn relevant skills in rapidly changing societies. This principle of developing skills as part of the education right has been acknowledged by rapporteurs on education from UNESCO, who within this field have identified several forms of educational equity⁸.

The fact that skill development is considered a part of what education seeks requires an inclusive perspective in the assessment practices. SDG 4.1.1 calls for disaggregated data by gender, and indeed this should be the first indicator to be produced. However, gradually the level of inclusion in testing programs should be expanded according to the perspective adopted here. First, whenever possible, other results should be provided to reflect levels of inequity. These include disaggregated results by levels of poverty, area of residence and type of school attended, among others. Second, gradually, efforts should be made to include groups that are often not considered in standardized evaluations, given the difficulty of testing them with suitable instruments and the extra costs to assess them. These marginalized groups in testing include children with disabilities, members of ethnic groups with a minority tongue, refugees, migrants or displaced children, and children in conflict zones⁹. Testing these students requires adaptations to capture their best performance and learning needs that should be developed. In other words, the field of testing needs to embrace notions of diversity and inclusion that are currently adopted in a variety of education programs. As a first step in the short run, it should be required that all test reports acknowledge the groups of children included in the testing program (current definitions of “exclusions” are often not detailed about vulnerable groups such as those described).


Finally, given that we are arguing for the right to learn and given that for the most part national evaluations only include children attending formal schooling¹⁰, all reporting of evaluations should include a note indicating what is the formal school coverage of the target age, to help with the interpretation of the results. This is information that national evaluations usually do not include.¹¹ The above view was frequently proposed by interviewees, who wanted to highlight the importance of measuring learning outcomes,

⁸ See for example <https://www.unesco.org/en/right-education/need-know>.

⁹ We will call these vulnerable groups here, as they are vulnerable to not being included in assessments and also face barriers to complete basic education in many countries.

¹⁰ There are a few home-based evaluations, regardless of whether the child is attending schools, notably ASER in India (<https://asercentre.org/ascer-survey/>), but these are usually not accepted by governments as official results.

¹¹ As an example, consider country 1 with 60% of students above a benchmark for satisfactory achievement, but with only 70% coverage in the target grade of which a majority are boys, versus country 2 with 50% above the same benchmark but with 95% coverage with equal gender participation.



particularly in LMICs and among vulnerable populations, to address systemic challenges and enhance the quality of education for all.

7. Designing a vetting mechanism for learning assessments


Regarding the assessment of the validity of the testing instruments and procedures, there seems to be consensus among interviewees that the vetting mechanism can be carried out by expert reviewers. We propose that these procedures be designed and implemented in a similar way to those used in the review of academic articles by professional journals, although they should be adjusted to provide support and be formative for participants, considering that one of the aims of this initiative is to increase country participation, particularly of LICs.

The procedure would start with the requirements that the TAG and current consultancy on benchmarking procedures (mentioned above) would be made public, similar to what in academic journals appear in “instructions for authors”, i.e. indicating what information should be sent and in what format for evaluation by reviewers¹². This information would be sent to evaluation country units, regional networks (e.g. from Latin America, Africa and others), and other relevant stakeholders (e.g. testing programs by IEA, OECD and international agencies supporting testing programs, particularly in LMICs). To expedite procedures, the information could be uploaded as requested in a user-friendly program, developed specifically for this vetting procedure, similar to software currently used in academic journals. However, only authorized evaluation officers from countries, networks or testing programs would be given access to upload this information.

The software for uploading testing information should give automatic feedback on whether the information is complete or point out what is needed. As mentioned above, the assessment of the quality of the testing instruments and procedures should be straightforward, but rigorous, with the report of reviewers being of great interest to countries and the whole vetting initiative. It is the assessment of the performance standard that provides more challenges.

Provided with all the above information, a team of reviewers would evaluate the report remotely (although a few random visits could, in principle, and funding allowing, be

¹² As an example of the types of reports, data and instruments that would be requested, there should be a general report on the testing goals and general procedures; test specifications and the tests themselves, in all the tongues that have been administered with a translation to English; sampling procedures and coverage of the target population, including information on vulnerable groups and exclusions; manuals of test administration; procedures for establishing benchmarks and the results of these; information on the reliability and validity of data, including analysis of bias, if available; and other relevant information. We do not anticipate requesting the data itself, as this may have sensitive or confidential information that the countries may not be able to share.




programmed for verification and refinement of procedures). The initial response would be, using the analogy to a review for a journal paper: Accept the information submitted and proceed to evaluating it; request further minor information before evaluating the materials (specifying which); request further major information (specifying which); or not accept the evaluation, requiring essentially a complete resend of the submission.

With a complete set of data, the institution in charge of organizing the vetting procedure (see below) would form a panel of reviewers that would issue a report analyzing the assessment and providing a general recommendation: the report complies with all technical requirements and thus should be fully considered for SDG reporting; should be considered for SDG with caveats on its interpretation; or does not comply with the criteria and should not be reported by UIS. The report could pay particular attention to Criterion 6, namely comparability to the international standard of the MPL, with particular caveats. The caveats could include: *the procedures in testing followed strictly the MPL in the design and administration of the instruments, but the results of this country may not be compared to other countries*; and the like.

As usually done in academic journals, reviewers should write a set of recommendations that would be internal, for the host institution responsible for the vetting mechanism (see below); these could include suggestions on the quality of the data requested for the assessment of the testing program and procedures in doing the review, that would help to continually improve the vetting mechanism. Also, the reviewers should write a set of recommendations to be given to the country(ies) submitting the information, written in a formative style, with suggestions on how to improve the testing program and pointing to resources that could help improve it. This feedback would not be public but sent only to the country officers or specialists of the evaluation. The goal of this procedure would be to enhance the quality of submissions and encourage participation, making it essential to provide formative-focused feedback rather than overly harsh criticism, which is typical of many academic reviewers.

On the panel of peer reviewers, we recommend that it includes experts on a variety of fields, including thematic (e.g. reading or mathematics), statistical procedures (for sampling, analysis and standard-setting procedures), on procedures of administration and inclusion of vulnerable students, and at least one expert in the education of the specific country or region. Several interviewees made the point that some countries, particularly LMICs, have tests administered in two or more local languages, thus generating the need to include experts in these languages and cultures. While high technical expertise is valued, so is familiarity with context, so that the report of the reviewers speaks to local needs. The number of reviewers by evaluation could vary between four and six, and they could work remotely, each assessing a specific part of the evaluation but being able to give an opinion on all sections of the report. There should be a leader for each panel of reviewers, responsible for assigning tasks for the reviewers, compiling the sections, preparing a report and presenting it.




On the selection of peer reviewers, it seems to us that standardization of procedures is key for the vetting mechanism to be valid. This would require that the reviewers be selected based on their expertise and credentials, according to the above, but also that they go through a training process prepared by the host institution conducting the vetting procedure (see below). The simpler mechanism would seem to be to prepare a set of documents, including all the criteria posed by the technical advisory group, and virtual tests on these that all reviewers should pass before being considered in a panel for an evaluation. Once they pass, their status as peer reviewers would remain for a given period, perhaps two years or when the criteria are significantly changed by the host institution, after which they would need to pass the virtual tests again. The panel of peer reviewers to evaluate any given assessment would come from this certified pool of experts, whose names would be public to enhance the credibility of the vetting mechanism.

The work of reviewers would be remunerated, with a specified time frame to complete the task. It seems difficult to establish the number of days of consultancy to be paid to the panel. The rates would vary if it were a national evaluation with one language, more than one language, if it is a regional evaluation or a cross-country evaluation (depending on the number of countries, more days of consultancy or more reviewers would be required) or any other additional complexity to the task. However, other than allowing for different rates depending on the scope of the job, the rates would be determined based on the number of days it would require to be completed, at a daily rate that would be the same for all reviewers; the leader would get extra days. This would be needed to create equity between reviewers from wealthy economies and those from developing economies, but also to increase transparency and reduce transaction costs. Each panel of reviewers should be given around six weeks to complete the internal and country reports mentioned above.

8. Designing a virtual fund to promote learning assessments in Low- and Middle-Income Countries (LMICs)

As in most indicators of educational quality, the field of testing and reporting on students' skills is marked by inequality. This inequality is characterized by lower performances of vulnerable children (as described above), but also by lack of information, which typically occurs in many low-income countries (LICs). Thus, the importance of adopting a rights-based approach to measuring and reporting on learning skills, as discussed above. Silvia Montoya and Luis Crouch have published a [blog](#) on the need for the fund and how it could be shaped and a [summary of current international assessments](#). UIS has also published an [inventory of learning assessments](#) that is relevant for this fund. The Global Partnership for Education (GPE) has recently [published a report](#) on the status of measurements across countries. GPE has been supporting national evaluations in LMICs. More information is available on higher primary grades and in secondary than in the first grades of primary, hence the urgent challenge to report on these.

The proposal to design a virtual fund to promote standardized evaluations in LMICS would have several goals: the first one is to promote that high-quality evaluations, aligned with



the MPL but also with local laws, plans and instruments (e.g. national curriculums) are implemented. In countries where there is no tradition of implementing standardized evaluations, international technical assistance would be required. However, many of our interviewees emphasized that this technical assistance should be implemented so that local capacities are developed, leading eventually to countries being able to carry out their own valid evaluations¹³. The evaluation plan should include an alignment with MPL and also provisions for equating results across time within the country and among countries.


Second, evaluations in LMICs should be carried out so that uses of the data are developed to improve learning skills of children, particularly of low performing students, thus reducing inequalities within countries. In reviewing the literature and from the interviews, two uses of data seem most promising: identify lower-achieving children, which would lead to higher investments in these children or specific programs developed or strengthened for them.¹⁴ The second promising use of data would be related to pedagogical issues, including revisions of national curriculum, developing or improving educational materials (e.g. textbooks, workbooks and concrete material, but also pedagogical uses of technology), and improving teachers pedagogical skills (pre-service or in-service).

For this virtual fund to work, a list of countries and evaluations in reading and mathematics per grade should be developed, including national and regional evaluations. This would help identify countries that have never participated in an evaluation, which would be the first candidates to be supported by the virtual fund. Second, a list of donors interested in funding evaluations of the type described here should be assembled, with comments about major regions of interest. Third, a list of experienced agencies and universities giving support to carry out evaluations as the ones described here should be assembled. Alternatives for testing include using or adapting an existing instrument or developing a new one¹⁵. A successful implementation of an evaluation project will depend on the matching of the three stakeholders mentioned for a specific country or region. We suggest that a simple program is developed with the three stakeholders. For the countries, this would include a listing of the national and international evaluations performed in the

¹³ Manos Antoninis, director of the Global Education Monitoring Report (GEM), has written a blog emphasizing that countries need to be supported in their needs for evaluation use and local capacity development (<https://world-education-blog.org/2024/03/26/on-the-way-forward-for-sdg-indicator-4-1-1a-supporting-countries-development-needs/>).

¹⁴ For example, evaluation data could be used to implement teaching at the right level interventions, which have been shown to be effective in LMICs (Snilstveit, 2015).

¹⁵ A “Buyer’s Guide” for Student Learning Assessments is currently being developed by Kevin Macdonald for UIS. This guide would be of great relevance for the virtual fund implementation. Also, UIS and the World Bank have produced a guide for cost-effective approaches to purchasing tests (Banerjee et al, 2023). There is also a “How to guide” on national evaluations that includes a description of existing testing programs that is relevant for the planning of the virtual fund (https://cdn.prod.website-files.com/61366d43ebd6df56d9b67a11/61748baba5e6696e8389b952_cBuATsusFbTaPCp9_Dcu8bhE1tIMN7Sci-Assessment%20at%20systems%20level.pdf).

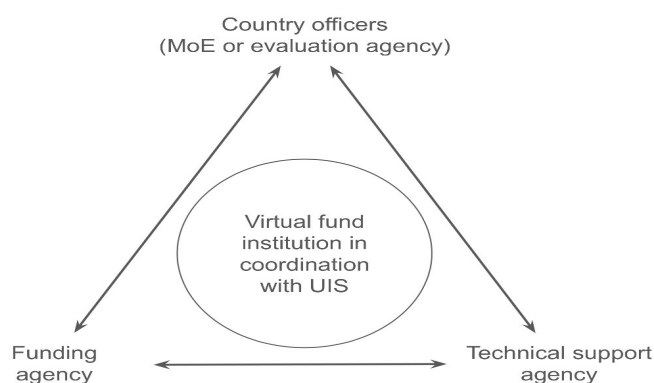


country, by year, area and grades, the external support received (if any) and the plans for future evaluations. For the funding agencies, the programs they have sponsored, by country, areas, grades and years, plus instruments used. For the technical agencies, similar to the above, the countries they have supported, in which years, grades and areas, and with whose support. The program should be able to automatically produce crosstabs with the above information, so that, for example, countries with no support are identified or in the opposite, countries receiving support from several agencies for similar purposes are identified. Additional key information to be included in this software is the number of students included in each evaluation and the cost of testing them (overall and per-students), as well as links to reports or any other products deriving from these efforts.

For a successful matching among the three stakeholders to occur, the reasons behind a country not having national evaluations should be attended. While our interviewees mentioned several and gave examples, they seem to fall under two main categories: lack of (or not enough) local funding and politics. The lack of funding seems to be easier to handle, as it would depend on the three-way matching described above (considering that procuring funding is always challenging). However, even a small financial commitment should be requested from the local government, to complement potential international funding. Also, a commitment to generate an evaluation unit within the country should be generated, with a long-term plan for evaluations and its uses. The political reasons for not having evaluations may be more difficult to attend and need to be discussed with local stakeholders. A typical objection is that the results of the evaluation “will make the government look bad”. This is more relevant in countries with a poor education system and a political leader or party in power for several years. For this, local allies working with the government may be useful, seeking to find arguments in favor of testing that will work locally.

Additionally, there is the issue of who starts the process of requesting support from the virtual fund. Ideally, it should be the country, but we can anticipate that in some cases, the virtual fund agency or UIS itself will want to work with countries not having evaluations ever to begin a process of international support. Many of our interviewees insisted that the testing program should not be “one size fits all”, nor “one test fits all”, but a collaboration between local officers a funding agency (if needed) and a technical agency for supporting local processed, with the goal of reporting data for SDG 4.1.1, but also fulfilling the expectations of the country. Figure 1 presents the expected interactions of the three main stakeholders, but the virtual fund institution, in coordination with UIS, should be in the middle of the process, identifying priorities for evaluation and funding, potential candidates for providing technical support and monitoring the process of implementation from a certain distance. this would all help to continually refine the process of the virtual fund.

Figure 1. Main stakeholders in virtual fund¹⁶



In terms of operations, we suggest as a first step that a software as the one described above is developed and a meeting between the UIS, the virtual fund institution, and potential donors is convened, to follow this up with interactions with selected countries to assess their interest in having their own evaluations.

9. Institutional background for the vetting mechanism and virtual fund

The work of professional, standardized reviewers to assess the validity of the evaluations and benchmarks for achievement, as well as the work related to the virtual fund, would need to be supported by an international institution. Such institutional support is necessary in the vetting mechanism for the peer reviewers to carry out their work, providing them with the needed material, contracts and guaranteed independence in their work. For the virtual fund, the role of the institution would not be only connecting stakeholders, as described above, but monitoring the implementation of the designed plans, seeking to refine them constantly. A political component is key in the definition of the plans of the virtual fund, for which the participation of UIS is particularly relevant. Assembling a team within UIS to carry out the vetting mechanism has been ruled out as an option from the beginning. However, UIS has to report on SDG 4.1.1, so the institution that would provide support for the work of the reviewers would be an intermediary between countries reporting results and UIS presenting them internationally. The virtual fund could be included completely within UIS, but we anticipate that the amount of work

¹⁶ In many countries, the evaluation offices are located within the Ministry of Education, while in others there is a specialized, somewhat independent evaluation office (for example in Brazil: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb>). While we have not found research on the pros and cons of each approach, it would seem that having an independent office frees it from the political pressures that come from being a part of the Ministry of Education.

would be significant and there are potential synergies between both mechanisms, which is why we are suggesting here that to have a single international institution carrying out both types of tasks, in close coordination with UIS.

In our interviews and review of similar mechanisms in education and other fields we asked what this institution would look like to maximize efficacy and efficiency. Several options appeared. These are presented and discussed in the table below.

Table 1. Institutional support for the vetting mechanism and virtual fund to assess evaluations

Type of institution	Comments	Recommended
UNESCO specialized institution (separate from UIS)	We discussed with stakeholders the possibility of establishing a specialized institute in educational evaluation, linked with UNESCO, similar to the Innocenti ¹⁷ center from UNICEF. This would not be created immediately, but in a few years, once the vetting mechanism is consolidated (for the immediate work, it would operate with another mechanism). The opinion of ALL interviewees was that this would not be desirable, as it would generate more bureaucracy and would not be efficient.	No, but UIS needs to work in close coordination with the international institution in charge of the vetting mechanism and virtual fund.
Existing evaluation company or agency (e.g. IEA or Pearson)	This option was suggested by one interviewee, who argued that these companies would have the “right incentives” to complete the work and procure funding. However, all other interviewees thought that it would be better to have an agency that has no interest in selling products or extend the work they were doing, so that they can have a neutral approach to all testing programs.	No
University or consortium of universities	The argument for having academics from universities in charge of the vetting mechanism was suggested by those who think that the work needs to be done in the most rigorous ways possible. However, this suggestion was discarded by most interviewees, who thought that a university would not be agile enough to provide results or compromise their academic integrity or reputation with results that would	No

¹⁷ See <https://www.unicef.org/innocenti/es>.

	<p>not always be as rigorous as desirable. Also, again, the issue of equating is likely to come up and compromise the reporting of students above a desirable international threshold.</p> <p>Finally, the role of connecting stakeholders, as required for the virtual fund, would not seem natural to facilitate if in charge of university academics or officers.</p>	
<p>Research association already working in the developing world</p>	<p>Some research associations have been working in the developing world, promoting evaluations. In some cases, they have interest in a particular type of evaluation or country (e.g. ACER) and thus would have a conflict of interest. Other types of research associations do not go beyond academic activities (e.g. CIES), and should not be considered, but should be among the relevant stakeholders of the two mechanisms described above. Thus, the advice is to stay in contact with these institutions, learn about their work supporting countries to carry out evaluations and consider them potential allies, particularly for the virtual fund, but they would not work well as the institution in charge of the vetting mechanism or virtual fund. These agencies also have interest in specific countries or regions that could bias the priorities of the virtual fund.</p>	<p>No</p>
<p>Institution collaborating with education in the developing world, particularly in low-income countries (e.g. GPE).</p>	<p>These institutions have provided funds for education in the developing world, and in some cases supported national evaluations. These programs in general are not as agile or focused as would be desirable, for a variety of reasons, and restarting them would mean resolving some of the institutional issues why they are not efficient mechanisms for this purpose.</p>	<p>No</p>
<p>Multilateral organization working in education (e.g. the World Bank)</p>	<p>While the WB and similar institutions have done very interesting work in education, including supporting assessment they would probably be too bureaucratic. They should be considered among the stakeholders for the virtual fund, but should not assume the general role of supporting the two mechanisms described above.</p>	<p>No</p>
<p>Existing non-profit organization, or</p>	<p>A non-profit organization would generate a sense of confidence that this is not a business-</p>	<p>Could be. We are thinking of institutions</p>

<p>network of organizations, with an international scope of work in education, not including assessment, but willing to generate an office for this work.</p>	<p>oriented activity, a relevant issue for many stakeholders. The institution would have a reputation for their good work in education, with many contacts in the education sector in LMICs, that could translate to the new areas of work. Since one of the goals (see below) is to promote evaluation in LMICs and particularly LICs with no history of evaluations, it would be good that the institution is from the Global South or does intensive work already in the South. It would also be desirable that they have fluent contacts with agencies and ministries in all regions of the developing world.</p>	<p>such as NORRAG or Southern Voice¹⁸, who already have a network of countries working with them in education.</p>
<p>Promote the creation of a non-profit institution or network specifically in charge of developing the work needed for the vetting mechanism and virtual fund.</p>	<p>The new Institution would have the advantages of being a new actor with a single goal, linked with the two mechanisms, and the task to prove itself in the short term. This could be done by recruiting existing institutions or experts to form a consortium representing all regions from the developing world. This would be an organization led by experts in education and assessment that can provide a balance between the rigorous and pragmatic views discussed above. The institution would need legal representation, so as to be able to sign contracts and receive funds of cooperation.</p>	<p>Could be; in this case or the above, a representation of the main regions of the developing world (i.e. Latin America, MENA, Africa and South-East Asia) should be demonstrated</p>

¹⁸ Disclaimer: GRADE is the host institution for Southern Voice.

10. More on the institutional selection and setup

Perhaps the best way to select an institution to carry out the vetting mechanism and virtual fund would be to conduct an open call for proposals, but if a more speedy procedure is required, an invitation could be sent to selected institutions to ask for a proposal and work plan for a period of three years or so. The reflections below are meant to describe the main functions the institution would carry out, again, in close coordination with UIS.

Interviewees emphasized the need for the institution to possess certain characteristics to ensure that it can operate with the agility required for timely reporting related to the vetting procedure and making connections among stakeholders in the virtual fund. Furthermore, the entity's constitution and the services it provides must be designed to not only ensure its sustainability but also create an environment that is encouraging for Low and Middle-Income Countries (LMICs). In this regard, perhaps the key characteristics of the selected institution would be to be credible because it uses state-of-the-art procedures to accomplish the vetting procedures and is a partner that can promote and continually improve fruitful alliances related to the virtual fund. Below we describe with more detail what could be the functions that this institution would be expected to accomplish:

Table 2. Institutional functions to carry out the vetting procedures and virtual fund

Function	Description
Vetting: Design and implementation of a submission portal	It will be necessary to develop a virtual portal for file submission of the requested information from national or regional evaluations. Country representatives would be authorized to upload the information and be able to check the status of their submission.
Vetting: Technical assessment of evaluations	When complete information is uploaded on the portal, the institution would assemble a panel of reviewers, with a lead for each team, hired from the pool of certified reviewers. The institution would then receive the report from the reviewers and if complete, would issue a recommendation for UIS to publish the results with or without caveats on the interpretation of the results, as well as a confidential report to the country or regional team with recommendations for future evaluations.
Virtual fund: develop a software to help prioritize countries, evaluations and operational plans for national or regional evaluations	The institution would need to develop a simple software of the work done in countries, funded externally by institutions and supported by technical agencies to help make a plan to connect stakeholders. A prioritization of the work would need to be done based on this mapping; given the concern for equity adopted here, we suggest to start with LICs with no history of evaluations that are willing to engage in this process. The institution should make it a focus of its work that the evaluation plan is developed to the satisfaction of local stakeholders. Eventually, we would expect that

	all countries see this type of monitoring as essential to educational development, hence including it fully in their national budgets.
Virtual fund: Promoting the use of evaluations	While many stakeholders argue for the need to have more standardized evaluations, it is interesting that many do not have specific examples on how to use them ¹⁹ . As described above, the potential uses include prioritizing populations within countries not included in evaluations, included but with low performance, and using the results in new or revised pedagogical plans (e.g. educational resources, teachers' pedagogical skills or revision of the curriculum). This would not be a role of the institution, but of the technical agency providing support to specific countries, but the institution could help as a collector and sharer of good international practices that could be adapted.
Vetting and virtual fund: Capacity building	All the plans for new assessments (virtual fund) should include a strong component to develop or strengthen local capacity to carry out valid and useful evaluations. However, given that the reviewers in the vetting mechanism will also include recommendations for countries in their confidential reports, the institution could support the development of local capacities and become a source of knowledge for a variety of partners.
Vetting and virtual fund: Communications	By this task, the institution will prepare a plan to engage with key stakeholders regarding both activities (vetting and virtual fund) through a variety of ways. The plan would need to differentiate activities for dissemination and engagement by types of stakeholders, with an emphasis in LICs. This plan would be based on a theory of change for both activities, seeking an improvement in achievement, a reduction of within-country inequalities and an increasing inclusion of all students in evaluations, based on the right to education approach described above.
Additional opportunities: policy-oriented research	Over the years, the institution will generate links with countries and international agencies in possession of databases with great relevance for policy-relevant research, as well as reports from reviewers and from collaborations linked with the virtual funds. We think that this opens an opportunity to use this information for research that would lead to a variety of lessons learned in the vetting mechanism and virtual fund. Doing this will require an amount of work that would not be possible in the short run (e.g.

¹⁹ A few examples are Bos, W. & Schwippert (2003); a study on the impact of TIMSS in low and middle income countries by Warwick B. Elley in 2002 (https://www.iea.nl/sites/default/files/2019-02/Elley_Impact_TIMSS-R.pdf); the impact of TIMSS and PIRLS in LMICs by Alison Gilmore in 2005 (https://www.iea.nl/sites/default/files/2019-04/Gilmore_Impact_PIRLS_TIMSS.pdf). More research on effective uses of standardized evaluation data are needed.



	generating anonymized data bases at the level of students, schools and provinces, and identifying regions and countries, as well as sociodemographic variables, that could eventually be uploaded to a public repository), getting consents from relevant actors as needed and other tasks. Databases could be released in a public repository. Furthermore, the institution could form alliances with researchers and institutions, defining an agenda of both academic and policy relevance.
--	--

Below we present some additional considerations on the profile of the institution:

Size: the institution would have a small size; the minimum personnel would include an executive director, a director of the vetting mechanism and a junior assistant, a director of the virtual fund and a junior assistant, one specialist in reading and one in mathematics, a specialist in psychometric and statistics, a specialist in capacity building related to evaluations, a communications professional, and a person for administrative support. There should be funds to hire persons or institutions for several of the activities, including the reviewers and support for the virtual fund.

Governance committee (GC): many interviewees pointed to the importance of having some representation from stakeholders on the methods and results. The governance committee would meet twice a year to approve the general procedures of the institution, but the approval of specific procedures would be defined by the institution, so as to promote efficiency. The number of representatives of the governing committee is to be determined, but it would seem convenient to have stakeholders from all major regions and evaluation networks, with a significant representation from the Global South. The members of the GC would also act as representatives of the institution and help with contacts in their regions of work.

Funding and incentives: the institution should be formed with some initial funding to start activities, although the amount needed, and source of these funds is yet to be determined. Below we present some of the stages of the initial development of the activities of the institution, which if completed successfully and on time, should lead to a renewal of its responsibilities.

Establishing indicators to effectively measure progress towards the SDG 4,1,1 presents significant challenges. As evidenced by interviews conducted, the responsible institution must carefully balance technical requirements with the practical realities of the work and the political pressure to report on SDG 4.1.1 as soon as possible. Therefore, robust mechanisms are crucial not only for reporting and monitoring these indicators but also for providing essential technical support and facilitating the development of necessary tools within each country. As one interviewee said: “there is no point in making this whole effort if everything we are seeking is to populate a cell with a number for an international report”. In this whole effort, the final goal of improving the skills of students using standardized, valid information, should not be lost.



11. Stages in the development of the institutional work

In the work of the institution, several initial stages can be anticipated:

Stage 1: Institutional setup (between six months and a year)

Vetting mechanism: Establishing procedures

The initial step involves setting up the procedures to process requests for validation of procedures and establishing benchmarks as defined by the TAG. It would also need developing a software for the automatic processing of requests, similar to the ones used by academic journals (as described above). Parallel to this, there would be a need to define a list of potential reviewers with their specializations and regions of expertise. These reviewers would need to be trained to certify that their reviews are standardized.

Virtual fund: Scoping countries and stakeholders

In the virtual fund, a list of countries with no evaluations aligned with SDG 4.1.1 would need to be defined. Along with this, a list of potential funders and expert agencies that could provide support would also be assembled. A software including this information would be set up, so that it helps prepare a plan with priorities.

Management: Preparing a work plan

After the institutional team is assembled, there would be a need to establish a work plan for at least two years, that could be revised every six months. The Government Committee would also be assembled, which would approve the work plan towards the end of this first period. It should have clear milestones and products for every six months of work. The communications work plan should also be approved, including a theory of change and deployed for action.

Stage 2: Deployment, fine tuning and expansion (two years)

Vetting mechanism: taking advantage of the low-hanging fruits

Once the procedures from stage 1 are in place, the institution should start a strong campaign to include testing programs, national and regional, involved in the TAG procedures mentioned above, so that SDG indicators can be produced. One issue in this line of work is that some regional programs already have their programs and standards and may not have incentives to go through a vetting mechanism that could result in recommendations that could be costly or not considered convenient for them. We thus suggest to start with the evaluation programs who volunteer to go through the vetting mechanism, in the expectation that participation will grow over time. However, given that going through the vetting mechanism will require additional work from local evaluation officers, the close support of UIS for these activities will be required, supporting the institution with contacts and messaging on the importance of the vetting mechanism



Virtual fund: *first collaborations implemented*

As explained above, we expect that the virtual fund will allow the collaboration of evaluation country officers, funding agencies and expert agencies in evaluation. The work should start with at least one LIC country with no previous evaluations performed. We suggest that the hubs of [KIX](#) countries are considered for this, as their hubs may facilitate prioritization.

Management: *close supervision, dissemination and monitoring*

During stage 2, management should monitor that the milestones and goals are reached, inform the Governing Committee of outcomes and barriers and refine the procedures for both vetting mechanisms and the virtual fund.

DRAFT



References

Arias, E., Dueñas, X., Giambruno, C., & López, A. (2024). El Estado de la Educación en América Latina y el Caribe 2024: la medición de los aprendizajes, IADB: Inter-American Development Bank. United States of America. Retrieved from <https://coilink.org/20.500.12592/5pxy7ii> on 19 Jan 2025. <http://dx.doi.org/10.18235/0013171>.

Banerjee, Abhijit; Andrab, Tahir; Banerji, Rukmini; Dynarski, Susan; Glennerster, Rachel; Grantham-McGregor, Sally; Muralidharan, Karthik; Piper, Benjamin; Jaime Saavedra Chanduvi; Yoshikawa, Hirokazu; Ruto, Sara; Schmelkes, Sylvia.

(2023). Cost-effective Approaches to Improve Global Learning : What does Recent Evidence Tell Us are Smart Buys for Improving Learning in Low- and Middle-income Countries? Washington, D.C. : World Bank Group. <http://documents.worldbank.org/curated/en/099420106132331608>.

Belafi, C., Hwa, Y., & Kaffenberger, M. (2020). Building on solid foundations: Prioritising universal, early, conceptual and procedural mastery of foundational skills. Research on Improving Systems of Education (RISE). https://doi.org/10.35489/BSG-RISE-RI_2020/021.

Bos, W. & Schwippert, K. (2003). The Use and Abuse of International Comparative Research on Student Achievement, European Educational Research Journal, Volume 2, Number 4, 559-573.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>

Cizek, G. J. & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards for tests*. Sage Publications.

Ho, A. (2022) "Specifying the Three Ws in Educational Measurement: Who Uses Which Scores for What Purpose?." J Educational Measurement 59, no. 4, 418-422. DOI: 10.1111/jedm.12355

Howells, K. (2018). The future of education and skills: education 2030: the future we want. OECD, Paris.

Ovsyannikova, O. (2019). GLOBAL PROFICIENCY FRAMEWORK: READING AND MATHEMATICS - Grades 2 to 6, UNESCO Institute for Statistics. Canada. Retrieved from <https://coilink.org/20.500.12592/rq3q3x> on 27 Jan 2025. COI: 20.500.12592/rq3q3x.

Snilstveit, B, Stevenson, J, Phillips, D, Vojtkova, M, Gallagher, E, Schmidt, T, Jobse,

H, Geelen, M, Pastorello, M, and Eyers, J, (2015). Interventions for improving learning outcomes and access to education in low- and middle- income countries: a systematic review, 3ie Systematic Review 24. London: International Initiative for Impact Evaluation (3ie).

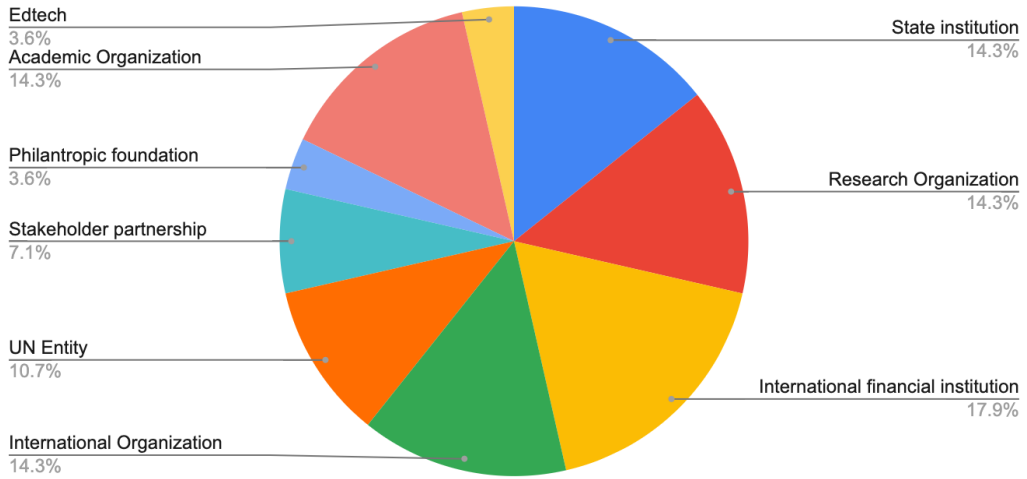
Appendix 1. list of key stakeholders interviewed

Nº	Institution	Interviewee	Role	Date of interview
1	NEQMAP - Institute of Informatics and Development (IID)	Ahamed, Syeed	Member of steering group	11/18/2024
2	UNESCO Global Education Monitoring Report	Antoninis, Manos	Director	11/22/2024
3	Inter-American Development Bank (IDB)	Arias, Elena	Senior Education Specialist	12/13/2024
4	GPE, Data and Evidence	Atis, Evans	Data & Evidence Lead	10/16/2024
5	PRATHAM	Banerji, Rukmini	Chief Executive Officer of Pratham Education Foundation	10/21/2024
6	Educational Quality and Assessment Programme (EQAP)	Belisle, Michelle	Director of EQAP	10/21/2024
7	World Bank	Benveniste, Luis	Global director of education	10/18/2024
8	FCDO	Berry, Chris	Senior Education Adviser	10/22/2024
9	Education Horizons	Centenera, James	Cofounder	10/23/2024
10	Gates Foundation	Dintilhac, Clio	Senior program officer in education	9/24/2024
11	Universidad de Pretoria - WERA	Ebersohn, Liesel	President WERA	1/14/2025
12	FCDO	Freya, Perry	Education research advisor	10/22/2024
13	IEAc	Hastedt, Dirk	Executive Director	10/17/2024
14	Harvard University	Ho, Andrew	Charles William Eliot Professor of Education	12/4/2024
15	Centre for Capacity Development in Africa - Stellenbosch University	Howie, Sarah	Director and Professor	12/20/2024

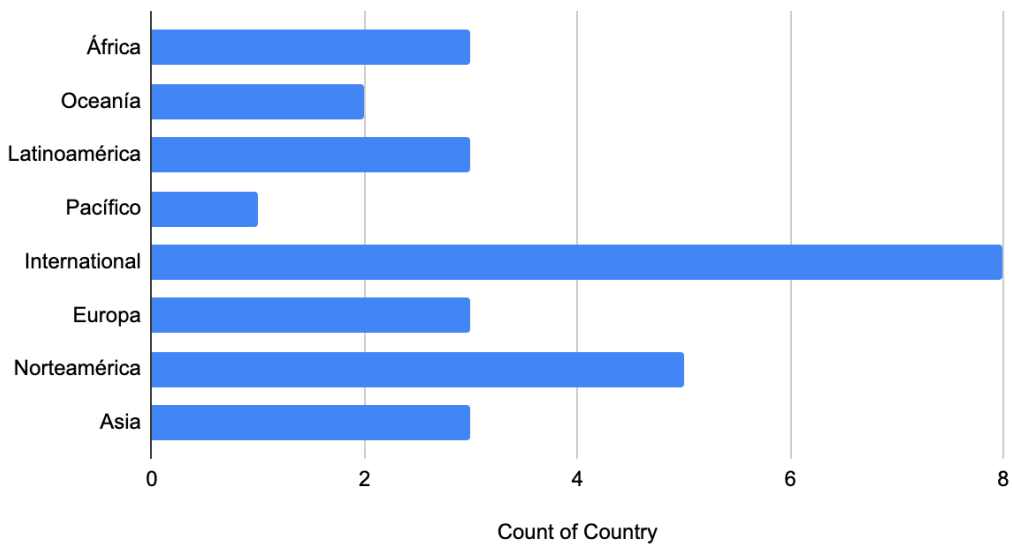
16	UNESCO Global Education Monitoring Report	Kiyenje, Josephine	Senior Project Lead	11/22/2024
17	World Bank	Luna, Diego	Senior education specialist	10/18/2024
18	Inter-American Development Bank (IDB)	Mateo, Mercedes	Director of Education	12/13/2024
19	Kenya National Examinations Council	Ngota, Epha	Coordinator-NAC - KNEC	12/16/2024
20	RTI	Quick, Angela	Senior Vice President, Education Practice Area	10/30/2024
21	Stanford University	Shavelson, Rich	Retired member	11/21/2024
22	Education Cannot Wait	Spoelder, Maurits	Educational planning	10/22/2024
23	RTI	Stern, Jonathan	Director, Education Research and Evaluation	12/5/2024
24	Inter-American Development Bank (IDB)	Suarez, Sonia	Education Specialist	12/13/2024
25	GPE, Data and Evidence	Vivekanandan, Ramya	Senior Education Specialist	10/16/2024
26	OECD	Ward, Michael	Senior policy analyst	10/23/2024
27	ACER Australian Council for Educational Research	Watson, Colin	Chief Executive Officer	11/18/2024
28	International Test Commission	Xiaoming Xi, Madeline	Hong Kong Examinations and Assessment Authority	10/31/2024




Type of agency



Main location





SCOPING STUDY ON A VETTING FUNCTION
AND VIRTUAL FUND FOR LEARNING ASSESSMENTS

Email:

uis.information@unesco.org

uis.director@unesco.org

uis.unesco.org

[@UNESCOstat](https://www.instagram.com/UNESCOstat)