United Nations
Educational, Scientific and
Cultural Organization

UNESCO INSTITUTE FOR STATISTICS

GLOBAL ALLIANCE TO MONITOR LEARNING

SUSTAINABLE DEVELOPMENT GOALS

Options for Reporting against 4.1.1 when using national assessment programs

GAML6/REF/3

Options for Reporting against 4.1.1 when using national assessment programs

July 2019

The ACER Centre for Global Education Monitoring supports the monitoring of educational outcomes worldwide, holding the view that the systematic and strategic collection of data on education outcomes, and factors related to those outcomes, is required to inform high quality policy aimed at improving educational progress for all learners.

Acknowledgements

This document was developed as a contribution by ACER-GEM in support of the UIS-led Global Alliance for the Monitoring of Learning (GAML). The document was prepared by Goran Lazendic, with support from Maurice Walker and Ray Adams.

Contents

Introduction

The sustainable development goals are 17 broad aspirations for social and economic development set by the United Nations in 2015. One of these goals, enumerated SDG 4, is to ensure inclusive and equitable quality education and promote lifelong learning opportunities for all.

The progress towards each goal is measured by a set of indicators. In monitoring progress towards SDG 4, the UNESCO Institute for Statistics (UIS) is working on an approach to measuring indicator 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.

The Global Alliance to Monitor Learning (GAML) is asked to consider that monitoring progress against Indicator 4.1.1 will require measurement of student outcomes at several different stages of learning in a broadly consistent way across education systems, to enable meaningful international dialogue about learning progress and how it may be supported. This is a challenge, given that learning and how it is measured varies across local contexts. Education systems interpret and make decisions about learning, how it is described in curriculum, and how it is assessed and reported.

The central requirement for an international alignment of learning outcomes is for countries to adopt fit-for-purpose approaches to monitoring learning outcomes that support consistency in reporting of international outcomes, while being flexible enough to accommodate national curriculum, pedagogical and assessment approaches, status and priorities. There are several methods of monitoring that meet these criteria. These methods include utilising the results from relevant international or regional large-scale assessments. Each of the appropriate methods report the estimated population who meet or exceed the indicator – known as the minimum proficiency level (ACER, 2019).

For the purpose of reporting against SDG indicator 4.1.1, countries that do not opt to participate in international or regional assessments have a range of further options. These options primarily fall into two categories policy linking methods, and empirically based equating methods. Both broad approaches require expert judgement and are formulated around linking existing or future assessment programs to the minimum proficiency levels.

This paper focuses on *empirical* options for countries wanting to align assessment program outcomes to minimum proficiency levels (MPLs), as described in the concept paper developed by ACER to support UIS and GAML assessment alignment efforts (ACER, 2019). The purpose of this paper is to present a set of empirical assessment alignment methods that can be considered for the alignment of assessment programs with the MPLs, so that reporting for SDG 4.1.1. is possible when countries opt not to participate in a regional or international assessment survey.

Aligning assessments to the minimum proficiency levels

In operational terms, an assessment alignment process will enable education systems to clearly, efficiently and consistently examine and report on the current level of alignment of their assessment programs against the MPLs.

Irrespective of format, an assessment typically has three key elements:

- a description that specifies the learning area and content of an assessment;
- a set of expectations regarding the range and trajectory of learning targeted by an assessment; and
- a reporting scale that provides empirical data regarding the distribution of students' performance measured by an assessment.

In principle, the alignment process should include examination of the assessment program with regard to each of these three elements. However, the availability of equivalent information in the assessment program or lack of resources and expertise to conduct such comprehensive evaluations means that a range of assessment and assessment outcome alignment methods must be offered. These evaluation activities can be grouped into three steps:

- learning area alignment
- policy alignment
- empirical alignment

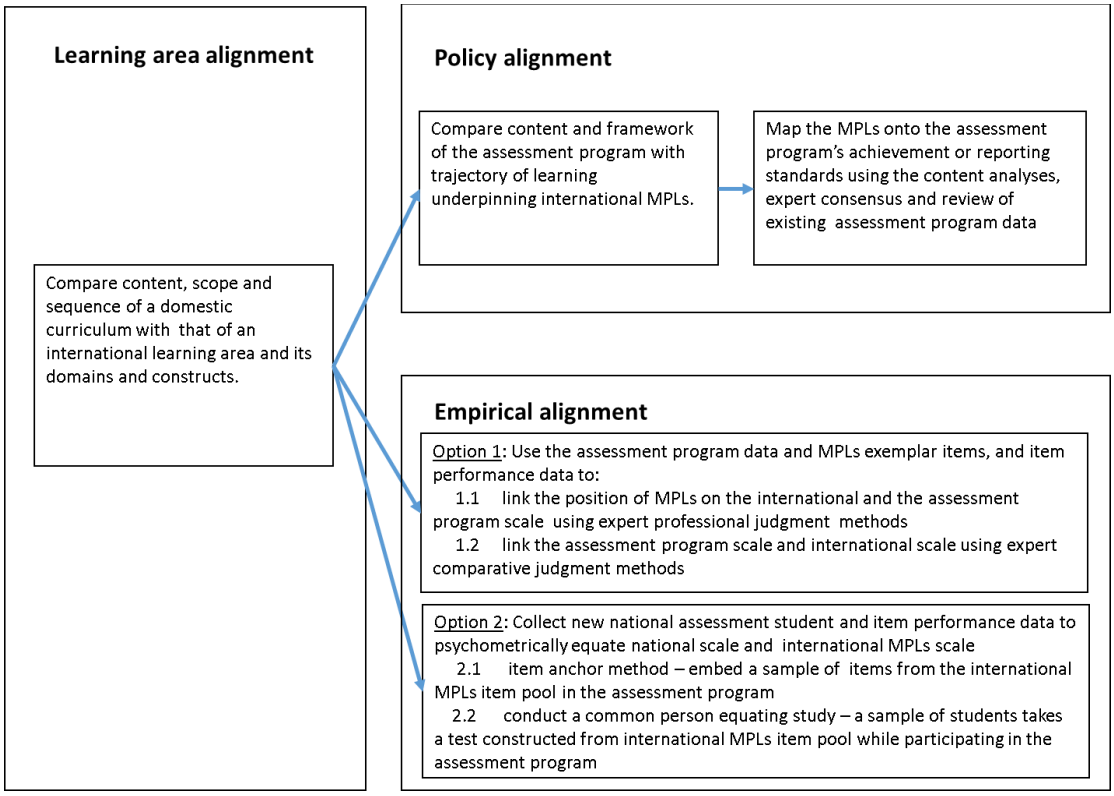
Learning area alignment is the systematic and structured comparison of the domestic curriculum, as a content source of its assessment program, and the learning area, domains and constructs underpinning the corresponding MPLs. The learning area alignment must be the initial step in any assessment alignment efforts. It provides a conceptual base for understanding and evaluating the scope and strength of the assessment alignment and its outcomes.

A policy alignment process is applicable to contexts where only descriptive achievement levels or proficiency standards are used in assessment reporting and educational policy monitoring. Policy alignment will include comparison of the explicitly or implicitly articulated trajectory of learning and progression of students' achievement captured by assessment program with that of MPLs. As an outcome of such comparisons it is possible to assess how well and how deep the existing national standards capture knowledge, skills and understandings contained in and conveyed by the MPLs. Such mapping can be conducted in a structured way following agreed observation and reporting protocols. Once such a correspondence between an assessment program and MPLs has been established, the existing student performance data can be used to ascertain the proportion of students that has attained each of the mapped MPLs.

Empirical alignment is the set of activities that link the assessment program scales and the MPLs. The empirical alignment methods can be classified as methods that do not

require collection of new data on students and items, and methods where it is necessary to collect new student and item data. The former set of methods will provide empirical alignment using different forms of expert judgments collected in the controlled, systematic and purposeful way. The latter set includes two well established psychometric test equating methods: item anchor and common person equating. Figure 1 shows the proposed assessment alignment methods and key logistical requirements.

Figure 1: An illustration of the relationship between three assessment alignment approaches



As already mentioned, the learning area alignment should be always conducted first in order to comprehensively assess the state of the educational system curriculum and assessment, and to inform and support the choice of the assessment alignment method. Policy alignment may be used if it is not logistically feasible to implement any of the empirical alignment methods.

It is important to note, however, that even when an assessment program does not have a formal scale and only descriptive achievement standards are available, one of the proposed comparative judgment methods can be used to conduct the empirical alignment of such assessment.

Empirical alignment

Two sets of the proposed empirical alignment methods are described in this section. The outcome of empirical assessment alignment process is by design quantitative. However, to fully understand the relationship between assessment program and MPLs and thus to accurately assess progress against SDG Indicator 4.1.1, the empirical alignment must be supplemented by systematic and structured comparison of learning trajectories

underpinning the learning areas, domains and constructs of the assessment program and MPLs.

Such qualitative and conceptual comparisons of domestic and international trajectories of learning should precede any empirical alignment as these will provide crucial information to guide alignment planning and implementation. Equally, the learning progress mapping will provide scaffolding in which the quantitative outcomes of empirical alignment can be validated in a structured and transparent manner. A set of processes and protocols should be developed to guide such a mapping exercise to ensure the consistency of its implementation and quality of its outcomes.

These protocols should form an integral part of a set of guidelines and technical specifications that will need to be developed for all proposed empirical assessment alignment methods. These will address planning, implementation, data collection and the reporting phases of the assessment alignment studies. The purpose of these protocols is to enable consistent and transparent implementation of all assessment alignment activities and to provide external and independent quality assurance framework for evaluation of the assessment alignment outcomes and countries progress against SDG Indicator 4.1.1.

Option 1: Empirical alignment – without new administration of an assessment program

This set of proposed assessment alignment methods will obtain new empirical data and evidence using systematically collected and analysed expert judgment data. Two sets of such assessment alignment methods are proposed:

- *Benchmarking:* linking the position of MPLs onto an assessment program scale, using expert professional judgment methods developed for setting standards and benchmarks in educational assessments.
- *Pairwise comparison:* linking an assessment program scale and a scale derived from items at and around each MPL, using expert comparative judgment methods that compare assessment tasks.

Benchmarking (Method 1.1)

The benchmark method will directly link the MPLs' scale parameters and the assessment program scale using expert professional judgment methods. This will be achieved by systematically collecting professional judgments about the relationship between assessment program items and the MPLs' exemplar items, and their ordering on the trajectory of learning set by the MPLs. Once the MPLs' positions are established as the cut-scores on the assessment program scale, then the existing assessment program student performance data can be used to calculate the proportion of students at and above each of the MPLs.

The benchmarking method stems from a large body of research and practical experience in setting the achievement standards and proficiency level cut scores in educational measurement. In terms of standard setting methods that are placing items in centre of standard setting activities (as opposed to student-centric methods) are Angoff (see

Angoff, 1971; Impara & Plake, 1998) and its variants, and the Bookmark method (see Cizek,Bunch, & Koons, 2004; Lewis, Green, Mitzel, & Patz, 1999). An essential feature of the latter method is the use of psychometric scales to place items onto a proficiency distribution where a standard is set.

Consequently, the Bookmark method provides a good fit with the current approach to the developing and setting of the minimum proficiency levels. The core of the Bookmark method is the ordered item booklet – a test booklet in which items are ordered in increasing difficulty. The task of judges is to locate an item in the item ordered booklet that corresponds to minimum expected competency for a student attaining the minimum proficiency level (or any other desired achievement standard). The ordered item booklet approach can be used to empirically link the assessment program to MPLs.

In one version of this novel method an ordered item booklet consisting of item from the assessment program is presented to experts who determine which item in the booklet meets the demand of the MPLs. In another version of this method a booklet of randomly ordered assessment program items is presented to experts who then rank each item relative to the exemplar booklet of difficulty-ordered MPLs example items.

A common factor between two versions of these benchmarking methods is that they require educators and subject experts who understand the test-taking population and assessed content. This is true for both the assessment program and MPLs items. The benefits of this method are that there are well-established practices regarding convening and training experts to participate in such standard setting activities. Furthermore, there is a robust body of research offering a range of methods for examining the reliability and validity of such standard setting procedures (e.g., Nichols, Twing, Mueller & O'Malley, 2010; Hamme, Shultz & Engelhard, 2011).

Pairwise comparisons *(Method 1.2)*

The pairwise comparisons will place the items for the MPLs item pool on the assessment program scale using an intermediate scale established by the comparative judgment methods. In the comparative judgments method the judges will compare the complexity and difficulty of the items from the assessment program and MPLs item pool, two items at a time. The outcome of this series of binary decisions, in which each item is compared to a random sample of other items between 20 and 40 times, are then scaled on to an interval, pairwise scale.

The items from the assessment program will therefore have two set of parameters: one on its original scale and another on the pairwise scale. This information can be then used to estimate the translation equating to place items from the MPLs item pool on the assessment program scale. Once this statistical linking is completed and the MPLs are translated onto the assessment program scale, the existing student performance data can used to calculate the proportion of students at and above each of the MPLs.

Pairwise comparison or comparative *judgement* methods exploit the finding that people are better at comparing two objects or examples of student work against each other, than at evaluating one object or piece of student work against criteria (Thurstone, 1927). Based on multiple comparative judgements, a rank order of tasks or examples of student

work is generated. This rank order is based on all decisions made across judges, and results in relatively reliable scales. Comparative judgments have been applied to a range of the assessments including mathematics, (Jones, Inglis, Gilmore, & Hodgen, 2013), oral language in early childhood (Humphry, Heltzinger & Dawkins, 2017) and writing (Mcgrane, Humphry & Heltzinger, 2018, Heldsinger & Humphry, 2010, 2013; Humphry & McGrane, 2015; Pollitt, 2012).

The advantage of pairwise comparisons is that there is no limit regarding the type and form of the assessment tasks. Furthermore, a large number of items can be can be included in the pairwise comparison as this is a relatively efficient judgment process. Thus, this method can provide robust and reliable empirical linking with MPLs. The robustness of pairwise scaling and statistical linking can be analysed and evaluated using standard IRT fit and reliability statistics.

Option 2: Empirical alignment – requiring new administration of an assessment program

This is a set of the most resource demanding and time-consuming assessment alignment methods that require the collection of new student and assessment item performance data to psychometrically equate MPLs with the assessment program. The purpose of this equating is to establish the exact location of all of MPLs' exemplar items on the existing assessment program psychometric scale, thus establishing an empirical cut-score for each of the MPLs on the assessment program scale. Such a robust equating will therefore provide an opportunity to directly estimate the proportion of students attaining each of the MPLs. Two standard psychometric equating methods could be used:

- *item anchor method:* embedding a sample of items from MPLs item pool in the assessment program
- *a common person equating study:* a sample of students takes the assessment program and a test constructed from the global item pool.

Item anchor method *(Method 2.1)*

The non-equivalent groups anchor test equating design (see Kolen & Brennan, 2004) is a well-established method of psychometric linking in educational assessments. Its key benefit is that all items from the assessment program will be placed on the same scale as one derived from the MPLs' exemplar item pool, thus providing comprehensive examination of the assessment alignment across the range of the exemplar pool, in addition to directly mapping the MPLs' cut-score on to the assessment program scale. Of additional benefit is that there is a well-established and well-researched set of statistical procedures and indicators that can be used to ascertain the robustness and uncertainty of empirical linking.

The key disadvantage of this method logistical burden to the assessment program in terms of the test development, the implementation costs to schools and students of data collection, and the time and costs needed to analyse the data and produce the alignment report. Furthermore, its use is limited to assessment programs that are similar in form to a set of national, regional and international assessments used to develop the MPLs' scale. A significant difference in item design and the interactions these items elicit from students

might lead to significant differential functioning of anchor items relative to assessment program's original items, thus producing biased or uninterpretable outcomes.

Common person equating *(Method 2.2)*

The common person equating or a single group equating design requires that the same students take a domestic test and a test that constructed using MPLs items. The key assumptions of this assessment alignment method is that it assumes that interactions with items that might affect student performance on national test when compared to MPLs tests are negligible.

This assumption represents a key disadvantage of common person method, particularly in relation to test order effect and its impact on student's engagement levels. The sample size requirements could be higher than that of the item anchor method, particularly where the domestic student population is heterogeneous.

Choosing an appropriate method to report assessment results against SDG 4.1.1

Participants, procedures, materials and student data requirements

This section provides an overview of key logistical resources regarding student and item performance data, materials, procedures and participants requirements for the four prosed empirical assessment alignment methods presented in this section.

Student level data

In providing this overview, it is assumed that there is an agreed method to determine the segment of student population targeted by each of the MPLS population. Such a segment of overall population is referred to as a *targeted population* in the rest of the paper. It is also understood that in each of the assessment alignment activities more than one targeted population might be subject of alignment activities.

In this paper the resource estimates are provided for a single targeted population assessment alignment study and thus total resource estimation should be increased if more than one tarted population is included in the assessment alignment study. However, some efficiencies in sharing the data, materials and participants across alignment activities for different targeted populations can be expected.

Both of the two empirical alignment options will make use of the assessment program student-level data. However the set of options that use the expert judgments require access only to the student level performance data to conduct the alignment and to evaluate the alignment outcomes.

In contrast, the set of psychometric equating will require firstly the collection of the new student-level data, and secondly, the rescaling of existing student responses. Therefore, the student-level data containing each student's item response string is required to conduct these assessment alignment activities. The existing student-level performance data will also be required, but only to estimate and evaluate the alignment outcomes. The

sample size and the scope of the new student-level data collection will be outlined for each of the two equating alignment options.

MPLs' Item pool

This paper assumes that a pool of items exemplifying progression of knowledge, skills and understandings required to attain each of the MPLs, including the regions closely-below and at-and-closely-above the cut scores for all MPLs, will be compiled and offered as an international resource. The MPLs' item pool will contain information necessary to conduct all proposed assessment alignment activities, including but not limited to item content and learning progressions metadata, publishing-ready images, scoring rubrics, and psychometric parameters and data. Different assessment alignment options will make use of some or all the elements of this global item pool resource.

Finally, a sufficiently large and representative sample of items from the international pool will need to be kept confidential to enable implementation of the proposed empirical assessment alignment studies.

Option 1.1 Benchmarking requirements

Participants

The participants will be comprised of subject experts from the host country. The research and practice has shown that a diverse sample consisting of teachers and stakeholder representatives (school and government administrators and policymakers) ensures that the standard setting process is sufficiently sensitive to and can account for needs of students and the education policy requirements. Hambleton and Pitoniak (2006) found that a sample size of 15 to 30 participants provides sufficiently a diverse and robust sample of experts for these item-centred professional judgment activities. The key condition is that is that all participants have good knowledge and understanding of national curriculum and assessments as well as that of the targeted student population. Ahead of the assessment alignment workshop, the study organisers should ensure that participants are knowledgeable of the learning area, domains, constructs and exemplars underpinning the MPLs.

Procedures

The benchmark assessment alignment activities are typically organised and implemented by the professional leaders as a multi-day workshop activity. In terms of steps, the participants have to receive the training in understanding and unpacking of a targeted MPL, training in judgment procedures, to make and record their judgments. An important briefing session also takes place at the end of these activities, which provides opportunities for participants to share impressions and observations that give significant insight for understating and interpreting assessment alignment outcomes. A three-day workshop is typically sufficient for these activities.

Materials

Materials include printed booklets and self-contained forms of items from the domestic and international item pool, including all relevant item stimuli and resources where applicable. Item metadata and performance information is necessary to prepare these materials and to implement the assessment alignment study. In terms of the number of items, the item section for the assessment program should be representative of the

assessment program test blueprint and content. Where possible, items from multiple assessment program cycles should be used to increase the sample of items across the whole scale range. Items from the international item pool will be included based on the same principle, with an increased number of items in the region of the target MPL and MPLs that precede and follow the target MPL.

Student data

Existing student-level test performance assessment program data will be used to evaluate the outcome of the assessment alignment studies and to ascertain the proportion of target population attaining the target MPL.

Option 1.2 Pairwise comparison requirements

Participants

The participants will be comprised of subject experts from the host country. The sample size requirements are closer to the lower end of 15 to 30 participants range suggested by Hambleton and Pitoniak (2006). A mix of teachers and stakeholder representatives is desirable, but not essential, as each participant makes a large number of decisions and it is thus possible to monitor and control any potential judgment bias in a way that is not possible in other expert judgment based alignment activities. The participants must have good knowledge of the assessment program and MPLs' item scoring protocols and processes. Familiarity with the progression of learning underpinning MPLs is desirable but not necessary.

Procedures

Participants' presence at the central location is required only for the induction and task training sessions which can typically be completed within a single day. The pairwise comparisons require specialist software; therefore access to computers and internet is necessary for this study. Pairwise comparison software is typically delivered via a web application and thus has low IT specification requirements. Where internet access is readily available the participants can complete the study at their computers in their usual work environment. Alternatively, a marking centre could be organised using existing marking centre facilities or even school facilities such as a computer laboratory or library. Black and Bramley (2008) found no difference in the outcomes of English paper comparative judgments between remote and in-centre participants' location.

The pairwise software handles the delivery and data collection for this assessment alignment study. This is an efficient method of empirical data collection and participants can process a relatively large number of items in a single day.

Materials

Online-print ready copy of all items from the assessment program and the MPL global item pool in common graphic format (e.g., PDF, JPEG) comprise the materials. Where items require a stimulus (e.g., a reading prompt for an item in reading test) the stimuli must be included in the item file.

Pairwise comparison can accommodate a relatively large number of items. Thus, it is recommended that all or the widest possible range of items from the assessment program and international item pool are used in the study. Items from more than one assessment program cycle should be included where available.

Item difficulty parameters are necessary to ensure that each item is paired with items within a similar range of difficulty to avoid trivial comparisons.

Student data

Existing student-level test performance assessment program data will be used to evaluate the outcome of the assessment alignment studies and to ascertain the proportion of target population attaining the target MPL.

Option 2.1 Item anchoring method

Participants

A random sample of 400 students represents a minimum sample size needed to obtain sufficiently stable IRT item performance estimates if a single test form is used. IRT item parameter estimation procedures are relatively robust and thus a convenience sample can be used if there are logistical difficulties in selecting and accessing a random sample of students. The robustness of the convenience sample can be increased by having a good cross section of students in terms of school and socio-educational background, and by randomly selecting 20 students from a school instead of including a whole class in the sample.

Where a country uses the spiralled test form or item block design then the sample size must be increased to ensure that 400 responses per item are captured.

Procedures

The assessment should be administered to students under the same conditions, protocols and procedures as the assessment program. It is desirable that external invigilators observe the study implementation to ensure the secure distribution and collection of tests, materials and responses.

Materials

The test prepared for this assessment alignment study will consist of approximately 75% domestic items and 25% items from the international pool. The selection of assessment program items should mirror the assessment program as close as possible in terms of content and composition, the range of item difficulty, and the distribution of items types Noting that the anchor test with restricted difficulty range, relative to the difficulty range of the domestic test, will still provide sufficiently robust equating outcomes (see Sinharay & Holland, 2007).. The length of the anchor item set will vary between the school grades of the targeted population. For example, the typical length of the tests in primary school grades does not exceed 40 items. So a set of 12 MPL anchor test items may be sufficient when a single MPL is placed on the assessment program scale. The number of items in the anchor set will progressively increase as proficiency demands of MPLs increase. The total number of items needed will grow by a factor of number of MPLs targeted in the alignment study.

Student data

Item response data from existing student-level assessment programs are required. These data will be rescaled following the equating of the assessment program scale and the new set of student ability estimates will be produced to determine the proportion of the target population that has achieved the target MPL.

Option 2.2 Common person equating

Participants

A two-stage stratified sample of the size that provides sampling accuracy equal to a simple random sample of 400 is required. For example, with a cluster size of 20 students and a moderate interclass correlation of 0.2, the required sample size is 1920 (see Ross, 2005). The assessment program should be used to calculate the actual interclass correlation and the sample size adjusted accordingly.

Procedures

The assessment should be administered to students under the same conditions, protocols and procedures as the assessment program. The order of domestic and international tests must be balanced across the clusters.

Materials

The intact version of the assessment program and the international MPLs' items test of a similar length will be used.

Student data

Item response data from existing student-level assessment programs are required. This data will be rescaled following the equating of the assessment program scale and a new set of student ability estimates will be produced to determine the proportion of target population that has achieved the target MPL.

Resource capacity and data readiness

In order to ascertain the feasibility of proposed empirical assessment alignment options, each country will need to assess the internal expertise or capacity to procure expert advice and services in relation to key elements of the proposed assessment alignment studies.

Common to all options is the availability of the assessment program data. This includes items, items' metadata, and performance information.

For the assessment alignment methods that rely on expert judgments, some key questions are:

- How well are the concepts of learning progressions and criterion-referenced assessments understood by the educational community?
- What is the level of assessment literacy among educators, administrators and policymakers?
- What is the level of formality, expertise and experience regarding the standard setting procedures?
- What is the level of formality, expertise and experience regarding assessment scoring and marking? Are there established marking operations in the country, whether permanent or seasonal?
- What is the capacity to organise and sustain the necessary workshop activities?
- Do educators and administrators have access to computers and internet?
- What is the status of the assessment program's item bank, item performance and metadata database?

Table 1 provides a comparison of the of the organisational, professional and logistical resource demands between two expert judgment empirical alignment methods.

Table 1

Requirements	Method 1.1 Benchmarking demand	Method 1.2 Pairwise comparisons demand
learning progression and criterion-referenced assessment understanding	High	Low
overall level of assessment literacy	High	moderate
expertise in standard setting activities	Moderate	Low
expertise in assessment scoring	Moderate	moderate
judges training	High	Low
implementation logistics	centralised - high level of supervisor	decentralised -low level of supervision
access to computers and internet	not required	essential
data capture and cleaning	high – to transcribe and collate judgment data	low – judgment captured by a system
psychometric expertise	low to moderate	moderate

For the assessment alignment methods that use psychometric equating methods it is important to know:

- the level of formality, expertise and experience regarding the collection, processing, cleaning and storing of assessment data?
- the measurement model is used to scale and report data for the assessment program? Are IRT item difficulty and person ability estimates available?
- whether tests within assessment program are vertically scaled across different school grades?
- whether tests are longitudinally equated across different assessment program cycles?
- the level of psychometric expertise and experience available, either internally or externally, to the organisation in charge of the assessment program?
- the status and availability of student-level item response data?
- the logistical and policy capacity and school burden to conduct the new assessment data collection or to expend the current assessment program data collection activities?

Table 2 provides a comparison of the of the organisational, professional and logistical resource demands between two expert judgment empirical alignment methods.

Table 2

Requirements	Method 2.1 Item anchoring demands	Method 2.2 Common person equating demands
new assessment material development	low	moderate to high
sample size	Low	moderate to high
implementation logistics	low to moderate	moderate to high
data capture and cleaning	low to moderate	moderate to high
psychometric expertise	high	high

These questions and summaries should be taken only as indicative and illustrative. The final list of resource considerations and demands for each the proposed empirical alignment methods will need to be developed in conjunction with the planned set of technical protocols and specification for the empirical assessment alignment with SDG Indicator 4.1.1.

References

ACER. (2019). *SDG 4.1, Indicator 4.1.1: Minimum Proficiency Levels described, unpacked and illustrated, Version 2*. ACER. Camberwell.

Angoff, W.H. (1971). Scales, norms and equivalent scores. *Educational Measurements*. Washington, DC: American Council on Education.

Black,B. & Bramley, T. (2008). Investigating a Judgemental Rank-Ordering Method for Maintaining Standards in UK Examinations. *Research Papers in Education, 23*(3), 357-373.

Cizek, G.J , Bunch, M.B. , Koons, H. (2004). A NCME instructional module on setting performance standards: Contemporary methods. *Educational Measurement, Issues and Practice*, *23*, 31–50.

Hame P., C. , Schulz,M. Engelhard, G. (2011). Reliability and Validity of Bookmark-Based Methods for Standard Setting: Comparisons to Angoff-Based Methods*. National Assessment of Educational Progress. Educational Measurement: Issues and Practice. 30.*

Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), Educational Measurement (4th ed., pp. 433-470). Westport, CT: Praeger.

Humphry, S., Heldsinger, S. & Dawkins, S. (2017), A two-stage assessment method for assessing oral language in early childhood. *Australian Journal of Education, 61*(2), 124-140.

Impara, J., & Plake, B. (1998). Teachers' Ability to Estimate Item Difficulty: A Test of the Assumptions in the Angoff Standard Setting Method. *Journal of Educational Measurement, 35*(1), 69-81.

Kolen, M. J. & Brennan.R. L. (2004). Test Equating, Scaling, and Linking: Methods and Practices (2nd ed.). New York: Springer.

Lewis, D. M., Mitzel, H. C., Green, D. R., & Patz, R. J. (1999). *The bookmark standard setting procedure.* Monterey, CA: McGraw-Hill.

McGrane, J. A., Humphry, S. M. & Heldsinger, S., 2 (2018). Applying a Thurstonian, Two-Stage Method in the Standardized Assessment of Writing. *Applied Measurement in Education, 31*(4), 297-311.

Nichols, P., Twing, J., Mueller, C. D., & O'Malley, K. (2010). Standard-setting methods as measurement processes. *Educational Measurement: Issues and Practice, 29*, 14–24.

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-forms of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*, 249–275.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review 34*(4), 273-286.