Linking Approaches
Working Paper for endorsement

Linking to a proficiency level scale

The linking of national and/or regional assessment to the global definitions would require in-depth enquiry into the items of the national and/or regional assessments. Linking is the general terms used to relate test scores on one test/form to another test/form. Different researchers have proposed different approaches and ways to link assessments. But, in general terms, linking tries to moderate differences between tests that were designed for completely different purposes to express them in the same scale in a way that allows some degree of comparability. In turn, this allows for fair inferences about the subjects (countries) compared. The process of making different tests comparable is in general denominated as "moderation".

Statistical moderation, based on the same sample of people taking two tests that is called the test-based approach, utilizes the score distribution of two assessments to construct concordance tables mapping the scores on two tests that do not measure the same constructs. Methods used for linking tests could be classified as equating, calibration, projection and moderation. Others classify into equating, scale aligning and predicting. Methods such as <u>calibration</u> (putting items and persons taking the one test form onto the same scale and setting a reference point), <u>equating</u> (putting different tests on a common scale, removing unintended differences in test form difficulties and setting up a common scale) are used alternatively but they are also other ways of linking. It is important to keep in mind the strength of linking depends on the assumptions made on the degree of similarity between inferences, constructs, populations and measurement conditions.

Non-statistical moderation has the same objective as statistical moderation but is obtained by matching up definitions of test by subjective judgement. In general, described as "social moderation", it uses judgement to match levels of performance of different assessments to a reference definition, which in the case of indicator 4.1.1 the reference definition is the Global Content Framework. Thus, social moderation calls for direct judgement about the comparability of performance levels between different assessments onto a reference scale.

As statistical moderation is based on comparability at a certain point in time of certain set of items or performance of individuals, social moderation comparability comes from the opinion of a group of people as the social moderators rather than a set of students or items at a certain moment in time. Nobody could solve the uncertainty of many of these choices (items, students, moderators) and there is always some subjectivity.

However, social moderation or policy linking could serve to define (and establish) broad standards for the knowledge and skills that students have to achieve. It can also monitor performance and help to understand the meaning of a minimum level of what students are expected to know and be able to do in relation to grade-appropriate contents. This lies at the heart of the curricular definitions in any country.

This "moderation" or linking is not an application of the principles of statistical inference but a way to specify the rules of the game. Establishing the rules of the game would help to establish agreement for comparing students that is not on information from tests that are built to measure different construct. Consensual processes and experts inputs are use as the way forward.

The proposals for linking to a common scale are not mutually exclusive and proposals described here are a combination that aims to establish some rules for comparing students, youth and adults. Alternative strategies to achieve comparability and assessing their effectiveness and efficiency are a matter of proof.

Scope of work of the UIS

a.   to define a set of cost-efficient linking strategies to maximise coverage in reporting

b.   to define an immediate/interim solution to reporting.

The UIS has taken a portfolio approach that includes two broad sets of possibilities: the non-statistical approach and the statistical approaches that rely on "hard-core" psychometric evidence to define comparability. ***Figure 1.*** Summarises the options below.

Strategy 1. The non-statistical approach: pedagogically informed recalibration of existing data

The approach involves using the proposed proficiency framework that describes the range of competencies that children/youth have at each level to locate proficiency levels from alternative assessment programmes based on the Performance Level Descriptors (PLDs). The approach is denominated social moderation (or policy linking) as linking is guided by experts' judgement. This proposal would allow the expansion of coverage in terms of educational systems reporting for SDG 4. For instance, coverage at the primary level would double, in terms of the population-weighted world, if national assessments were included.

Strategy 2. The statistical approach

2.a. Psychometrically informed recalibration based on common items

✓   implies the use of common items in different assessment programmes

✓   one version has been proposed by the Australian Council for Educational Research (ACER) as part of an overall proposal of progression in learning but options are not exhausted.[1]

✓   has proven to face many difficulties in implementation from technical to political.

2.b. Recalibration by running a parallel test on a representative sample of students

✓   The IEA outlines the "Rosetta Stone" solution that deals only with the primary level and allows two assessments, one international, the other regional to be expressed on the same scale. Concretely, the proposal states that sub-samples of students in three to five countries per programme would write not just the regional tests, but also IEA's test.

✓   This would produce a "concordance table" based on psychometric modelling.[2] The table is not the reporting scale but it facilitates by expressing a larger number of countries in the same scale.
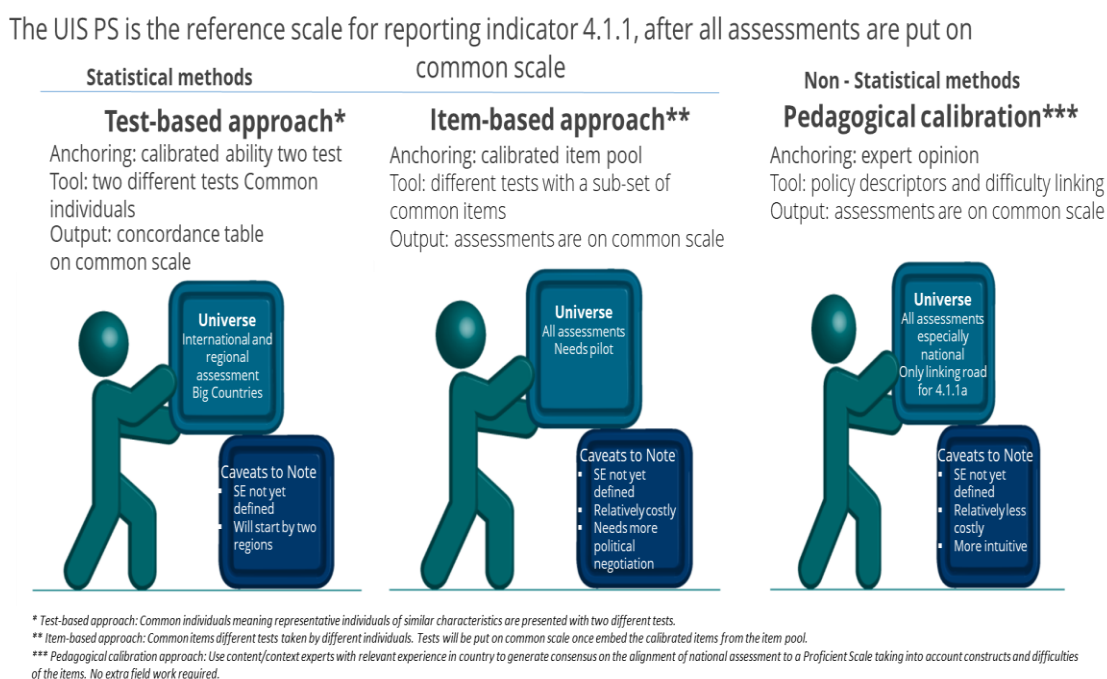
---

[1] Note that the reference scale is built from items coming from various assessments.
[2] For countries the option is to either participate in a regional programme or in a global programme (something that might be difficult or not possible if the region does not have a regional initiative).

2.c. Recalibration of existing data

✓ This approach relies largely on statistical adjustments[3] taking advantage of the fact that some countries, referred to as "doubloon countries", participate in more than one cross-national programme. Using several such overlaps has allowed for the identification of roughly comparable proficiency thresholds. It could serve as a validation but it is unlikely to have political buy-in.

Figure 1 - Linking strategies

The UIS PS is the reference scale for reporting indicator 4.1.1, after all assessments are put on common scale

* Test-based approach: Common individuals meaning representative individuals of similar characteristics are presented with two different tests.
** Item-based approach: Common items different tests taken by different individuals. Tests will be put on common scale once embed the calibrated items from the item pool.
*** Pedagogical calibration approach: Use content/context experts with relevant experience in country to generate consensus on the alignment of national assessment to a Proficient Scale taking into account constructs and difficulties of the items. No extra field work required.

Source: UIS

Weighing options

The efforts described in ***Table 1*** should be taken more as complementary routes than as alternative options in order to minimise risk if some of the approaches prove to be too costly, the margin of error too high, politically unfeasible or a combination of all these. The strategies help each other to build a sustainable reporting strategy where it is easier to see stepping stones between Strategy 1 and Strategy 2a and complementarity between Strategy 2b and Strategy 1, such as the Rosetta Stone which needs to be expressed in a proficiency framework. Strategy 2c has a potential use as a check to compare statistics based on national assessments (Treviño and Ordenes, 2017). [4]

---

[3] See Altinok, N. (2017).

[4] A third strategy could be a new test that everybody takes for reporting using a common comparable tool but this is neither politically feasible nor cost-efficient so it has not been pursued.

Table 1. Relationship between linking strategies and coverage of assessment type

	Statistical linking			Pedagogical linking
	Recalibration through parallel tests	**Psychometrically informed recalibration**	**Statistical recalibration of existing data**	**Pedagogically informed recalibration**
PISA, TIMSS and PIRLS	Will be used	Could be used	Yes	Yes
Regional cross-national assessments	Will be used	Could be used	Yes	Yes
National assessments	Could be used	Could be used	Not clear how	Yes
National examinations	--	--	Not clear how	To be used

Source: Gustaffson (2018).

Objective

Given the variation in learning assessments in turn of content, construct, and methodology used for data collection and reporting, the portfolio approach aims to have an extensive look at various ways of linking existing assessments.

Decisions for plenary endorsement: Linking Portfolio

Please provide your feedback by completing the questions that follow. Thank you.

Your name (please print): _____

Name of your organization: _____

1.	Do you agree to use statistical and non-statistical approaches in linking learning assessments?	YES	NO
2.	Do you agree with strategy 1 - The non-statistical approach: pedagogically informed recalibration of existing data (or policy linking), as an option for linking in		
a.	Grade 2/3?	YES	NO
b.	End of Primary?	YES	NO
c.	End of Lower Secondary?	YES	NO
3.	Do you agree with strategy 2a – The statistical approach: Psychometrically informed recalibration based on common items (or item-based approach), as an option for linking in		
a.	Grade 2/3?	YES	NO
b.	End of Primary?	YES	NO
c.	End of Lower Secondary?	YES	NO
4.	Do you agree with strategy 2b – The statistical approach: Recalibration by running a parallel test on a representative sample of students (or test-based approach), as an option for linking in		
a.	Grade 2/3?	YES	NO
b.	End of Primary?	YES	NO
c.	End of Lower Secondary?	YES	NO
5.	Do you agree with strategy 2c – The statistical approach: Recalibration of existing data, as an option to validate the data?	YES	NO
6.	On a national level, do you agree that you can use different approaches to report on different education levels? For example, Country X can use social moderation (policy linking) to report on Grade 2/3 for Indicator 4.1.1, but the test-based approach to report on End of Primary for Indicator 4.1.1.	YES	NO
7.	On an international scale, do you agree that countries can use any approach to report on an education level? For instance, Country X can use social moderation (policy linking) and Country Y can use test-based approach to report their data on End of Primary for Indicator 4.1.1.	YES	NO

We would appreciate any comments that you wish to make:

Comments: