SDG indicators deserve the best methodologies: IEA cautionary statement on political linking

International adoption of the UN's Sustainable Development Goals (SDGs) marks a watershed moment for the future of our world. SDG 4, which aims to "ensure inclusive and equitable quality education and promote lifelong learning opportunities for all," recognizes the vital role of education in society.

Measuring progress towards achieving this goal is of critical importance, and IEA believes this should be done with great care. IEA fully supports the UN's 2030 Agenda and affirms that the measurement of progress toward SDG 4 should be based on agreed international standards and, wherever possible, existing measures. Our common goal is to ensure a timely global response to achieving the Agenda targets.

Regarding indicators of student levels of competency, one of the challenges is to be able to build tables of correspondence between different tests in order to construct indicators that are comparable from one country to another. Three methods are currently being considered: policy linking[1] (discussed here), item linking using calibrated item banks, and student linking with the Rosetta Stone[2] approach, which compares the results of the same sample of students on different standardized tests.

IEA would like to draw attention to the methodological weaknesses of the first method, policy linking. IEA recognizes political linkage as a method to investigate the item difficulty of specific items and provide an approximation of the difficulty of an assessment, but not to make serious international comparisons.

Firstly, the very principle of this method is questionable when it comes to international comparisons. This approach is inspired by so-called "standard setting" methods (also sometimes called Angoff methods) and consists of **relying on a panel of experts** who judge the difficulty of test items in order to set up thresholds on a scale according to specific standards. Being dependent on the judgement of the particular group of experts chosen and those leading the group, **these methods are vulnerable to experts' opinions**, which can differ substantially, running counter to any scientific and objective approaches.

Indeed, **these methods rely on subjective elements, such as national or political agendas**. The interpretation of the same framework is very different from one country to another. However, the

[1] http://tcg.uis.unesco.org/policy-linking/

[2] https://www.iea.nl/studies/additionalstudies/rosetta

ambition of the SDGs is to set targets on indicators whose definition does not depend on country or political agendas, but rather can be guides for all countries, regardless of the widely varying contexts.

In addition, this approach sets aside a very important element: **an assessment is more than a set of items**. Items are bundled into booklets and the assessment includes guides and manuals for all procedures used in the study (sampling, translation of items, test administration, scoring, data entry, etc.). The way items are administered can make a substantial difference to the probability of solving an item correctly. Aspects that need to be considered include:

- Motivation of students to do well on the test—especially high stakes vs. low stakes assessments.
- Timing of the test: the amount of time given to solve the items and the total test length.
- Administration mode of the test—especially computer-based assessment vs. pencil-and-paper.
- Appropriateness of the language used in the test and familiarity of the students with the terms used.
- Embedding of the items—especially if items appear at the end or at the beginning of the test but also which items were administered before each item.

To give an example from IEA's Trends in International Mathematics and Science Study (TIMSS):

TIMSS bundles items to blocks and the blocks of items are assembled to booklets in such a way that each of the blocks appears in the beginning as well as at the end of the test to accommodate for position effects. For TIMSS 2007, the test design was modified and the test became somewhat longer than in previous cycles, which resulted in significant differences in item parameters if items appeared at the end of the test.[3] This was taken into account when scaling the data and reliable trend estimates could be established.

This example shows that item difficulties and results from assessments depend on many different aspects. It is undoubtedly agreed that the probability of items on high stakes assessments will be higher than on low stakes assessments. **Judging individual item difficulties cannot lead to reliable comparisons of test results**.

Of course, these methods may seem attractive because they are quite inexpensive and very easy to implement. However, if used, **their methodological weaknesses would call into question the validity of the SDG indicators and therefore the SDG approach itself**. It would be a great shame to sacrifice the ambition of the SDGs by choosing the poorest methodological path.

IEA strongly recommends not to use political linkage in order to report results and trends on SDG 4 based on assessments that are not internationally or regionally standardized.

[3] See pages 264 ff in https://timssandpirls.bc.edu/PDF/t03_download/T03_TR_Chap11.pdf

IEA

Researching education, improving learning