unesco
Institute for Statistics

"Filters" and requirements for database setting benchmarks and precursor skills weights

The TAG agreed to carry out a benchmarking analysis to fill out the table that has been discussed in various communications.

The aim of this note is to ask for collaboration from those with providers of assessment databases to:

1. To find out how many databases are well-conditioned enough to be used to set benchmarks and weights, using any "sufficiently rigorous" method mixed with expert judgment.
2. And then ask that the data be sent by the time agreed in the cover note accompanying this file.

These are ***logically*** distinct steps though they could be simultaneous in time.

The note does two things: set 1) filters on the data, and then 2) request that the data be well-structured or well-conditioned so as to make it easier to analyze, and 3) a decision tree for how to use a combination of existing benchmarks, newly analyzed data, and expert judgment for setting the benchmarks.

As an aide memoire, the table that has been proposed in various communications (e.g., e-mail from Luis Crouch on 12 Feb 2024) (and with the column for weights added) was:

Skill (or domain or subconstruct)[1]	***Weight***	***Benchmark lang 1 (could be lang group)***	***Benchmark lang 2 (could be lang group)***
Listening comp	Y	X	X
Letter sounds	Y	X	X
Accuracy and/or fluency	Y	X	X
Reading comp	Y	X	X
Comprehensive score	NA	Weighted sum Y*X	Weighted sum Y*X

Note: this would make the comprehensive score become compensatory, but weighted. The TAG provided enough "sense of the meeting" that this is worthwhile, and provided some examples of how it could be done, such as DIBELS in the US. As for linking from one type of benchmark to

[1] Important: this list is not exhaustive. Furthermore, the analysis itself will show the ones that are most closely correlated with RC and with each other, so that the experts can focus on the most robust ones. The role of the GPF and MPL also will be considered.

another (e.g., RC to ORF), there is also Ferdous and Muller (2024), and there are examples in the literature of how this sort of thing can be done, such as here. One needs to mind that one cannot naively mix rates and percentages and IRT scores, but one can use IRT to equate according to the parameter of latent ability in IRT. One also needs to mind that such a comprehensive score may not have much meaningful utility at the classroom level or for lesson design, but it can be useful as a generalized ranking tool. All this can then be combined with expert knowledge.

Filters

1. Nationally representative sample, in order to guarantee spread of skills (and not because we are reporting, yet)
2. Clarity on stop rule and data that follow especially on any oral reading passage and associated comprehension questions. (Such as whether students were allowed enough time to answer all questions, even if there was a marker at 60 seconds to detect fluency.)
3. Must provide the codebook or a "meta" codebook for each of the country application that are provided in the database.
4. Must provide any existing analysis for each of the country applications that are provided in the database.

Dataset condition

The idea behind asking for a well-conditioned dataset is to make the analysis easier. But the analysis would be done separately for each country and language, by using the identifying fields mentioned in point 5 below. We are hypothesizing that in providing the "dossiers", various institutions organized their datasets in such a way that it'd be easy to provide these data.

1. Follow a convention on "omitted" or "not-measured" (blank), missing (. or .. or NA), true zero (0), and imputed zero or non-eligible because of the stop rule (could be 99999). Those conventions are suggestions but make it clear in the codebook (see point 4 above) what these are.
2. Provide a standard field name for all skills, from each "family" of assessments (e.g., the field for Oral Reading Fluency should always be called ORF in EGRA).
3. If there are items that are actually lists or sets of words, it would be beneficial if in addition to reporting the fluency or accuracy, there could be a field next to the fluency or accuracy variable that states the number of words in the passage.
4. Have exactly the same number and order of fields for all country applications of the assessment in question:
    a. If a given country case of the assessment did not assess a given skill, just leave that entire field blank (non-measured)
5. All assessments for any one family (EGRA, MICS/FLS, say) are in one flat file, but with standard fields that would state: a) country, b) grade, c) year, d) LOI (if avail), e) lang of assess.
6. Metadata on gender and age or grade (if HH survey), and ideally on LOI, home language, and language of assessment.