Single universal assessment? Or scales for comparison?

CIES Conference
Mexico City
March 2018

Luis Crouch
RTI International

Session:
Constructing Global Metrics of Learning: Are They Possible?

Background paper
https://shared.rti.org/content/universal-assessment-and-bottom-pyramid
Bottom line

- Answers to previous questions
- Not to be facetious…
- Depends on what one means by “metric”
- One extreme: Do we mean single global test?
- But this is only the most extreme: paper tiger?
- And if that’s what we mean, then everything becomes quite iff.

- Do we mean *some* form of comparability?
- Then: we have a gamut
- We have more choices

- To anticipate bottom line: single global assessment is probably not technically possible, not politically feasible, and not necessary
- But there are other interesting options
Motivation
Concern with global cognitive inequality, need for accurate measurement “at the bottom”
Tradeoffs and issues with notion of “a single assessment”
Reasonable compromises – for now
Conclusions
Motivation

- Governments are not always benevolent or honest with their own citizens
- Some form of globally-vetted, and globally reported measurement might help
- Helped in the battle to get kids in school

- But we have a huge gamut of choices
  1. No reporting at all
  2. Reporting w.o. comparability, and accepting whatever country sends
  3. Reporting w.o., comparability, and accepting what country sends, but footnote/qualify
  4. Reporting w.o. comparability, accepting what country sends only if meet X standards
  5. Reporting w. comparability, some form of equivalency/linking
  6. Reporting w. comparability using a regional or global assessment of choice
  7. Reporting w. comparability according to an evolved single global assessment

Want me to guess where “we” will end up? Most likely an evolution along the arrow.
Special foci of measurement?

- A concern with “bottom of the pyramid issues”
- Some hints that by paying attention to “bottom of pyramid” is how countries’ averages progress
Motivation: within- and between-country inequality

- “250 million not learning” – what does that mean?
- Median poor country way below 5th percentile in OECD

**Comparative Performance at Same Percentiles PISA Reading 2015**

And, the worst off in PISA are very far from the worst off in the world. (True, the best off are the 10 best off.)

5th percentile is close to “special needs.”
But not just equity: the path up is through the middle?

Percentage of Students by TIMSS Proficiency Level
(countries ranked by average proficiency)
But not just equity: the path up is through the middle?

Percent at each proficiency level, PISA 2015, reading

Download of PISA Data
And, about half the global inequality is **within-country**.
Single national measurement?

- If countries make progress from the bottom by addressing the bottom of their distributions...

- If half of global inequality is within countries, and the inequality between countries is hard to address with education-specific policies...

- What does this say about measurement?
  - SDGs require tracking % of students minimally proficient
    - For this to be meaningful, “proficient” should mean more or less the same to everyone
  - Education Commission called for “global lead indicator”
    - Idea did not progress yet, but probably won’t die

- What do these factors together imply for measurement?
  - For a single global assessment?
Accuracy at the bottom: PIRLS

Figure 5. Accuracy of PIRLS Scale in two countries, and at different levels within each country

Mullis, Martin, and Foy (2013)
TIMSS information function at the bottom

Test Information Function: TIMSS (math grade 8)

Sandefer (2016)
Accuracy declines with % correct—but % correct = f(income)

- IEA experts: inaccurate below 30% correct— but I extrapolate that to be A LOT of countries.
Normal items, low-performing country  
Easier items, low-performing country

PISA consultants: are even the easier items easy enough? See info function.

Adams and Cresswell (2016)
Conclusion thus far

- Single global assessment perhaps **not even possible**
- There may be limits to how “easy” you can make these assessments and still have them useful for the top or even the middle
  - So how likely is a global test?
- Not surprising given how low performance at the bottom is
  - In principle, super-long assessments, or with computer adaptation to difficulty, are possible, and could solve some of the problem
    - But how likely how soon in the poorest countries?
    - And how long can one “tolerate” the assessments to be?
- Do we need assessments oriented at the bottom?
“Bottom” oriented assessments

- ASER, EGRA, etc.: started as simpler assessments, not that informed by psychometrics and IRT
- But well-informed by:
  - assessment practices on precursor skills
  - “Reading science” (or pre-Reading science)
- As popularity spread, curiosity: how rigorous?
- Classical measurements of reliability
  - Some pretty good, others need improvement
  - Some measures of concurrent validity: not bad
  - Not much on predictive validity (but then who ultimately does? Does PISA predict personal income?)
  - Inter-rater: can be pretty good
- Other areas of concern:
  - Reliability/comparability over time: critical even if no global comparison intended or between-language comparison intended
  - Most eschew comparisons across languages anyway
- In spite of all this, EGRA and ASER finds lots and lots of zero skills
- Same problem as TIMSS, PISA even though much easier (because earlier grades) – but really not much to be done

Mendelovits, Munro-Smith, Murphy, and Waters (2014); Results for Development Institute (2015)
Conclusion on “bottom” measures

- “Rigorization” might not make “bottom” measures globally comparable in the sense of a single EGRA easily comparable across languages and skill levels, say
- Some rigorization necessary even for tracking progress within country
Last resort? Global framework or scale

- Many roads blocked.
- Single assessments not likely
- But maybe global scale, not single assessment:
  1. Establish a global framework of appropriate skills that correspond to proficiency levels
  2. Link via items and descriptors for the levels
  3. Accept that comparability not perfect
  4. But different methods of linking imply increasing comparability
  5. Validate empirically in a variety of countries, including the poorer ones, so as to make sure that bottom of the pyramid issues are covered.
  6. Set examples of benchmarks on the proficiency scale that countries could use, but without creating universal expectations.
**Luis Crouch**
Senior Economist
International Development Group
RTI International
+1 202 728 2058
lcrouch@rti.org

**Paper version:**
https://shared.rti.org/content/universal-assessment-and-bottom-pyramid