UNESCO Institute for Statistics

Education Data and Statistics Commission

Summary of key points around "GAML/EDSC criteria for use of an assessment to report on SDG 4.1.1"

GAML Meeting, 25-26 February 2025

Luis Crouch

Global Alliance to Monitor Learning

#25YearsOfDataInsights

Background

GAML Meeting December 2023 calls for clearer criteria on which assessments can report on 4.1.1.

- Autumn 2023 “demotion” of SDG 4.1.1a due to insufficient numbers
- Many in community claim that there are plenty of measurements
- UIS contends: “not sufficiently rigorous”
- UIS Calls for GAML meeting to discuss
- Sense of the meeting: need for clear criteria for what is acceptable, UIS please provide
- Draft document by March 2024 (note: special application to 4.1.1a)
- Many (hundreds) comments received
- TAGs in March and May 2024 to review, comment for further drafts
- Current version is Version 4 and incorporates the clarifications sought by commentators

Criterion 1: Alignment to the MPL (using 4.1.1a as case, others included) (“valid.”)

- Reading:
  - At least 20 “items” or score points from GPF
  - Must have at least 10 comprehension items; may have more
  - May have up to 10 items on “precursor” skills (accuracy, oral reading fluency, etc.); may have more
  - Only comp items will count towards % meeting the MPL but UIS will benchmark the “precursor” skills (called for by Montoya, after GAML meeting of December 2023)
- Mathematics
  - At least 20 “items” or score points from GPF
  - Of these, minimum of 10 in “numbers and operations”, incl. 3 out of the 4 subconstructs
  - Minimum of 10 “score points” in measurement, geometry, statistics and probability, and algebra
  - For now only numbers and operations proposed to be counted towards MPL, may change
- Note conceptual asymmetry reading vs mathematics: in maths, issue is not “precursors” but clarity on minimum acceptable scores

Criterion 2: Item Quality

- Judged appropriate by relevant experts for inclusion in the assessment
- Developed under advice from subject matter experts (SMEs)
- Discussed and vetted by local experts to ensure suitability for the local context
- SMEs responsible for items trained in item development principles and procedures
- Field tested on rep sample (note nuance on multi-country assessment)
- The scoring guides align with measurement intent
- Psychometric item analysis must be conducted on the field test data using at least CTT
- Item difficulty (e.g., item facility (CTT) or item location on the ability scale (IRT)) must be reviewed following the field trial and deemed appropriate, ideally have a diff> 0.20 and < 0.90
- Discrimination (for each item must be reviewed following the field trial

Criterion 3: Sampling appropriateness

- Target pop. relative to a, b, or c specified
- Exclusions shown, specified, justified
- Sampling approach documented (stratified, cluster, etc.)
- Sample size must allow for 95% CI of plus or minus 5 percentage points, show and defend calculations
- Sampling frame documented
- Sampling weights explained
- Response rate > 0.7, documented by stage
- Less than 15% schools substitution allowed

Criterion 4: Assessment administration

- Standardized manual must exist, suff clear for others to administer / replicate and get same validity and reliability results
- Process for selecting, training, qualifying, and replacing enumerators and supervisors detailed enough and robust
- Training protocols for the above explained and proof of application
- Explicit data Q&A plan including privacy protections, incl % of supervisor direct observation and/or re-visits
- Incident reporting procedure clear, decision rule for exclusion, exclusion procedure must not affect representativeness of the sample

Criterion 5: Reliability

- Must carry out and show item difficulty, discrimination and DIFF following live admin.
    - IRT preferred but may use CCT equivalent if plan to use IRT later
- Live assessment must have Cronbach alpha at least 0.8, for the relevant single-grade or singe-age group
- If items are oral or constructed response, then inter-rater reliability must be at least 0.8 kappa (or equiv.) in live application
- Items with weak reliability at live application can be removed but only with appropriate justification
- (Note: recommend full reliability analysis during field test as per above; if not perform, then run risk.)
- DIFF for gender and other SES to be analyzed and documented
- Where stop rule used, clarify whether missing for other reasons, missing because did not get to it, or incorrect: that is, items beyond those asked must be clearly interpretable

Criterion 6: Benchmark-based link to MPL

- Appropriate statistical link to MPL with clear benchmarks (what is considered “M” in an assessment to fit the M in MPL)
    - Suff items previously shown to link (common item)
    - Admin along another assessment (common person)
    - Other methods such as Benchmark being studied by UIS, with examples (see other presentation) to analytically link
- If common-item, must use IRT
- If common-person, may use CCT but prefer IRT
- May use “pairwise” comparison, extra details apply here as per the document

Criterion 7: Maintaining standards over time

- Items not in public domain may be re-used in future applications
- Items released to public domain may not be
- Process and metod in place, and documented, for ensuring equal difficulty over time, either common-person or common-item
- If common-item must document delta analysis and which are the common items
- If common-person, show concordance table with at least 95% CI

Related work and presentations (esp. on benchmarking, criterion 6)

- "Technical"
  - Analyze and create exemplars of items and their difficulties from "newer" assessments from reading science and maths-teaching science viewpoints (see presentations Linan-Thompson, Ketterlin-Geller)
  - Same from linguistic point of view (on reading) (see Nag presentation)
  - Using large scale data analysis from "newer" assessments (see Ferdous)
  - Test benchmarking method for linking directly to MPL (see Sandoval presentation)
  - Blueprints (see ACER presentation)
  - Note: esp if national assessment
- Institutional (see Cueto presentation)
  - "Vetting:" How are these criteria applied, what is the mechanisms? A lot of work required!
    - Esp. as national assessments come in
    - How do you submit? (See Colin Watson presentation)
  - "Virtual fund:" How do we ensure sufficient measurement in thus-far "orphan" countries

Gratitude to:

- All who commented on early drafts as part of GAML
- Colin Watson and colleagues at ACER for much of the heavy lifting and original sketch of list of criteria
- Abdullah Ferdous for collaboration in first drafts

unesco
Institute for Statistics

25 YEARS

Education
Data and
Statistics
Commission

THANK YOU

Learn more:

uis.unesco.org

databrowser.uis.unesco.org

@UNESCOstat

#25YearsOfDataInsights

Global
Alliance
to Monitor
Learning