AIR®
Advancing Evidence.
Improving Lives.

Quality Measures of Individually Administered Assessments for SDG 4.1.1a Reporting

Dr. Abdullah Ferdous
Principal Researcher/Psychometrician
American Institutes for Research, International Development Division
aferdous@air.org

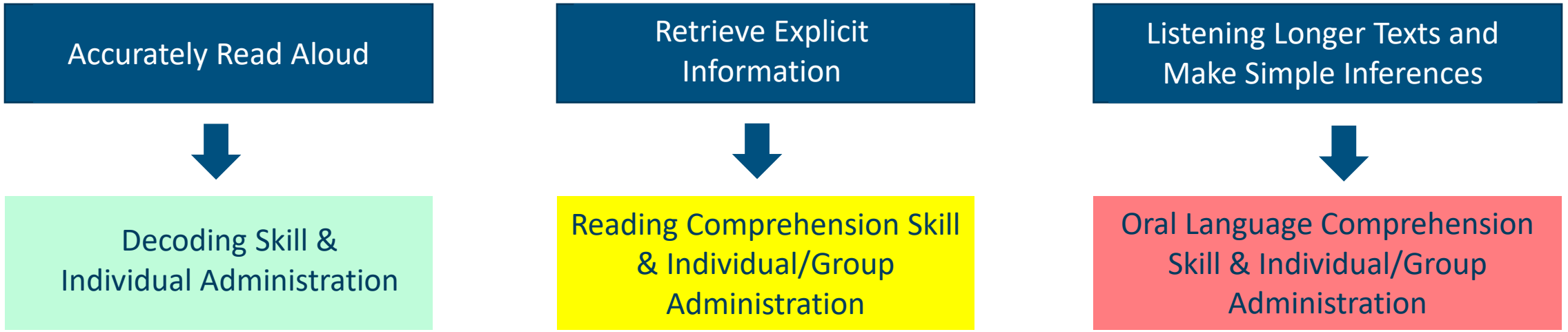
Global Alliance to Monitor Learning (GAML), Paris | December 2023

Background

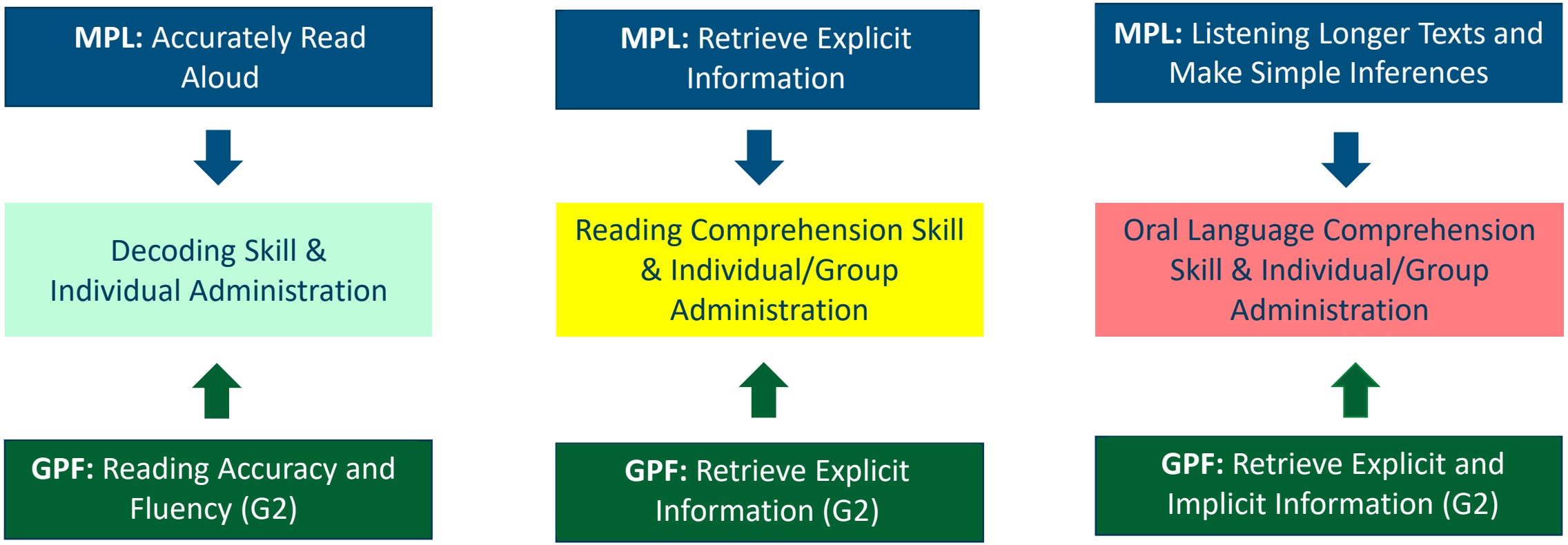
- **SDG 4.1.1:** Proportion of children and young people: (a) in **grades 2/3**, (b) at the end of primary, and (c) at the end of lower secondary achieving **at least** a minimum proficiency level in (i) reading and (ii) mathematics, by sex.
- The data for this indicator are expected to be:
  - **Comparable** across countries that have linked their assessments to a common scale.
  - **Aggregating** assessment results across countries to report.
  - **Tracking** assessment results over time to monitor progress.

Background

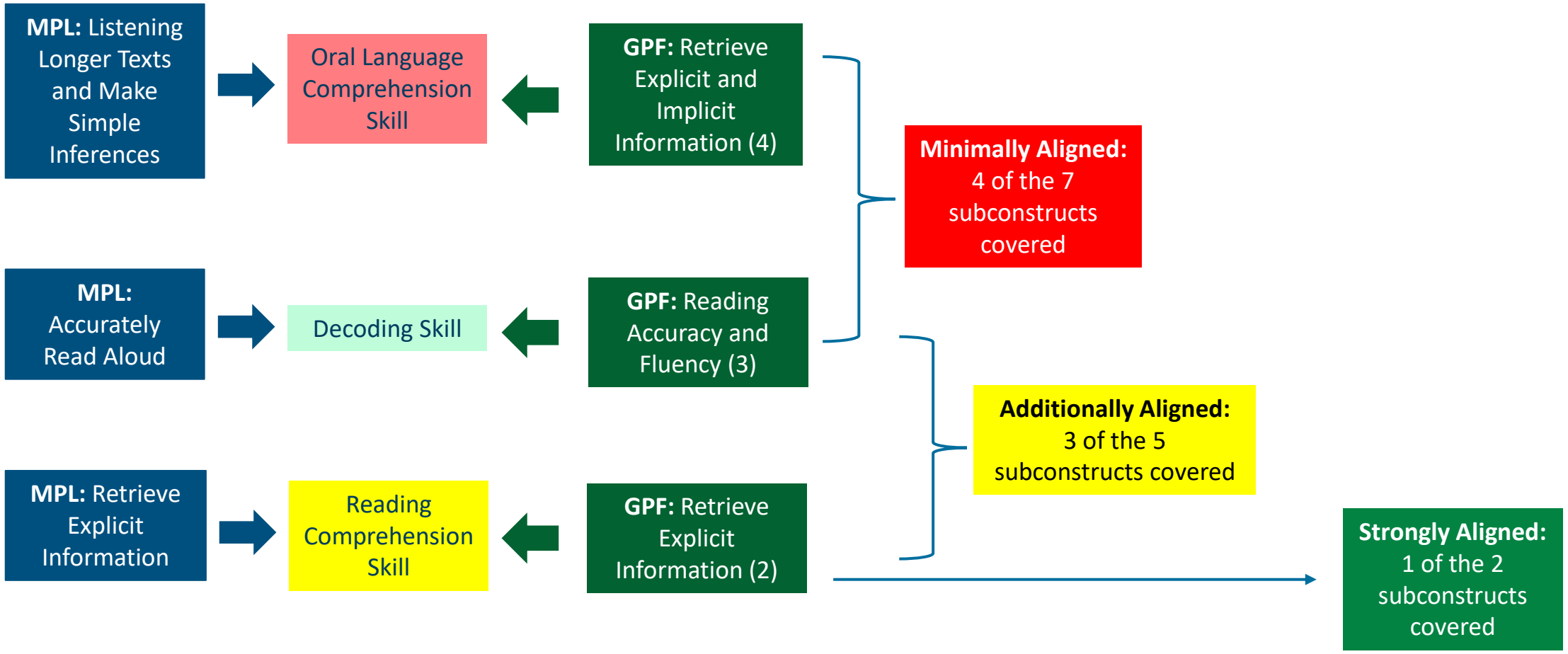
Minimum Proficiency Level (MPL): Students accurately read aloud and understand written words from familiar contexts. They retrieve explicit information from very short texts. When listening to slightly longer texts, they make simple inferences.



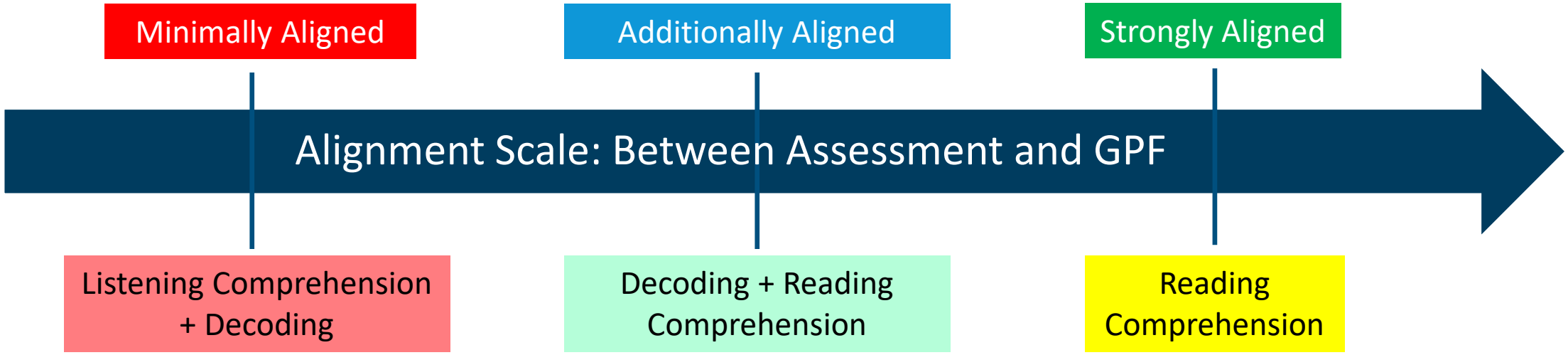
Acceptable Alignment between Assessments and MPL/GPF



Acceptable Alignment between Assessments and MPL/GPF



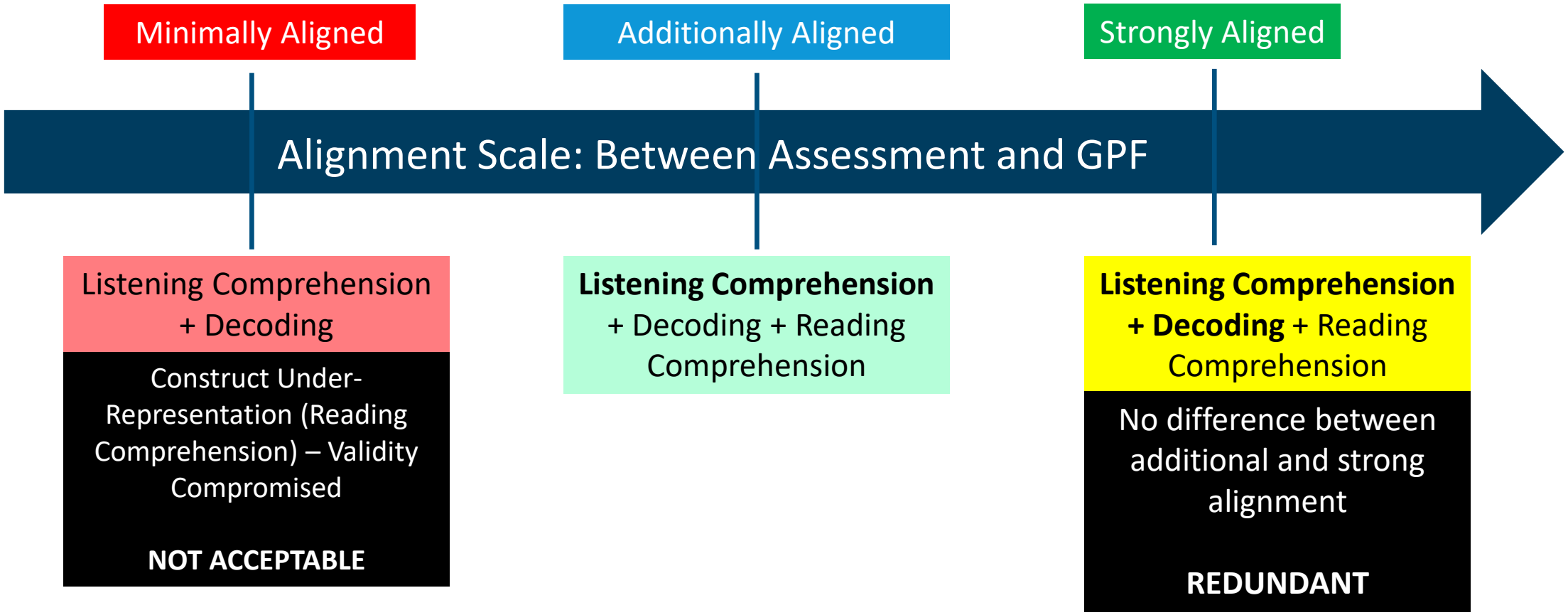
Acceptable Alignment between Assessments and MPL/GPF



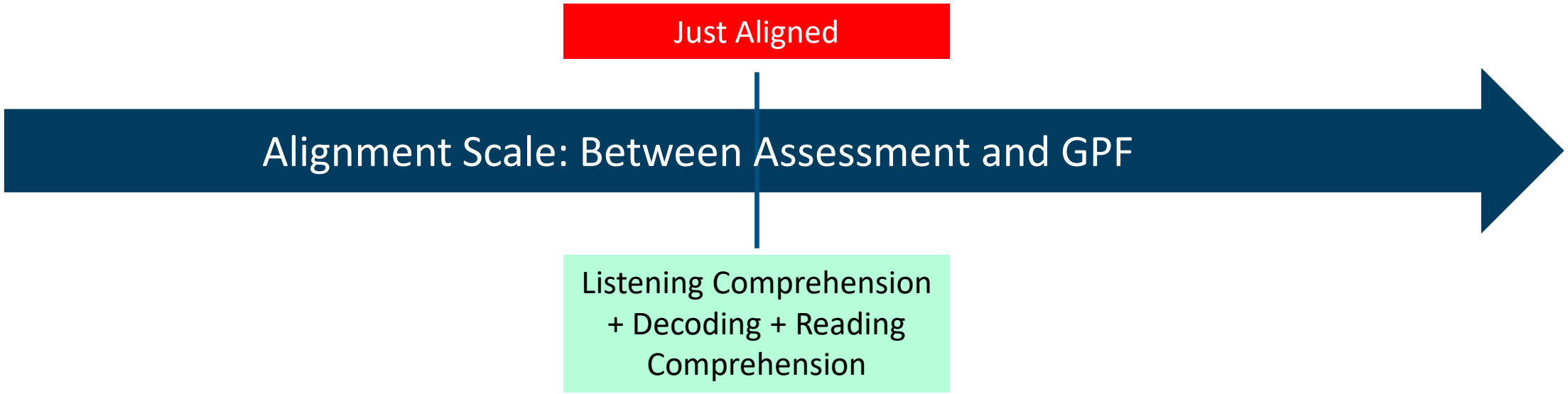
Characteristics of an Ordinal Scale

1. Monotonically increasing
2. Meaning attached to each scale point
3. The description of each point is additive
4. Distance between two points is not necessarily equal

Acceptable Alignment between Assessments and MPL/GPF



Acceptable Alignment between Assessments and MPL/GPF



Recommendation: Instead of three levels on an ordinal alignment scale, a single level could be proposed, termed “Just Aligned," provided the assessment covers the listening comprehension, decoding (reading accuracy), and reading comprehension domains, along with at least one half of their respective subconstructs within each domain.

Questions

- **Criterion 1:** Are assessments sufficiently aligned with the MPL/GPF?
- **Criterion 2:** Have the items of these assessments been reviewed qualitatively and quantitatively to determine their suitability for inclusion?
- **Criterion 3:** Is the sample of learners that took the assessment representative of the population against which the results will be reported?
- **Criterion 4:** Is there evidence that the assessment was administered in a standardized way?
- **Criterion 5:** Are the outcomes of the assessment sufficiently reliable?

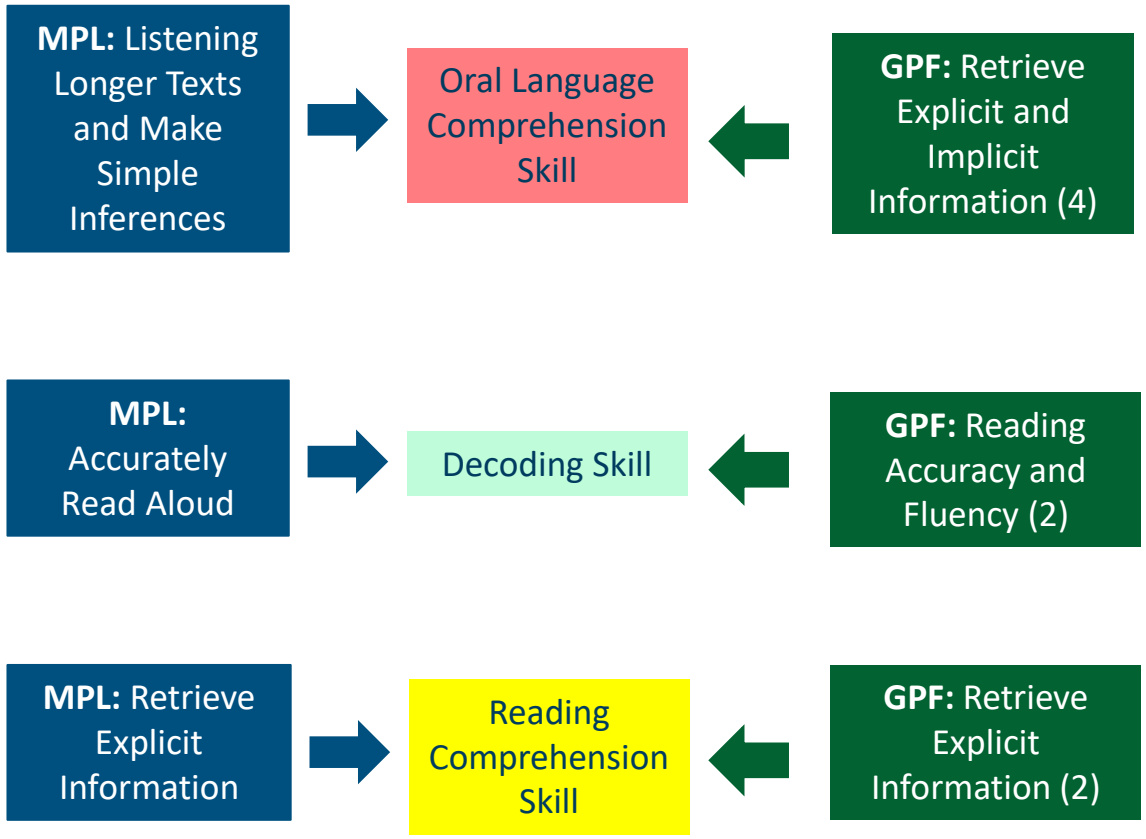
Questions

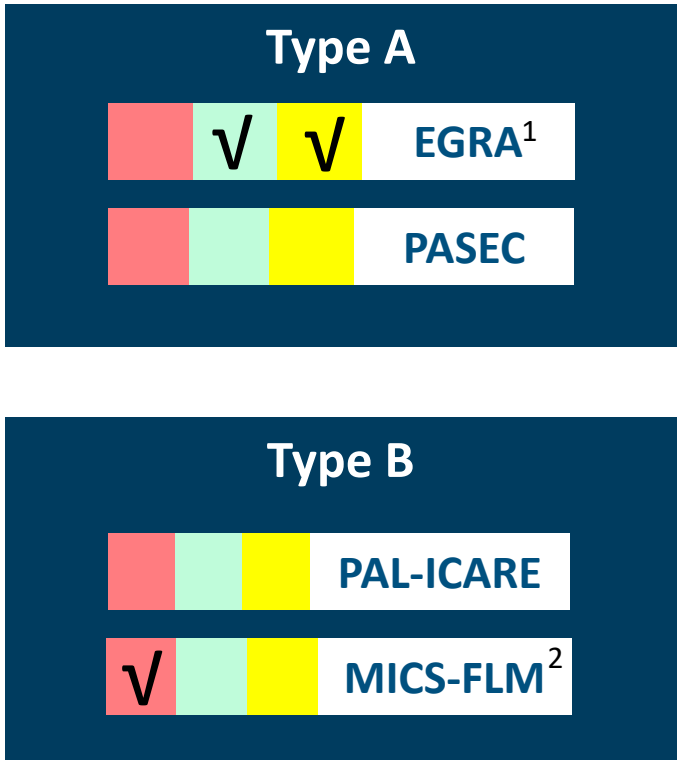
Criteria	Domains/Subdomains to Cover the MPLa	Type A		Type B	
		EGRA	PASEC	PAL-ICARE	MICS-FLM
Criterion 1	Listening Comprehension (LC)				
	Reading Comprehension (RC)				
	Oral Reading Accuracy (Decoding)				
Criterion 2	Qualitative and Quantitative Review of the Items				
Criterion 3	National Representative Sample				
Criterion 4	Standardized Test Administration				
Criterion 5	Assessment Reliability				
	Benchmark Method				

Materials Reviewed

- Assessment frameworks
- Test development and adaptation guidelines
- Psychometric item analysis reports
- Sampling reports
- Assessor and supervisors training guidelines
- Supervisor monitoring guidelines
- Policy linking toolkit
- Global proficiency framework (GPF)

Criterion 1: Are the assessments sufficiently aligned with the MPL/GPF?





1. Provided it uses the core subtasks for reporting, 2. latest version of the MICS-FLM

Criterion 1: Are the assessments sufficiently aligned with the MPL/GPF?

- Each of these assessments (EGRA[1], PASEC, PAL-ICARE, and MICS-FLM[2]) exhibits the essential alignment with MPL/GPF.
- The interpretations of test scores for these assessments are both consistent and robust, serving the intended purpose of facilitating cross-country comparisons.
- Countries implementing these assessments must produce an **assessment specification** or **test blueprint** document to provide evidence of this criterion.
- They should prepare assessment specification document that describes the definitions of domains measured and interpretations for intended uses (Standard 4.1). It should also define the content of the test, the length, and the item formats (Standard 4.2).

1. Provided it uses the core subtasks for reporting, 2. latest version of the MICS-FLM

Criterion 1: Are the assessments sufficiently aligned with the MPL/GPF?

Criteria	Domains/Subdomains to Cover the MPLa	Type A		Type B	
		EGRA	PASEC	PAL-ICARE	MICS-FLM
Criterion 1	Listening Comprehension (LC)	✓	✓	✓	✓
	Reading Comprehension (RC)	✓	✓	✓	✓
	Oral Reading Accuracy (Decoding)	✓	✓	✓	✓
Criterion 2	Qualitative and Quantitative Review of the Items				
Criterion 3	National Representative Sample				
Criterion 4	Standardized Test Administration				
Criterion 5	Assessment Reliability				
	Benchmark Method				

Criterion 2: Have items been reviewed to determine suitability for inclusion?

- These assessments (EGRA, PASEC, PAL-ICARE, and MICS-FLM) implements international best practices in test development.
  - Items are constructed based on item specifications and guidelines.
  - Subject matter experts (SMEs) participate in item writing and adaptation.
  - Item face validity is checked (Qualitative Reviews).
  - Items are field tested to examine psychometric properties (Quantitative Reviews) using classical test theory (CTT) and/or Item Response Theory (IRT).
  - Countries implementing these assessments should document the test development or adaptation process in detail (Standards 4.7 - 4.10) to meet this criterion.

Criterion 2: Have items been reviewed to determine suitability for inclusion?

Criteria	Domains/Subdomains to Cover the MPLa	Type A		Type B	
		EGRA	PASEC	PAL-ICARE	MICS-FLM
Criterion 1	Listening Comprehension (LC)	✓	✓	✓	✓
	Reading Comprehension (RC)	✓	✓	✓	✓
	Oral Reading Accuracy (Decoding)	✓	✓	✓	✓
Criterion 2	Qualitative and Quantitative Review of the Items	✓	✓	✓	✓
Criterion 3	National Representative Sample				
Criterion 4	Standardized Test Administration				
Criterion 5	Assessment Reliability				
	Benchmark Method				

Criterion 3: Does the sample of learners represent the population?

- The policy-linking toolkits makes recommendations on nationally representative samples, with two specific considerations for individually administered tests.
  - Instruments are administered in multiple languages
    - An example of the Foundational Learning Study in India
  - Comparing household-surveys to school-based surveys.
    - Report statistical power of the studies separately for children attending formal schooling and those not attending
- Future efforts could explore the feasibility of expanding the surveyed population in non-nationally representative studies.

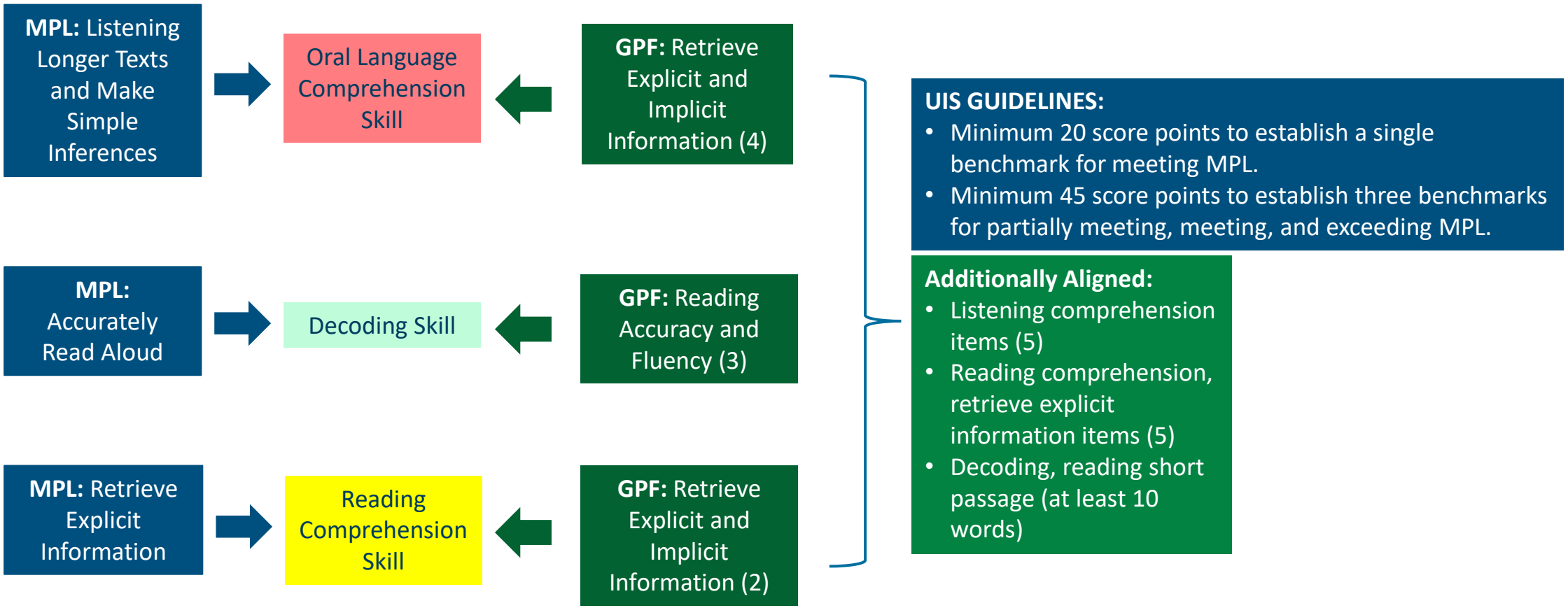
Criterion 4: Was the assessment administered in a standardized way?

- The **EGRA, PASEC, PAL-ICARE,** and **MICS-FLM,** toolkit or manual offers comprehensive guidance, covering all aspects from assessor recruitment to data upload onto servers.
- Assessment booklets include precise instructions for assessors to follow as they administer the tests.
- A training program is conducted throughout the country, including field practice to ensure that surveyors have a complete and consistent understanding of the test administration procedure.
- Supervisors are trained to monitor surveyors' pre-survey performance evaluation during training, during survey monitoring, and post-survey recheck based on monitoring and recheck guidelines.
- Separate manuals are created for assessors and supervisors to cater for their specific roles and responsibilities.
- Countries implementing these assessments must produce a standardized test administration manual that adheres to Standards 6.1, 6.3, 6.5, and 6.7.

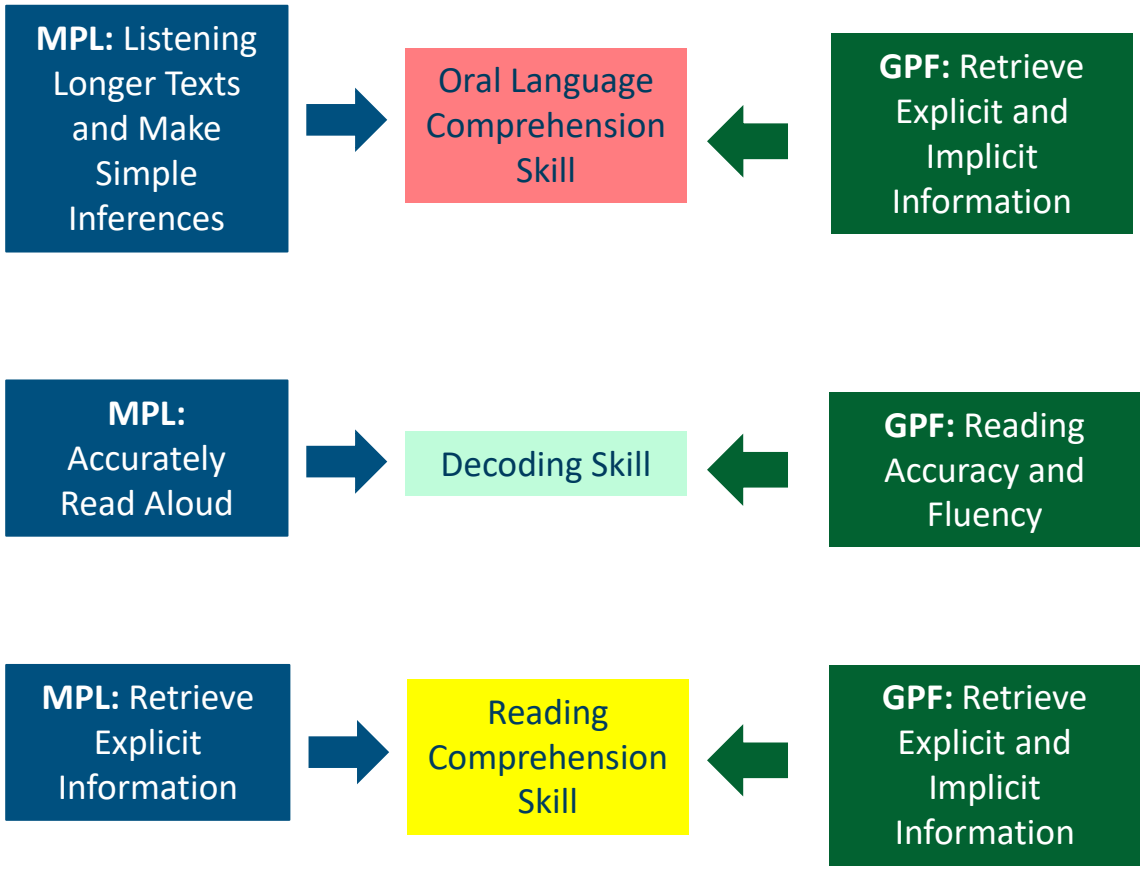
Criterion 4: Was the assessment administered in a standardized way?

Criteria	Domains/Subdomains to Cover the MPLa	Type A		Type B	
		EGRA	PASEC	PAL-ICARE	MICS-FLM
Criterion 1	Listening Comprehension (LC)	✓	✓	✓	✓
	Reading Comprehension (RC)	✓	✓	✓	✓
	Oral Reading Accuracy (Decoding)	✓	✓	✓	✓
Criterion 2	Qualitative and Quantitative Review of the Items	✓	✓	✓	✓
Criterion 3	National Representative Sample				
Criterion 4	Standardized Test Administration	✓	✓	✓	✓
Criterion 5	Assessment Reliability				
	Benchmark Method				

Criterion 5: Are the outcomes of the assessment sufficiently reliable?



Criterion 5: Are the outcomes of the assessment sufficiently reliable?



Type A

12 items	50 words (single passage)	15 items	EGRA
16 items	At least 12 items	18 items	PASEC

Type B

6 items	At least 45 words (two passages)	5 items per passage	PAL-ICARE
5 items	30 words (single passage)	10 items	MICS-FLM

Criterion 5: Are the outcomes of the assessment sufficiently reliable?

- Three distinct domains: listening comprehension, decoding, and reading comprehension – measurement of each domain should demonstrate adequate reliability.
- A minimum reliability coefficient of internal consistency (i.e., Cronbach alpha of at least 0.70) should be attained for each domain/entire assessment.
- Each domain in these assessments (EGRA, PASEC, PAL-ICARE, and MICS-FLM) includes the requisite number of items for the reliable measurement of the knowledge and skills outlined in the MPL.
- Countries implementing these assessments usually report the reliability of internal consistency (Cronbach alpha) for each subtask separately.

Criterion 5: Are the outcomes of the assessment sufficiently reliable?

- If a test is used for a range of ages (e.g., MICS-FLM and PAL-ICARE), reliability coefficients should be reported for each domain and for each age range, not just all ages combined (e.g., 7-9 and 10-14 years for FLM-MICS) (Standard 2.12)

Criterion 5: Are the outcomes of the assessment sufficiently reliable?

Criteria	Domains/Subdomains to Cover the MPLa	Type A		Type B	
		EGRA	PASEC	PAL-ICARE	FLM-MICS
Criterion 1	Listening Comprehension (LC)	✓	✓	✓	✓
	Reading Comprehension (RC)	✓	✓	✓	✓
	Oral Reading Accuracy (Decoding)	✓	✓	✓	✓
Criterion 2	Qualitative and Quantitative Review of the Items	✓	✓	✓	✓
Criterion 3	National Representative Sample				
Criterion 4	Standardized Test Administration	✓	✓	✓	✓
Criterion 5	Assessment Reliability	✓	✓	✓	✓
	Benchmark Method				

Setting benchmarks

- Students meeting global minimum proficiency level should be determined either using a compensatory or a conjunctive scoring method.
  - In a compensatory model, it is assumed that strong performance in one skill can make up for weak performance in another skill.
  - In a conjunctive model, students must achieve a specified level of performance on each skill to be classified at that proficiency level.

Setting benchmarks

Skill	No. of Items	Benchmark	Student 1	Student 2
Listening Comprehension	5	3	4	3
Decoding	30	18	24	22
Reading Comprehension	5	2	0	2
Total Score	40	23	28	27
Meeting MPL: Compensatory			**Yes, without demonstrating reading comprehension skill**	**Yes**
Meeting MPL: Conjunctive			**No**	**Yes**

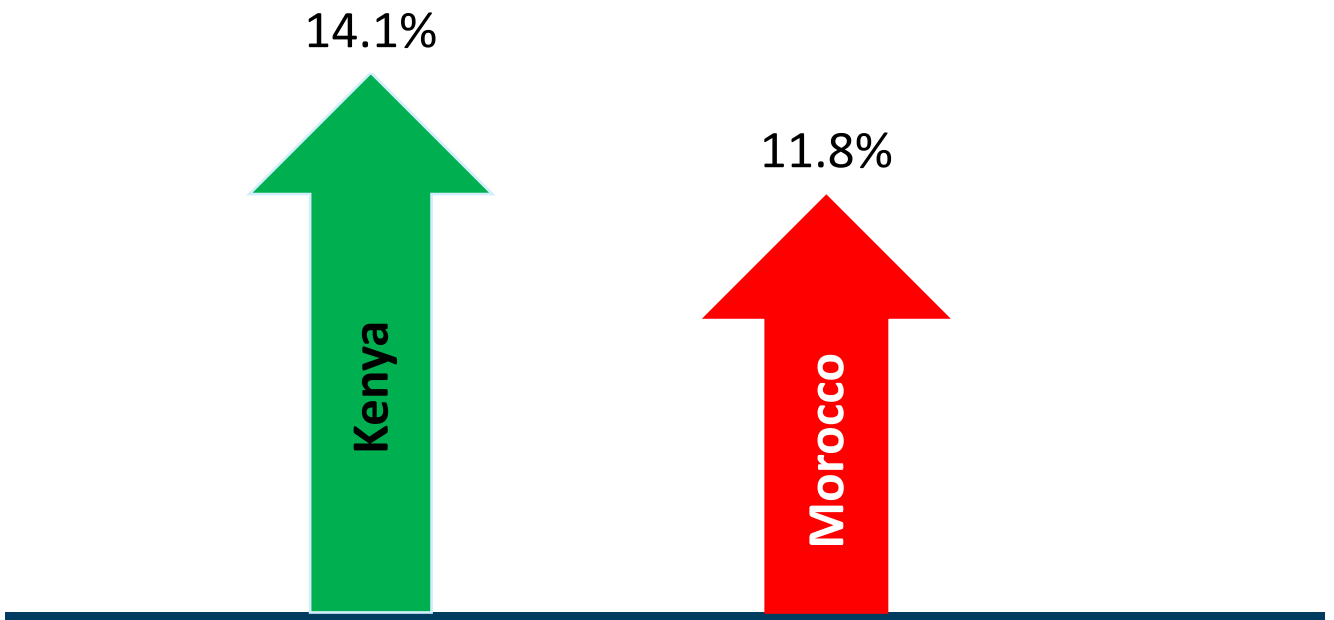
Setting benchmarks

Kenya: Grade 2	Benchmark	% of Students	Compensatory	Conjunctive
Listening Comprehension (5)	4	27.6%	35.3% students obtained at least 31 out of 78 marks	14.1% Students obtained at least 4 in LC, 24 in OR, and 3 in RC
Oral Reading (68)	24	60.5%		
Reading Comprehension (5)	3	28.7%		
TOTAL (78)	31			

Morocco: Grade 2	Benchmark	% of Students	Compensatory	Conjunctive
Listening Comprehension (6)	3	29.2%	32.2% students obtained at least 34 out of 67 marks	11.8% Students obtained at least 3 in LC, 28 in OR, and 3 in RC
Oral Reading (55)	28	36.7%		
Reading Comprehension (6)	3	28.3%		
TOTAL (67)	34			

Setting benchmarks

- A conjunctive scoring model is the most ideal scoring approach to ensure that results from diverse assessments aligned with MPL/GPF can be compared effectively.
- It will classify students with the same minimum knowledge and skills across countries into the meeting global MPL, irrespective of the differences in their learning assessments and conditions of their learning.



Proportion of children in **grade 2** achieving **at least** a minimum proficiency level in reading (accurately read aloud and understand written words from familiar contexts. They retrieve explicit information from very short texts. When listening to slightly longer texts, they make simple inferences).

Criterion 5: Are the outcomes of the assessment sufficiently reliable?

Criteria	Domains/Subdomains to Cover the MPLa	Type A		Type B	
		EGRA	PASEC	PAL-ICARE	FLM-MICS
Criterion 1	Listening Comprehension (LC)	✓	✓	✓	✓
	Reading Comprehension (RC)	✓	✓	✓	✓
	Oral Reading Accuracy (Decoding)	✓	✓	✓	✓
Criterion 2	Qualitative and Quantitative Review of the Items	✓	✓	✓	✓
Criterion 3	National Representative Sample				
Criterion 4	Standardized Test Administration	✓	✓	✓	✓
Criterion 5	Assessment Reliability	✓	✓	✓	✓
	Benchmark Method	✓	✓	✓	✓

Caveat : Other pre-requisites for all tools reporting

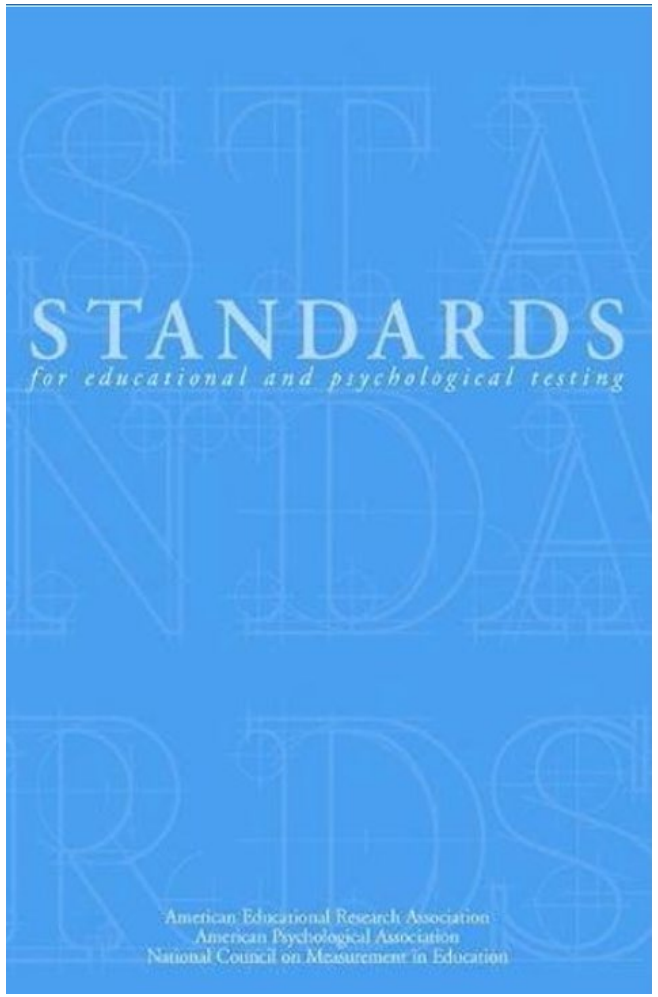
- No stop-rules
- Government ownership
- Instrument should not be public for school-based assessment
  - A possible solution of pre-calibrated instruments
- If these criteria are not met by an implementer of an assessment (e.g., insufficient documentation or non-nationally representative sample) then that administration cannot be used for reporting.
  - Detailed guidelines based on Standards should be prepared.

Conclusions

- Assessments for SDG 4.1.1a reporting must measure all three domains – listening comprehension, decoding, and reading comprehension.
- Given that cross-country comparisons are integral part to SDG 4.1.1a, assessments' validity and reliability, along with consistent score interpretations, become crucial.
- Once benchmarks are set on assessments using policy linking or other standard setting method, a classical- and item-response-theory-based approached can be employed to track and report student progress over time. It is important to note that benchmarks are set only once during the life of the assessment.
- To enhance the security of test instruments, PAL and MICS may consider developing multiple pre-calibrated forms. These forms could be utilized at different administration times.

Conclusions

- The development and implementation of EGRA, FLM-MICS, PAL-ICARE, and PASEC are poised to meet all the requirements outlined in the policy linking toolkit for SDG 4.1.1a.
- A users' guide to the conjunctive scoring model should be prepared and disseminated through different venues.
- Countries undertaking these assessments are required to thoroughly document the processes and procedures involved in test development or adaptation, as well as sampling and the standardized test administration procedures.
- The *Standards* book offers comprehensive guidelines, specifying the detailed requirements for assessments to ensure their reliability, validity, and fairness.
- The combined use of these assessments holds the potential for substantial data coverage in support of SDG 4.1.1a.



ABDULLAH FERDOUS, PH.D.

Principal Researcher/Psychometrician

WhatsApp: +1 508 969 7033

Email: aferdous@air.org

LinkedIn: LinkedIn.com/in/abdullah-ferdous-9719a124