

EGRA and the SDG 4.1.1 Criteria - a crosswalk



This document summarizes the technical requirements in the update EGRA guidance in an easy to reference format. This document is not intended to be used alone; rather, it is a reference guide to those carrying out an EGRA to check its alignment to key parameters of quality.

Summary of Technical Requirements in the EGRA Toolkit 3.0

Parameter	steps required	reporting required	COMMENTS
-----------	----------------	--------------------	----------

**Alignment to the
GPF and construct
validity**

During the design and adaptation phases, conduct an alignment exercise.

The alignment exercise, which is based on expert judgement, demonstrates the relationship of each item and subtask to the GPF constructs and sub-constructs, for each grade level assessed.

Results from the alignment exercise are documented in the final report including purpose of the assessment, definition of the domains measured, constructs and subconstructs and learning outcomes measured, judged level of alignment by an expert for each item and GPF construct and subconstructs, item formats.

Alignment report summarizes the relationship between the assessment items and the proficiency levels the assessment seeks to measure. When relevant, the alignment report demonstrates appropriate use of benchmarks (how many benchmarks are feasible, for whom and for what purpose).

When developing the EGRA instrument, reading and curriculum experts can conduct an alignment exercise that examines each item and its adherence to measuring the constructs of interest. During adaptation workshops, the constructs and sub-constructs, in addition to curriculum standards and teaching and learning materials, should be used to guide item development. During piloting, ensure that there are sufficient high-quality items per sub-construct.

Ensure sufficient comprehension items and leveled passages are developed and documented.

Scoring guide with a minimum of 20 total score points aligned to the GPF in reading for the grade assessed. Of these 20 points, a minimum of 10 points assess reading comprehension.

When assessing Grade 2, the assessment covers both reading comprehension subconstructs.

When assessing Grade 5, the assessment covers 4 of the 8 reading comprehension subconstructs.

When determining the purpose of the EGRA, if there is a high value placed on comprehension, ensure there are sufficient items to inform decisions while maintaining a reasonable length of the test. This may mean fewer precursor skills assessed if more information on comprehension is required.

During pilot, gather data on sufficient items and reading passages to support selection of items and proposed scoring.

During pilot and operational data analysis, follow the scoring guide and document where scores are imputed; disaggregate by domain, construct and subconstruct according to alignment exercise.

Ensure reading comprehension is assessed using sentence verification, reading passage, or both.

Both fluency and accuracy scores may be reported. Accuracy scores should be based on an assessment where learners have sufficient time to complete the passage (typically 2-3 minutes total).

Reading comprehension scores should be based on a child being able to attempt all items. To reduce testing burden on lower performers, pilot results support the imputing of zero scores for any items.

Multiple reading passages may be combined with sentence verification or other well-piloted comprehension tasks. Ensure both reading comprehension subtasks are covered. Users may need to consider a word comprehension subtask. During piloting, the quality of the tasks to measure comprehension among the population of interest – without too many ceiling or floor effects – must be documented and reported. Any skip patterns should also be noted and accounted for in the analysis and ensure all children attempt a sufficient number of items.

**item content
and quality**

Ensure each item has been reviewed and judged appropriate by reading and curriculum experts to be included in the assessment.

Report the design and adaptation processes, who is involved and what their roles are. Include adaptation workshop participants list and summary of workshop. Training should also include details on what experts are looking for when developing and reviewing items to ensure their quality and lack of bias.

Workshop summary report should include: participants list, workshop process for training participants on item development principles and procedures, qualifications of the subject matter experts (SMEs), summary of discussions (including concerns and resolutions).

Plan for field testing items during workshop and results of the field test.

Draft scoring guide, to be revised and finalized following pilot.

The alignment exercise (described above) can be combined with this step. Typically, this qualitative review by experts is part of the adaptation workshop. If using an existing tool, this review by experts – for the purpose and population currently being assessed – is critical and should not be skipped.

Note that the field test is not the same as the pilot. The field test is part of the development process of the tool and provides information as item developers are working to create appropriately leveled items. The pilot is an exercise of testing how well the developed items do, as a complete instrument. Typically, the pilot should include at least 2-3 versions of the complete instrument to allow selection of the best performing instrument for the operational data collection.

	Pilot study that includes sufficient assessment items to select from	Pilot report with the following sections, at a minimum: study plan, assessment items that were tested, who was tested, how the tested population is representative of the population to whom the test will be administered, who conducted testing and their training, who they were tested with, results. Pilot results including the following for each item: item difficulty, item discrimination, item-total correlation, and any qualitative observations of item performance during pilot	During design and adaptation, the psychometrician and/or data analyst will work with the reading experts and other SMEs to design the pilot and conduct relevant analysis. Results of the pilot study will indicate which items are appropriate to use in operational data collection because they will meet the required thresholds. Inclusion of poor-performing items (such as excessive floor and ceiling effects) should be carefully considered. Additional testing for revised items should be conducted and documented.
	Final assessment tool based on pilot results	Include a table in the pilot report that shows that all items meet the following requirements: • difficulty level between 0.20 and 0.90, with exceptions documented and justified. • Item discrimination values judged appropriate for the population and purpose • Item-total correlations of at least 0.20	The data analyst and psychometrician will advise on which items are appropriate to retain and use in the operational data collection and which items may need to be dropped and/or revised. The final pilot report should include a summary table that is easily accessed and provides the necessary item quality results. Item quality is again reviewed during final data analysis to confirm that items behaved as expected with the full population. Any discrepancies are addressed and discussed in the analysis section of the report.

population coverage and sampling	Gather necessary data on the population to which the results will be generalized to	Document the population frame including data on all variables for disaggregation and stratification. Describe the desired target population and disaggregations of interest. Documentation on subgroups (or characteristics, such as distance) excluded and justification for the exclusions.	School-based assessments typically use schools as the unit of analysis for the population frame and for sampling. Household- or community-based assessments (i.e, outside the school system) should demonstrate that the population frame is complete and adequately captures information on the population of interest.
	Select a sample (if not conducting a census assessment)	Document the sampling approach and design including stratification, clustering and descriptions of how representation by gender, language, geography, socio-economic status and other variables of interest, will be represented fairly in the sample. Document likely design effect. All sample size calculations are documented and verified to be appropriate for the given purpose of the assessment. Depending on purpose, the sample size must be appropriately selected. For indicator reporting such as SDG 4.1.1, the sample should achieve a 95% CI of at most 5 percentage points.	The documentation of the sampling approach also provides the necessary documentation for validity of interpretation of the results. The results can only be generalized to the population that is represented adequately by the sample. The documentation should note whether learners from specific sub-groups are included or not: learners with disabilities, IDPs, learners who don't speak the language of the assessment, out-of-school children, learners from specific geographical locations or ethnic minorities. Reasons for exclusion should also be documented. Be sure to engage with a statistician to ensure the adequacy of the sample size and stratification (or layers) used to select the sample. Oftentimes, unnecessarily large samples are used because the appropriate statistics have not been computed, resulting in unnecessary costs and time.

Compute sampling weights
(if appropriate)

Calculate sampling weights and document calculations for the overall sample and sub-groups.

Determine the strategy for non-response or substitution. Document how substitution will be done and training to be provided to conduct substitution appropriately during operational data collection.

Statisticians can typically provide this information at the time of sampling. The weights will be applied at the time of analysis, and may be after statistician engagement has ended. Be sure to gather detailed documentation from the statistician on sampling weights.

assessment administration and data custodianship	Develop and operationalize a standardized test administration manual	Test administration manual with sufficient clarity that others can reproduce the administration conditions.	
	Recruit and select enumerators and supervisors	Recruitment advertisement for enumerators and supervisors, selection process, and structure of teams to be deployed.	During recruitment, the number of teams should be based on the budget, number of days expected in field, and the length of the assessment for each child/school/day. The whole EGRA team will be involved in this process.
	Training enumerators and supervisors/Quality Control Officers with a focus on quality during field test and pilot	Training manual with sufficient clarity to recreate the training process. Manual must include at least: training agenda, training materials, training videos or supplementary resources used, process for identifying and addressing training gaps, process for selecting final enumerators based on a training completion assessment, safeguarding protocols, consent, and standardized administration and recording practices for each item and each task. Training manual for supervisors and quality control officers and post-training qualifying conditions required. Plan for number of IRR tests to be conducted per team to demonstrate sufficient monitoring of quality.	Training typically is conducted by a training expert over a 3-5 day training workshop. The workshop should include extra enumerators in the beginning, and as the workshop progresses, quality control protocols should allow selection of the final enumerators that meet specifications. Depending on the enumerators engaged during field test, the training required for pilot may be shortened. Workshops can also be used to assess which enumerators will be able to serve as supervisors/Quality Control Officers in the field. Dry run practices for data collection should be planned between the conclusion of training and the beginning of operational data collection.

Monitor quality during operational data collection.

Develop and train all enumerators and supervisors/QCOs on the data quality assurance plan.

Document response rates by sample stage and substitutions, if any.

Document discrepancies between enumerators, between sampling guidance and on-the-ground realities, and resolutions made. Document concerns or analytical steps to be taken during analysis to address discrepancies. All incidents of inappropriate, incomplete, or stopped test administration should be documented and immediately reported to allow for additional checks and/or substitution of the enumerator or the sampled unit.

The plan for IRR during operational data collection should demonstrate sufficient monitoring of quality and reasonable cost. Conducting IRR for all teams each day represents a significant number of supervisors and can be cost prohibitive. On the other hand, conducting IRR only during training means there is no monitoring of the live data collection.

A good plan for monitoring should use templates, regular feedback loops with enumerators and supervisor/QCOs, and easy-to-access documentation. For example, paper-based documentation may require additional steps to aggregate the information into a spreadsheet, which can then be acted on by analysts during data analysis. Templates and regular feedback discussions with QCOs can reduce the burden of documentation and increase timely corrections in errors. Templates can be shared electronically and regularly to ease aggregation and increase awareness of issues as they arise.

Data custodianship and data flow plans and systems are in place

Data collection software and hardware specifications are documented, including procedures for programming.

Data storage, access and security protocols are documented and access is appropriately given (i.e., only those covered under the consent statements have access).

Data flow from the field to centralized servers is documented and security protocols are in place; any security or data breaches are recorded and resolutions are documented.

Data custodianship involves a clear plan, from the beginning, regarding data security and storage. Whether using paper or tablets to administer tests, or both, a data custodianship plan will articulate who access which data and for what purpose, what data are saved in the software and ensure adherence to consent statements. For example, if a child declines to participate, the software should not retain any information about them.

Data storage and access plans also should be developed from the beginning. These plans can help clarify how data are to be stored and shared (how often should enumerators upload from their tablets), how data are stored when aggregated and who has access and for how long. These plans also rely on well written consent statements.

reliability	Conduct Item analysis for live data collection	Table of all results from item analysis including but not limited to: item difficulty, discrimination and differential item function (if appropriate)	Item analysis is conducted during field test, pilot and on the operational data. This ensures that items behaved they way they were expected to even during the assessment. In the event items did not function as expected, additional analyses can be conducted – were the issues with specific enumerators, types of learners, or otherwise. Item analysis can help ensure the summary results are reliable measures of learning among the population of interest.
	Cronbach's alpha	Cronbach's alpha value of at least 0.80 for a reading comprehension total score.	A score that meets this threshold of 0.80 indicates that it represents children's ability levels on a single underlying construct. When field testing and piloting, the alpha values should help determine when the instrument is sufficiently measuring the skill of interest. Note that the EGRA is intended to be used to understand performance at the skill level and therefore is not suitable for a total score across all subtasks. Typically, combining like items – such as 15 reading comprehension items – is recommended. When out-of-school students or students of varying ages are included, the Cronbach's alpha may need to be computed for both the full sample and for those who are within the age/population ranges of interest.

	Conduct live data collection Inter-Rater Reliability	Conduct, compute and report IRR values for operational data collection based on the planned frequency of IRR.	As discussed above, this relies on a strong and reasonable data monitoring plan.
	Ensure imputation of zero scores is appropriate and as planned	Based on the assessment plan and programming of items, and where a stop rule might be applied, determine scoring rules for items not administered, response missing and/or response provided but incorrect. The decisions for scoring are included in the scoring plan.	The scoring plan ensures that children are attempting the number of items required. It will also outline how skip patterns are to be handled. Skip patterns are carefully developed during field testing and piloting to ensure that, if a item or subtask are programmed to be skipped in the test, there is sufficient confidence (based on the students responses thus far on the test) that they are likely to get a zero on the skipped items. As such, a skipped item would be scored as zero during analysis and a child's score is computed out of the total number of items (and not just those that they attempted). This facilitates reducing burden on a struggling reader while also ensuring scores are comparable across all learners in the dataset.
benchmarks	Determine benchmarking procedures to be used	Document process for consideration and selection of statistical linking and or non-statistical linking methods.	Forthcoming guidance
	Conduct benchmarking exercise		Forthcoming guidance
	Report results against established benchmarks		Forthcoming guidance

Maintaining standards and comparability over time	Plan for maintaining benchmark standards to be relevant and appropriate over time	Ensure test security and procedures for updating tests for each administration are in place.	Forthcoming guidance
	Plan for test equating and/or item bank to support future, comparable, test versions	Document process for test equating and item bank creation, security and selection of items for future, comparable assessments	Forthcoming guidance