AHEAD

African Harmonized Early-Grade Assessments Dataset

Reference Guide

Most recent release date: **02 July 2026**

Primary investigators

Cally Ardington, Chifundo Kanjala, Siphiwe Bogatsu, Mikaela Daries

CONTENTS

1. Introduction

This reference guide describes the harmonization framework developed for the African Harmonized Early-Grade Assessments Dataset (AHEAD), which combines datasets from United States Agency for International Development (USAID)-funded Early Grade Reading Assessment (EGRA) and Early Grade Mathematics Assessment (EGMA) projects conducted across Africa. The data harmonization undertaken in this project is retrospective. Retrospective harmonization (also referred to as ex-post or output harmonization) involves harmonizing datasets after they have been collected. The goal is to develop a flexible, scalable, and analytically robust database that supports longitudinal, cross-sectional, and comparative analyses across diverse educational contexts.

Recognising that EGRA and EGMA assessments have been implemented over many years in different countries and programmes using different assessment instruments, languages, sampling designs, and variable naming conventions, this harmonization framework provides a systematic approach to identifying and standardising comparable measures while preserving the integrity and context of the original datasets.

This reference guide introduces the harmonized dataset and the long-term harmonization programme. It describes the structure and content of the pooled dataset; and provides guidance on variable naming conventions, assessment variables, identifiers, survey weights, missing data, and other key considerations for using the harmonized data.

1.1 Using this Guide

This user guide should be used alongside the complementary resources available through the DataFirst Open Data Repository and the AFLEARN website.

The harmonized dataset has been designed to preserve traceability to the original source data. Users who require additional variables, documentation, or project-specific information should consult the accompanying resources listed below.

- The **DataFirst Open Data Repository** hosts the pooled dataset and its structured metadata, including the DDI codebook (study- and variable-level documentation, labels, frequencies, project information).

- The *linking-to-source* file is provided alongside the pooled dataset; this file contains the actual mapping between harmonized and original identifiers. It uses the school_id and student_id keys, organized by country, project, and round, to let you join non-harmonized variables directly.

- **AHEAD Study Catalogue** is the official registry of early-grade assessment surveys included in the AHEAD dataset. For each assessment survey, the catalogue provides country, year, project code, round/wave, languages, grades, study type, sub-task availability, survey-design setup (Stata svyset and R svydesign), and guidance for linking harmonised records back to original source files

Note: Understanding Projects and Constituent Studies

Throughout this guide, a distinction is made between projects and constituent studies. A project refers to a broader assessment programme, which may include one or more rounds of data collection. Each round is treated as a separate constituent study. Individual constituent studies are uniquely identified by the combination of country, project, and round, and this combination should be used when distinguishing between specific survey iterations. While harmonized variables share common names across projects, the values they contain remain specific to each constituent study and should be interpreted within their original survey context.

2. Objectives of Harmonization

The harmonization programme has three primary objectives:

- Build a flexible database that can integrate new studies as they become available, including non-USAID-funded national EGRA/EGMA assessments over time.
- Ensure consistent variable naming, structure, and documentation across datasets so that comparable fields mean the same thing in every project.
- Enable comparative and pooled analysis at multiple levels: pupil, school, project, and country.

3. Guiding Principles

The harmonization is guided by a set of principles that underpin the design and ongoing development of the pooled database. These principles define the intended characteristics of the harmonized dataset and guide future releases. These guidelines are aligned with the goals of the harmonization.

Table 1: African EGRA/ EGMA data harmonization principles

Principle	Meaning for users
Scalability	The harmonized database structure is designed to support the addition of variables (horizontal expansion) and new surveys (vertical expansion) in future releases.
Transparency	Harmonization decisions, variable mappings, and metadata are fully documented to ensure that users can understand how project variables have been transformed and interpreted. Harmonization decisions and metadata are documented through the portal **DDI codebook** and this guide.
Traceability	Every harmonized variable can be traced back to its original project dataset and supporting documentation. Each project links back to original microdata and documentation on DataLumos via the Project-Level Descriptions on the AFLEARN website.
Analytic flexibility	The harmonized dataset is designed to support a wide range of analytical applications, including cross-sectional pooling where variables align, within-project round comparisons, and school-panel analysis where programmes resurvey the same schools.

4. Prioritization Criteria for Inclusion

Project datasets are prioritised according to their potential analytical value, comparability with existing data, and contribution to the long-term objectives of the harmonization programme.

When expanding the harmonized pool, projects may be prioritised by:

- Recency of data collection
- National representativeness
- Availability of longitudinal follow-up at the child level
- Extended assessment formats (e.g. three-minute oral reading fluency instead of the standard one-minute task)
- Inclusion of EGMA modules alongside EGRA

5. Context: demographics, survey and study designs

The pooled dataset contains both harmonized EGRA and EGMA assessment variables and a set of harmonized contextual variables describing pupils, schools, survey design, and study characteristics.

While the assessment variables are discussed later in this document, this section focuses on the contextual variable groups. These variables provide the information required to identify individual projects, account for differences in sampling and assessment contexts, and support appropriate comparative analyses.
The following subsections provide guidance on their interpretation and use, while detailed variable labels, coding schemes, and frequencies remain available in the accompanying DDI codebook.

5.1 Demographics and pupil context

The harmonized dataset contains demographic and pupil context variables describing learner characteristics, assessment context, and identifiers.

Table 2: Demographic variables in the dataset

Variable	Description
gender	Pupil gender
grade	Grade / class level
age	Pupil age (continuous)
location	Urban / rural
lang_home	Home language
lang_assessment	Language of assessment
lang_instruction	Language of instruction
lang2_assessment	Second language of assessment
year, month	Assessment year and month
school_id, student_id	School and pupil identifiers (unique within each country, project and round combination)
consent	Consent / assent flag

5.1.1 Important considerations

Pooling across studies: Use the harmonized country variable for country-level analyses. Individual constituent studies are uniquely identified by a combination of country, project, and round, which should be used when distinguishing between specific survey iterations.

Sparsity: Not every study collected every contextual variable. Before conducting pooled analyses, users should consult the interactive data browser on the AFLEARN website.

Grade: Although grade values have been harmonized to a common set of labels, the distribution of grades varies across projects or individual studies. Users should make comparisons with reference to each project's stated target population and design. Note

that the same grade label does not imply equivalent curricula or learning expectations across countries.

Age: Age is retained as a continuous variable wherever possible. Missing or non-response values follow the harmonized missing-value conventions described in Section 10.

Location: Urban/rural coding is harmonized where collected. However, not all projects distinguish location.

Identifiers: The harmonized variables `school_id` and `student_id` are copied directly from the original datasets and are meaningful only within their individual studies. They should not be used to link observations across studies or survey rounds. The pooled dataset includes the harmonized identifier `harm_student_id`, described in Section 8.

5.2 Survey design, weights, and geography

The dataset includes a set of variables describing the sampling design, geographic coverage, and survey structure of each project. These variables enable users to correctly account for survey design when conducting weighted analyses and to identify the geographic context of each project.

Table 3: Harmonized survey-design and geography variables

Variable(s)	Description
country	Country
year	Year
admin_div1–admin_div3	Administrative division codes are provided for up to three hierarchical levels. The administrative units represented at each level vary by country—for example, provinces and counties in Kenya, and divisions and subdivisions in the DRC
admin_div1_type–admin_div3_type	Name of each administrative level in this study (e.g. province, county, division, region, district) — separates the code from what the level is called.
sample_design_stages	Number of sampling stages in the survey design.
stage1–stage4	Sampling unit (SU / cluster) at each stage.
strata1–strata4	Stratification grouping at each stage.
fpc1–fpc4	Finite-population correction (FPC) factor at each stage.
wt1–wt4	Stage weights
wt_final	Final weight — use this for design-based estimation.

5.2.1 Important considerations

Harmonized names, project-specific contents: Survey design variables share common names across projects, but values are individual study-specific.

Within-survey rule: Design variables and weights must be used one study at a time. Never declare a survey design on the full harmonized file.

Multistage design variables: The number of applicable stages, strata, fpc, and stage-weight variables depends on the value of sample_design_stages. Variables for stages beyond the number used in a project's survey design will not apply. For design-based analysis, use `wt_final` rather than the individual stage weights.

Geography (admin_div*): Administrative divisions are harmonized to `admin_div1`–`admin_div3`. Companion `admin_div*_type` fields document the local label (province, county, region, district, etc.) because administrative levels differ across countries.

No pooled representativity: Pooled descriptive statistics that ignore survey design are **not** nationally representative of any single country. Details of each study's sampling design are available in the **interactive bowser. Per-study survey set-up:** The correct weighting and clustering specification for each study is listed in the EGRA/EGMA Project and Coverage Explorer. Do not apply one design to the entire pooled file.

Known gaps: Some studies lack weights or complete design documentation; see the Release Notes for the current release for known gaps.

5.3 Study Design

These concepts describe **how a study was designed and fielded**. Full project narratives and sampling descriptions, collection occasions and links to original files are in the interactive browser on the AFLEARN website.

Table 4: Study design variables in the dataset

Concept	Meaning
study_type	Classifies the unit tracking structure of the study. Cross-sectional (cross sectional) = independent samples drawn at each round. School panel (school_panel) = the same schools tracked across rounds. Pupil panel (pupil_panel) = the same pupils tracked across rounds.
study_design	Describes the methodological design of the study. Pre – post (pre_post) = measures taken before and after an intervention without a control group. Quasi_experimental = non-randomised comparison group. RCT = randomised controlled trial. Descriptive = no intervention, monitoring or baseline only. Unknown = design not ascertainable from available documentation.
round	An integer indicating the sequential data collection occasion within a given project (1, 2, 3, ...), assigned in chronological order.
Treatment indicator	Identifies whether a record belongs to a treatment or control group. Control = in control group. Treatment = in treatment group. Treatment2 = in 2nd treatment group. Treatment3 = in 3rd treatment group. Not applicable = study has no intervention.
Level of randomization	The unit at which treatment and control groups were randomly assigned in experimental studies. Used to determine the appropriate level at which to cluster standard errors in treatment effect estimation.

6. EGRA and EGMA subtask variables

EGRA and EGMA modules make up the bulk of the harmonized variables. The harmonization framework standardises the naming and organisation of assessment variables across projects while preserving meaningful differences in assessment design and administration.

6.1 EGRA subtask descriptions

In Table 5 below, pedagogical descriptions of the EGRA subtasks are adapted from Dubeck and Gove (2015, Table 1, p. 318). Subtasks labelled as *core* represent the standard EGRA assessment modules, while the remaining subtasks reflect optional or extended modules included in selected EGRA implementations.

Table 5: Pedagogical descriptions of the EGRA subtasks

Subtask	**Subtask prefix**	**Skill**	**Description**
Letter name identification *core*	letter_name	Alphabetic principle	Learners identify upper- and/or lower-case letters by name within 60 seconds (LNCPM). Discontinued if no letters in the first line are correct.
Letter sound identification *core*	letter_sound	Alphabetic principle	Learners identify letters by corresponding sound within 60 seconds (LSCPM). Discontinued if no letters in the first line are correct.
Non-word reading *core*	invent_word	Alphabetic principle	Learners read aloud pseudowords within 60 seconds (CWPM). Discontinued if none of the first five words are correct.
Oral reading fluency *core*	oral_read	Oral reading fluency	Learners read a grade-appropriate passage within 60 seconds (CWPM). Discontinued if none of the first ~10 words are correct.
Listening comprehension *core*	list_comp	Oral language comprehension	Learners listen to a passage and answer comprehension questions. Untimed; no discontinuation rule.
Initial sound identification	pa_init_sound	Phonemic awareness	Learners isolate and pronounce the first sound of a word. Untimed.
Initial sound discrimination	pa_df_init_snd	Phonemic awareness	Learners identify the word with a different beginning sound among three words. Untimed.

Phoneme segmentation	pa_num_sound	Phonological awareness	Learners break words into individual sounds or syllables. Untimed; discontinued if no points in first five items.
Syllable identification	syllable_sound	Alphabetic principle	Learners identify syllables by sound within 60 seconds. Used in syllabic languages.
Familiar word reading *core*	fam_word	Word reading	Learners read high-frequency words within 60 seconds (CWPM).
Dictation	dict	Alphabetic principle	Learners write letter sounds, words, and/or sentences dictated by the assessor. Untimed.
Vocabulary (pointing)	vocabA	Receptive vocabulary	Learners point to body parts named orally. Untimed.
Vocabulary (spatial)	vocabB	Receptive vocabulary	Learners follow spatial instructions with objects. Untimed.
Vocabulary (naming)	vocabC	Expressive vocabulary	Learners name pictured objects. Untimed.
Reading comprehension *core*	read_comp	Reading comprehension	Learners answer questions on a passage; may allow look-backs. Untimed.
Maze	mazeA, mazeB	Comprehension	Learners select the best-fitting word from three options at regular intervals in a text.
Cloze	cloze	Comprehension	Learners supply missing words in a passage without options. Untimed.

6.2 EGMA Subtask Descriptions

In the table below, pedagogical descriptions are adapted from the *Early Grade Mathematics Assessment (EGMA) Toolkit* (RTI International, 2022). All subtasks represented are core EGMA assessment modules.

Table 6: pedagogical descriptions of the EGMA subtasks

Task	Subtask prefix	Skill	Description
Number identification	num_id	Number competence	Learners read numerals aloud (single-, two-, and three-digit) within 60 seconds.
Number discrimination	quant_comp	Numerical magnitude	Learners identify the greater number in 10 pairs of increasing difficulty. Untimed; discontinued after four successive errors.
Missing number	miss_num	Mathematical patterns	Learners identify missing numbers in sequences. Untimed; discontinued after four successive errors.
Addition level 1	addlvl1	Arithmetic	20 addition problems (addends ≤ 10, sums ≤ 19) within 60 seconds.
Subtraction level 1	sublvl1	Arithmetic	20 subtraction problems (values ≤ 19) within 60 seconds.
Addition level 2	addlvl2	Arithmetic	5 harder addition problems (sums ≤ 70). Not administered if Addition level 1 score is zero. Untimed.
Subtraction level 2	sublvl2	Arithmetic	5 harder subtraction problems (values ≤ 70). Not administered if Subtraction level 1 score is zero. Untimed.
Word problems	word_prob	Arithmetic and reasoning	Six oral word problems of increasing difficulty with counters available. Untimed; discontinued after four successive errors.

6.3 EGRA and EGMA subtask naming conventions

Every EGRA/EGMA subtask shares a **subtask stem**; individual fields add an item number or metric suffix. Within each **subtask stem**, you will typically find:

Table 7: EGRA/ EGMA subtask naming conventions

Field type	Pattern	Examples
Item-level	{subtask_stem}_{n}	letter_sound_1, num_id_3
Subtask summary	{subtask_stem}_{metric_suffix}	oral_read_score, oral_read_pm, fam_word_time_remain

The full set of metric suffixes is shown in the table below.

Table 8: Standard metric suffixes

Metric suffix	Description	Timed tasks	Untimed tasks
_score	Raw score	all	all
_pm	Correct items per minute	all	—
_score_pcnt	Percent correct	—	all
_attempted	Items attempted	all	reading comprehension
_attempted_pcnt	Percent correct of items attempted	all	reading comprehension
_score_zero	Indicator: score equals zero	all	all
_time_remain	Time remaining (seconds)	all	—
_auto_stop	Auto-stop rule triggered	all	—
_max	Maximum items in subtask	all	all
_time_allowed	Time allowed (seconds)	all	—

Item-level variables use the `corr` value-label family in the DDI codebook (typically 0 = Incorrect, 1 = Correct, with non-response codes).
Summary metrics (scores, rates, times) are usually continuous and unlabelled.
A full list of subtask stems and metric suffixes is provided in Annex A.
All variable-level codes and labels are documented in the portal **DDI codebook**.

7. Multiple instances of the same subtask

EGRA assessments sometimes administer the same subtask more than once, for example, in multiple languages, in timed and untimed formats, or using alternative reading passages.

Subtask administered in multiple languages
Where the same subtask is administered in multiple languages, a prefix of `l2` is added to indicate the second language of assessment. This second language of assessment is then specified in the `lang2_assessment variable`.

Different versions of subtask within the same language
Where a subtask is administered more than once in the same language (e.g. two different reading passages), the letter B (or C, D etc.) is added to the end of the subtask stem. For example, `oral_readB_score` indicates the number of correct words on the second reading passage.

Subtask variants

In cases where a standard subtask is administered in a substantively different way, a new subtask stem is created. For example, untimed oral reading is indicated by the `unt_oral_read` stem.

7.1 Important considerations

Availability: Not every subtask was administered in every project, and item counts differ. Use the interactive browser and **DDI codebook** to confirm coverage for your project of interest before pooling.

Wide versus long data: The harmonized dataset is distributed in **wide format**, with each harmonized variable represented by its own column. This structure supports straightforward pooling across projects. Analysts wishing to compare multiple versions of the same assessment (for example, first- and second-language oral reading tasks) may reshape the data into **long format** where appropriate for their analysis.

8. Identifiers

The harmonized dataset preserves the original individual study identifiers while also providing harmonized identifiers and project metadata to support data management, reproducibility, and traceability.

The pooled dataset includes the following identifiers and metadata:

Table 9:Identifiers in the pooled dataset

File	Identifiers and metadata
Main pooled data file	country, year, original school_id, original student_id, harm_student_id, country_iso, project_code, study_type, study_design, round, school_code (harmonized school id), pupil_code

The interactive browser provides more context on the project related variables.

8.1 Original identifiers (school_id, student_id)

school_id and student_id are retained in string format. See Section 5.1.1 for identifier scope and linkage limitations.

Although some projects followed the same schools over multiple rounds, school_id and student_id alone do not constitute pupil-level longitudinal identifiers. Whether longitudinal linkage at the pupil level is available depends on the release; consult the Release Notes for the current release for details.

Use the interactive browser to trace back to the original DataLumos files and their codebooks.

8.2 Harmonized student ID (harm_student_id)

The pooled dataset includes harmonized identifiers to facilitate data management within the pooled database. The variable `harm_student_id` provides a unique, fixed-width identifier for each pupil record. It encodes the project structure, the round of data collection, school, and pupil codes, creating a stable key for uniquely identifying records within the harmonized dataset.

The structure of `harm_student_id` is described in the table below:

Table 10: Harmonized identifier components

Position	Component	Width	Meaning
1–2	country_iso	2	ISO-3166 alpha-2 (e.g. KE, RW)
3–5	project_code	3	Zero-padded project code within country
6	round	1	Collection round within project (1–9)
7–9	school_code	3	Harmonized school rank within country+project+round
10–12	pupil_code	3	Harmonized pupil rank within school

Assignment logic (interpretive): `school_code` and `pupil_code` are derived from original `school_id` and `student_id` where present; missing pupil IDs receive sequential placeholders within the school. Codes are assigned **within** each `country_iso` + `project_code` + `round` combination - the same original school in two rounds receives a **different** `harm_student_id` because the round digit changes. By construction, `harm_student_id` is unique within each `country_iso` + `project_code` + `round`.

9. Missing values and unresolved codes

Missing-value and non-response conventions for each variable are documented in the portal **DDI codebook** (value labels and frequencies).

Annex A — Variable naming conventions

Harmonized names use **snake_case**. Every EGRA/EGMA subtask shares a **subtask stem**; individual fields add an item number or metric suffix.

Table 12. Standard EGMA subtask prefixes

Task	Core stem
Number identification	num_id
Number discrimination	quant_comp
Addition level I	addlvl1
Addition level II	addlvl2
Subtraction level I	sublvl1
Subtraction level II	sublvl2
Word problems	word_prob
Missing number	miss_num
Multiplication	mult
Fractions	frac
Shape identification	shape_id
Written exercise	we_add, we_sub, we_mult, we_div

Table 13. Standard EGRA subtask prefixes

Task	Core stem	Alternate prefixes in project data
Listening comprehension	list_comp	—
Letter name knowledge	letter_name	letter
Letter sound knowledge	letter_sound	—
Invented word reading	invent_word	—
Familiar word reading	fam_word	—
Oral reading fluency	oral_read	—
Reading comprehension	read_comp	—
Letter dictation	dict	dict_let, dict_letter, letter_dict
Word dictation	dict_word	word_dict
Sentence dictation	dict_sent	—
Vocabulary	vocabA, vocabB, vocabC, vocab_word	vocab
Oral vocabulary	oral_vocab	—
Maze / cloze	mazeA, mazeB	maze

Phonemic awareness — initial sound	pa_init_sound	—
Phonemic awareness — different initial sound	pa_df_init_snd	pa_diff_init_sound
Phonemic awareness — different final sound	pa_df_fnl_snd	—
Phonemic awareness — segmentation	pa_num_sound	pa_phon_sound, pa_sounds_id
Syllable reading	syllable_sound	syll_sound
Silly sentence	silly_sentence	—

Table 14. Common project-specific prefix patterns

Pattern	Typical meaning	Harmonized treatment
l2_, mt, mt1–mt3	Second language / mother tongue	Separate l2_* stem
unt, k_unt	Untimed administration	Separate stem (e.g. unt_oral_read)
lb, ext_lb, ilb	Look-back comprehension	Separate stem or read_compB variant
B, A on stem	Alternate passage or section	Separate stem (oral_readB, vocabB)
eq, k_eq, r_eq	Baseline-adjusted equivalent score	**Excluded**
i, i1–i3	Imputed values	**Excluded**
t, ht, sso	Teacher / head teacher / staff	**Excluded** (non-pupil)