

Data Merge Guide

The Thrive by Five Index 2024 study includes variables on early childhood development outcomes and primary caregivers, as well as assessed children's home environments. For children enrolled in Early Learning Programmes (ELPs), the dataset also contains information on ELP quality, principals, practitioners, and facilities.

Data were collected electronically using computer-assisted personal interviewing (CAPI) across multiple survey instruments ("forms"), each capturing information at different levels: child-level, ELP-level, and primary caregiver-level.

To reduce file size and complexity, the public-use data for each instrument are shared separately. Researchers wishing to analyse variables across multiple instruments will need to merge the relevant datasets using the relevant identifier, `id_child` or `id_ecd`.

Levels of Data Collection

- Child-level datasets: One record per assessed child
- ELP-level datasets: One record per early learning programme
- Each ELP may be associated with multiple children, resulting in different merge structures depending on the datasets being combined.

Merge Structures

- One-to-one (1:1) merges are required when combining child-level and primary caregiver-level datasets (e.g. ELOM 4&5 and Primary Caregiver Interview).
- One-to-many (1:m) merges are required when combining child-level datasets with ELP-level datasets (e.g. ELOM 4&5 with LPQA v2), as multiple children are assessed within the same ELP.

Important Notes:

Practitioner analysis:

- At ELPs where the principal was also the sole practitioner, only the Principal Interview was conducted, and the Practitioner Interview was not administered. Many variables included in the Practitioner Interview are also captured in the Principal Interview. Consequently, to analyse practitioner-level variables, analysts should merge the relevant variables from the Principal Interview, **for principals who were the sole practitioner at their ELP only**, dataset with the Practitioner dataset to avoid unnecessary missing data.
- To support analysis of practitioner-level variables, we have included a Stata do-file titled *Program - Merge practitioner and sole principal.do*. This script merges the Principal

Interview and Practitioner Interview datasets, incorporating practitioner-related information captured in the Principal Interview for cases where the principal was the only practitioner at the ELP. This helps ensure that these variables are not unnecessarily missing from the final practitioner-level dataset.

- In the Principal Interview dataset there are **165 observations** where the principal is also the only practitioner at the ELP.

Non-enrolled children:

- Data collected for non-enrolled children (household-based ELOM 4&5, SEF assessments and Primary caregiver interviews) cannot be merged with ELP-level datasets, as no ELP information was collected for these children as they were not enrolled.

Dataset Identifiers

The table below lists each dataset, its level of data collection, and the identifiers available for merging.

Instrument	Dataset name	Level of data collection	Identifiers
ENROLLED SAMPLE:			
ELOM 4&5 Years Assessment Tool (ELOM 4&5)	2024tb5_elom_sef_pub	Child	id_child, id_ecd
Social-Emotional Functioning (SEF) Rating Scale	2024tb5_elom_sef_pub	Child	id_child, id_ecd
Learning Programme Quality Assessment (LPQA v2)	2024tb5_lpqa_pub	ELP	id_ecd
Facility Observation Form	2024tb5_facility_pub	ELP	id_ecd
Principal Interview	2024tb5_principal_pub	ELP	id_ecd
Practitioner Interview	2024tb5_practitioner_pub	ELP	id_ecd
Primary Caregiver Interview	2024tb5_pcg_pub	Child	id_child, id_ecd
NON-ENROLLED/HOUSEHOLD SAMPLE:			
ELOM 4&5 Years Assessment Tool – Non-enrolled children	2024tb5_elom_sef_hh_pub	Child	id_child
Social-Emotional Functioning (SEF) Rating Scale – Non-enrolled children	2024tb5_elom_sef_hh_pub	Child	id_child

