Sampling Weight Guide

The Thrive by Five Index 2024 is based on a statistically representative sample of:[1]

- South Africa's children aged 4–5 years who are enrolled in early learning programmes, and
- South Africa's ELPs.

To ensure representativeness, each public-use dataset includes one or more sampling weight variables. These weights must be applied during analysis to produce unbiased population estimates and correct standard errors for inference.

The survey used a three-stage, stratified, cluster sampling design, with weights constructed as the inverse of the probability of selection at each sampling stage.

This guide is intended to assist researchers in correctly selecting and applying the sampling weights when analysing the data. Sampling weights should be used for all analyses aimed at producing population-level estimates. For detailed documentation of the sampling design and weight construction, readers are referred to the *Thrive by Five 2024: Sampling Strategy* document and the *Thrive by Five Index 2024: Technical Report*.

Sampling design and selection probabilities

Sampling occurred in three stages:

1. Ward selection (within strata, in most cases a ward constitutes the primary sampling unit, but in a very small number of cases, two or more ward were combined to form a primary sampling unit)
2. ELP selection (within wards)
3. Child selection (within ELPs)

Ward selection probability:

$$P(ward\ selected) = \frac{\#\ wards\ selected\ in\ stratum \times \#\ Grade\ 3\ learners\ in\ ward}{Total\ \#\ Grade\ 3\ learners\ in\ stratum}$$

[1] The non-enrolled child and primary caregiver samples are not nationally representative, and therefore do not have survey weights. See *Tb5 2024 Non-enrolled Selection Note*.

ELP selection probability:

$$P(ELP\ selected) = \frac{\#\ ELPs\ selected\ in\ ward}{Total\ \#\ ELPs\ in\ ward}$$

Child selection probability:

$$P(child\ selected) = \frac{\#\ children\ selected\ in\ ELP}{\#\ eligible\ children\ in\ ELP}$$

Final selection probability:

$$P(child\ selected) = P(ward) \times P(ELP) \times P(child)$$

The overall probability that a child is selected is the product of the probabilities at each stage. The child-level sampling weight is the inverse of this probability.

Weights by level of analysis

- Child-level datasets use weights that incorporate all three sampling stages.
- Primary caregiver-level dataset uses weights that incorporate all three sampling stages.
- ELP-level datasets use weights based on ward and ELP selection only.

The clustering structure is hierarchically nested:

$$id\_child \subset id\_ecd \subset id\_ea$$

Weighting variables included in each dataset

Instrument	Dataset name	Dataset name	Sampling units	Stratification	Weight
Child	ELOM 4&5 Years Assessment Tool (ELOM 4&5) Social-Emotional Functioning	2024tb5_elom_sef_pub	id_ea – Ward (Primary sampling unit) id_ecd – ELP, sampled	stratum	weight_child

	(SEF) Rating Scale		within EAs (Secondary sampling unit)		
Primary caregiver	Primary Caregiver Interview	2024tb5_pcg_pub	id_child – Child, sampled within ELPs (Tertiary sampling unit)		weight_pcg
ELP	Facility Observation Form Learning Programme Quality Assessment (LPQA v2) Practitioner Interview Principal Interview	2024tb5_facility_pub 2024tb5_lpqa_pub 2024tb5_practitioner_pub 2024tb5_principal_pub	id_ea – Primary sampling units id_ecd – ELP: Sampled within EAs		weight_elp

Working with the sub-sample of children who have primary caregiver interviews:

Not all child observations have an accompanying primary caregiver observation due to unit non-response, that is, not all primary caregivers of assessed children were interviewed. For this sub-sample, analysis of either primary caregiver data on its own or analysis of child-level data combined with primary caregiver data, weight_pcg must be used.

The non-representative sample of non-enrolled children and their primary caregivers:

The data collected from non-enrolled children were not designed to be nationally representative. Accordingly:

- No survey weights are provided
- These data cannot be used to produce representative population estimates.

Worked examples:

Here we provide a worked example of how to apply the survey design settings for the child level analysis (e.g., ELOM 4&5) and for ELP level analysis (e.g., principal interview) using either R or Stata:

Unit of analysis	R	Stata

Child	library(survey)
	
	design <- svydesign(
	  ids = ~id_ea + id_ecd + id_child,
	  strata = ~stratum,
	  weights = ~weight_child,
	  data = tb5_analysis,
	  nest = TRUE
	)
		svyset id_ea [pweight = weight_child], strata(stratum) || id_ecd || id_child

ELP	library(survey)
	
	design <- svydesign(
	  ids = ~id_ea + id_ecd,
	  strata = ~stratum,
	  weights = ~weight_ecd,
	  data = tb5_analysis,
	  nest = TRUE
	)
		svyset id_ea [pweight = weight_elp], strata(stratum) || id_ecd

Primary caregiver	library(survey)
	
	design <- svydesign(
	  ids = ~id_ea + id_ecd + id_child,
	  strata = ~stratum,
	  weights = ~weight_pcg,
	  data = tb5_analysis,
	  nest = TRUE
	)
		svyset id_ea [pweight = weight_pcg], strata(stratum) || id_ecd || id_child