Technical Note: Analytical approach for the analysis of the learning outcomes in Nyanda

Date: 24th October 2025
Prepared for: Ministry of Education, Technical Team
Author: UNICEF Innocenti, Education Unit
Source: Education dataset

Executive summary

The objective of the analysis is to equipe the Ministry of Education of Nyanda with the capacities to assess and monitor the performance and improvement in reading and mathematics score for grade 1-6 students. The analysis focuses on understanding the main inequality issues among students in Nyanda.

Data preparation and transformation

The source of the data is the education dataset of Nyanda for the period 2023-2025. The data used for the analysis is composed of four distinct datasets (district, school, students, teachers) in wide format covering the years 2023 to 2025.

The first step for the analysis consists in necessary data quality checks, followed by transforming each dataset into long format to prepare them for the merge into a complete student-year panel dataset.

Two major data transformations were implemented before proceeding with the analysis:

1. <u>Students weighting:</u> the dataset has weights for the year 2023. These weights were used for all years to make the analysis representative of the population. The decision to use the 2023 weights for all years was taken because weights are strongly negatively correlated with the income level variable (correlation: -0.9), suggesting that lower income students are underrepresented. To ascertain whether it was plausible to use the same weights for all three years, attrition and sample replacement were inspected. The results of the analysis show that neither of the two are present and therefore the weights were used for all three years.
2. <u>Standardization of test scores:</u> Reading and mathematics scores were standardized to make them comparable across subjects, grades, years and to be able to interpret regression results in standard deviation units instead of row test score points that would make comparability more difficult.

Description of the models and specifications used

Two models were used for the analysis:

1. Simple OLS model at the student level with year fixed effects and cluster at school level for both test scores

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_1 X_{1it} + \cdots + \beta_1 X_{kit} + \varepsilon_{it}$$

Y_{it} = standardized test score for student i in year t
X_{it} = explanatory variables for student i in year t
ε_{it} = error term capturing unobserved factors for student i in year t

   a. The OLS model with year fixed effect was used to estimate baseline relationship between student characteristics and test performance. The fixed effect was included to control for time-specific shocks (e.g., curriculum changes, policy reforms and others). To account for

within school correlation among students, the OLS regression included cluster standard errors.

2. Multilevel model with random intercept for school, district, and student-level variance

$$Y_{isd} = \beta_0 + \beta_1 X_{isd} + u_d + v_s + \varepsilon_{isd}$$

Y_{it} = standardized test score for student i in school s and district d

X_{isd} = explanatory variable for student i in school s and district d

u_d = district level random effect, representing district level variation

v_s = school level random effect, representing school level variation

ε_{isd} = student level residual, representing individual unexplained variation for student i in school s and district d

    a. The multilevel model was used to explicitly model the hierarchical structure of schools which is defined by students (level 1), schools (level 2), districts (level 3). The model also includes random interceptions for schools and districts to account for variation in test scores between schools, districts, and within schools (between individuals in the same school and district). These specifications are very important because they help quantify some of the drivers of the variance in students' test scores that are not accounted for by an OLS regression.

Limitations of the OLS model:

1. Independent assumptions: The OLS assumes that all the observations are independent, but this is rarely the case for educational data where students share schools, district environments (teachers, infrastructure, peer effects among others). This can lead to biased standard errors.
2. Hierarchical structure of the education system: the educational system is hierarchical in nature, and the OLS cannot account for the variation between schools and within schools.
3. Omitted variable bias: OLS can attribute variation to individual level variables if important district and/or school level variables are omitted.
4. Uniform assumptions of the effects across contexts: The OLS ignores that effects on test scores may vary by context and treat relationships between the dependent (test scores) and the independent (all covariates) variables as the same in all schools and districts.

Limitations of the multilevel model:

1. Complexity and interpretation: this model is more complex in the interpretation of the variance component because of the hierarchy intrinsic in the model. This makes it more complex to interpret how much of the total variance is attributable to each single covariate.
2. Data requirements: to have reliable estimates the model needs a large sample at each level.
3. Model specification sensitivity: the results depend heavily on the hierarchy specified and misspecification can bias the results.
4. Still correlational: like OLS, multilevel models are not causal models. They can identify association and correlations but not causality without strong assumptions or an experimental design.

	OLS - Reading (z)	OLS - Math (z)	Multilevel - Reading	Multilevel - Math
(Intercept)	-0.264***	-0.275***	-0.267***	-0.276***
	(0.006)	(0.006)	(0.007)	(0.007)
factor(household_income_quintile)2	0.171***	0.184***	0.169***	0.186***
	(0.005)	(0.005)	(0.005)	(0.005)
factor(household_income_quintile)3	0.351***	0.354***	0.351***	0.354***
	(0.006)	(0.006)	(0.006)	(0.006)
factor(household_income_quintile)4	0.532***	0.534***	0.533***	0.541***
	(0.008)	(0.008)	(0.008)	(0.008)
factor(household_income_quintile)5	0.711***	0.710***	0.714***	0.718***
	(0.012)	(0.012)	(0.012)	(0.012)
female	0.006	0.003	0.007	-0.002
	(0.004)	(0.004)	(0.004)	(0.004)
rural	-0.264***	-0.262***	-0.255***	-0.257***
	(0.004)	(0.005)	(0.005)	(0.005)
disability_statusYes	0.017	0.018+	0.017	0.018+
	(0.011)	(0.011)	(0.011)	(0.011)
parent_education_levelPrimary	0.009+	0.017**	0.007	0.017**
	(0.005)	(0.005)	(0.005)	(0.005)
parent_education_levelSecondary	-0.005	0.004	-0.005	0.003
	(0.006)	(0.006)	(0.006)	(0.006)
parent_education_levelTertiary	-0.009	-0.009	-0.010	-0.009
	(0.010)	(0.010)	(0.010)	(0.010)
attendance_centered	0.849***	0.858***	0.818***	0.817***
	(0.036)	(0.036)	(0.035)	(0.035)
factor(year)2024	0.119***	0.122***	0.119***	0.122***
	(0.005)	(0.005)	(0.005)	(0.005)
factor(year)2025	0.240***	0.244***	0.239***	0.244***
	(0.005)	(0.005)	(0.005)	(0.005)
SD (Intercept school_id)			0.315	0.321
SD (Intercept district_id)			0.000	0.000
SD (Observations)			0.955	0.956
Num.Obs.	192000	192000	192000	192000
R2	0.089	0.089		

- p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001