Northern Uganda Literacy Project
8-year Follow-up
Analysis Plan

Julie Buhl-Wiggers, Jason Kerwin, Jeff Smith, and Rebecca Thornton
February 17, 2022

This document spells out the main analyses we plan to perform using the 8-year follow-up data on the Northern Uganda Literacy Project (NULP). In addition to these prespecified analyses, we also expect to conduct further exploratory analyses beyond the ones described here. Those analyses will flow from one or more of further reflection on our part, developments in the related literature, or unexpected patterns in the data. They may also involve extending the sample to the other cohorts in our data and may not consist solely of estimating experimental mean impacts.

<u>Sample</u>

We will perform the majority of our planned analyses using only Cohort 2 students—students who entered grade one in 2014, who were treated in grades P1 to P3 (1^{st} to 3^{rd} grade) in 2014-2016. This plan is for analyzing the 8-year followup data on student from that cohort, which will be collected in early 2022. Our analytic sample will include all Cohort 2 students for whom we have data.

<u>Main Outcomes</u>

The outcomes we plan to analyze in our long-term impact analysis are as follows:

1. Leblango EGRA (Early Grade Reading Assessment) score (in SDs)
   - Score is weighted by the first principal component of the control-group data across all Leblango EGRA components we tested in this wave of data collection, for every student in Cohort 2
   - SDs are relative to the control-group distribution
2. English EGRA score (in SDs)
   - Score is weighted by the first principal component of the control-group data across all English EGRA components we tested in this wave of data collection, for every student in Cohort 2
   - SDs are relative to the control-group distribution
3. EGMA (Early Grade Mathematics Assessment) score (in SDs)
   - Score is weighted by the first principal component of the control-group data across all EGMA components we tested in this wave of data collection, for every student in Cohort 2
   - SDs are relative to the control-group distribution
4. Attended school in 2021 (percentage points)
   = 1 for attended any school, including primary or secondary

= 0 otherwise

5. Attended secondary school in 2021 (percentage points)
   = 1 for attended a grade higher than P7
   = 0 if attended P7 or below, or did not attend school
6. Ever had sex (percentage points)
7. First had sex at age 13 or below
   = 0 if never had sex
   (The median student in the sample turned 13 in 2020, when school closures first began)
8. Worked outside of the home in 2021 (percentage points)
9. Worked outside of the home in a non-agricultural sector in 2021 (percentage points)
   = 0 if did not work outside of the home

All of these outcomes will be measured for Cohort 2.

Following common practice in the U.S. educational program evaluation literature we divide our planned analyses into confirmatory and exploratory analyses.

Specifically, we divide these outcomes into three groups:

Group 1: confirmatory academic outcomes (1, 2, 3)
Group 2: confirmatory downstream outcomes (4, 5)
Group 3: exploratory outcomes (6, 7, 8, 9)

* numbers in parentheses correspond to the numbered outcome variables in the list.

** This is not a comprehensive list of all of the outcomes we will examine.

Obtaining impact estimates

We will obtain experimental impact estimates for each of our outcomes y_{ij} via the following parametric linear model estimated by ordinary least squares:

$$y_{ij} = \beta_0 + \beta_1 FC_j + \beta_2 RC_j + Z_j'\tau + X_i'\gamma + \varepsilon_{ijt} \quad (1)$$

where i indexes students, which are nested within their original schools (as of P1) indexed by j. FC_j and RC_j are indicators for a school being randomly assigned to the Full-Cost and Reduced-Cost program, respectively. Z_j is a vector of indicators for the stratification cells used in the lottery that assigned schools to study arms. (1) is the specification we will use for all our confirmatory analyses, and we will also use it as our default approach for estimating average treatment effects in any exploratory analyses.

In (1), X_i is a vector of control variables; we will control for an indicator for being male, age as of the baseline exam inputted as categorical indicators for each age[1], and baseline test score indices for three exams (the Leblango EGRA, a math assessment, and an oral English exam). The baseline exams were conducted at the beginning of the 2014 school year. For students with missing values of the baseline exam score, we will replace the missing values with zero and include a separate indicator variable for the baseline exam score being zero.

We will construct each of the baseline score indices in a way that parallels the outcome measures: we will take the first principal component of test scores across all test modules for the control group data for Cohort 2 students at the 2014 baseline. We will then use those weights to construct weighted average scores for the entire sample. We will standardize the baseline score indices relative to the control-group distribution.

Inference

We will conduct inference on our estimates via randomization inference. Specifically, we will randomly permute the study arm assignments of each school within the stratification cells used in the original lottery. We will implement this in Stata via the `ritest` command, following the approach in Kerwin and Thornton (2021).

Null hypotheses

We plan to consider three null hypotheses for each outcome we study:

$H_0^1: \beta_1 = 0$

$H_0^2: \beta_2 = 0$

$H_0^3: \beta_1 = \beta_2$

We will likely consider further nulls (e.g. whether the population value of one or both of the mean impacts exceeds the level required to pass a cost-benefit test, for example) in our exploratory work.

Confirmatory and exploratory analyses

Our confirmatory analyses will consist of estimating impacts using equation (1) for the outcomes in Groups 1 and 2 and testing null hypotheses H_0^1 and H_0^2 using these estimates.

Our exploratory analyses will include two types of tests. First, we will test null hypothesis H_0^3 using experimental estimates of impacts on the outcomes in Groups 1 and 2 obtained using

[1] The age categories we will use are 5 and below, 6, 7, 8, 9 or above, and missing. Values below 5 will be bottom-coded as 5; values above 9 will be top-coded as 9. Missing age information at the baseline survey will be coded as a separate category.

equation (1). Second, we will estimate experimental mean impacts on the outcomes in Group 3 using equation (1) and use those estimates to test all three hypotheses for these outcomes.

Multiple Testing

We will take account of multiple hypothesis testing for conducting our confirmatory analyses using the Benjamini, Krieger, and Yekutieli (2006) method to compute sharpened *q*-values that control the false discovery rate (FDR). We will use the Anderson (2008) implementation of their approach, which computes the lowest value of the sharpened *q*-value for which we can reject the null, so that our *q*-values can be interpreted in the same way that conventional *p*-values are.

We plan to use the method separately by domain. That is, we will do multiple testing corrections across one group of six tests—two tests for each of the three outcomes in Group 1—and one group of four tests—two tests for each of the two outcomes in Group 2.

We will not undertake formal multiple testing procedures for our exploratory analyses, but will remind readers of the issue in interpreting those analyses.

References

Anderson, Michael. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103(484): 1481-1495.

Benjamini, Yoav, Abba Krieger, and Daniel Yekutieli. 2006. "Adaptive Linear Step-Up Procedures that Control the False Discovery Rate." *Biometrika* 93: 491–507.

Kerwin, Jason, and Rebecca Thornton. 2021. "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures." *Review of Economics and Statistics* 103(2): 251–264.