Complementary Components? Experimental Evidence on Bundling Health and Education Interventions

Pre-Analysis Plan

Noam Angrist* Dahyeon Jeong† Simone Lombardini‡
Theresa Lubozha§ Dan Maggio¶ Hannah Uckat‖

August 19, 2026

Abstract

At scale, government-implemented educational programs are plagued by low attendance rates that call into question whether estimates from the academic literature generalize. Within 125 primary schools in Zambia's Luapula Province, we find that only 23 percent of children attended a Catch Up session in the prior week and that 34 percent of schools held no session at all. To address these compounding compliance frictions, we cross-randomize Zambia's Home Grown School Meals program with a small teacher-level session incentive program that reimburses teachers with mobile airtime in exchange for holding TaRL sessions. We expect these interventions to boost TaRL session attendance by both improving children's school attendance and increasing the fidelity with which teachers hold sessions. We estimate the effects of both interventions and their complementarities on children's attendance and foundational literacy and numeracy scores.

*Oxford University noam.angrist@bsg.ox.ac.uk.
†World Bank dahyeonjeong@worldbank.org.
‡World Food Programme simone.lombardini@wfp.org.
§VVOB – Education for Development Theresa.Lubozha@vvob.org.
¶**Corresponding Author:** Rutgers University daniel.maggio@rutgers.edu.
‖World Bank huckat@worldbank.org.

1 Introduction and Overview of Study Design

The objective of this project is to implement a cluster randomized controlled trial analyzing whether two complementary interventions, a school meals program and a small teacher incentive, can increase children's attendance in Zambia's national Teaching at the Right Level (TaRL) program, known locally as Catch Up, and thereby improve foundational literacy and numeracy. The trial is embedded within the scale-up of Zambia's Home Grown School Meals (HGSM) program in Luapula Province, in partnership with the Zambian Ministry of Education, the World Food Programme, the World Bank, and VVOB – Education for Development. This document is intended to pre-register our analysis for the trial.

The motivation for this trial is that, while TaRL has repeatedly been shown to be effective, attendance is often low at scale within government programs. We confirm this fact with our baseline survey, which took place in 125 schools between November and December of 2025. In our survey, only 23 percent of sampled children attended a Catch Up session in the seven days prior. Non-attendance seems to reflect compounding constraints that prevent children from attending school (34 percent of children report being absent from school at least one day in the past week) and teachers from holding class (schools held an average of 1.6 TaRL sessions per week against 2.8 planned, with 43 of 125 schools holding none at all).

Our two interventions target these frictions in isolation, and we cross-randomize to estimate the extent to which it is necessary to overcome frictions at both levels to increase compliance with TaRL programming.

In the first intervention, the HGSM program, participating schools receive maize flour from the National Food Reserve Agency, together with funding to purchase perishable items such as beans on the spot market, and provide each enrolled child one mid-day meal each school day. The program is being scaled nationally with support from the World Food Programme and the World Bank. Logistical constraints prevent immediate saturation, which created the opportunity to randomize the scale-up. Within our sample, 63 schools were early receivers of the program beginning in March 2026, while 62 schools were randomly assigned to be enrolled beginning in January 2027.

The second intervention provides Catch Up teachers with a small incentive for every TaRL session they hold. Teachers submit a photograph of the learning-level register that accompanies each Catch Up class, through a WhatsApp-based chatbot, as verification that the class took place. In exchange, they receive 20 ZMW (approximately 1 USD) of mobile airtime per session, for up to three sessions per week, paid bi-weekly. We randomly select 20 percent of submitted photographs to be manually verified by our research partner VVOB, and a quarter of that subset is additionally verified through phone calls to the school's head

teacher. Unsatisfactory submissions may be resubmitted, and failed resubmissions are not paid.

2 Sample and Data Collection

2.1 Sample and Randomization

The study sample includes 125 government primary schools in three districts of Luapula Province slated for the scale-up of the HGSM program: Mwense, Chembe, and Samfya. Within each school, we randomly sampled ten grade three children to participate in our baseline survey, conducted between November and December of 2025; in total, the baseline includes 1,249 children and 125 head teachers. Our study intends to create a panel of children and, as such, will aim to resurvey the same children.

Our cross-randomization took place in two parts. First, schools were randomly assigned to be enrolled in the HGSM program immediately (63 schools, from March 2026) or in January 2027 (62 schools, our control), stratified by district and urban or rural location. Second, we randomly assigned schools to the teacher incentive intervention, crossed with the first randomization and stratified by district, urban or rural location, and HGSM assignment. This results in four roughly equal-sized arms of 31 to 32 schools each:

- **Pure control** — enrolled in HGSM in January 2027; no incentive intervention.
- **HGSM only** — enrolled in HGSM in early 2026; no incentive intervention.
- **Incentive only** — enrolled in HGSM in January 2027; receives the incentive intervention.
- **HGSM + incentives** — enrolled in HGSM in early 2026; receives the incentive intervention.

Across a range of baseline indicators at both the school and child level, there is no evidence of imbalance across arms. The one exception is that schools enrolled early in HGSM held an average of 1.89 TaRL sessions per week at baseline, relative to 1.37 in schools slated for later enrollment, and children in early-enrollee schools attended 0.563 sessions per week relative to 0.403. Because this imbalance falls along the exact margin our interventions target, all treatment effect estimation uses an ANCOVA specification that controls for the baseline value of the outcome, as described in Section 3.

2.2 Data

Baseline survey. Conducted in all 125 schools between November and December 2025. A head teacher survey covered school characteristics, Catch Up implementation, and school feeding. A child-level survey covered attendance and food security, and included foundational skills batteries: versions of the Early Grade Reading Assessment (EGRA) and Early Grade Mathematics Assessment (EGMA) adapted to the local context.

Endline survey. The endline will repeat both instruments in all 125 schools in November and December 2026, adding anthropometric measurement. Because our interventions specifically target attendance, a school-based survey creates the possibility of differential attrition; to minimize it, the endline includes 24-hour revisits so that children are not excluded due to one-off non-attendance. Robustness checks for this potential differential attrition are pre-specified in Section 3.7.

3 Analysis

3.1 Treatment Effects

Our main treatment effect estimation follows an ANCOVA specification:

$$y_{ij1} = \alpha + \beta_1 M_j + \beta_2 I_j + \beta_3 (M_j \times I_j) + \gamma y_{ij0} + \delta_d + \varepsilon_{ij} \tag{1}$$

where y_{ij1} is the endline outcome for child i in school j; M_j is an indicator for assignment as an early enrollee of the HGSM program; I_j is an indicator for enrollment in the incentive intervention; y_{ij0} is the baseline value of the outcome, included where the outcome was collected at baseline; and δ_d are strata (district, urban/rural) fixed effects. Standard errors are clustered at the school level, the unit of randomization for both interventions. For the number of TaRL sessions held, which is measured at the school level, we estimate Equation 1 at the school level with Huber-White heteroskedasticity-robust standard errors.

Our parameters of interest are β_1 and β_2, which identify the effect of each intervention relative to the pure control, and β_3, which identifies the complementarity between the HGSM and incentive interventions. We will also report the linear combination $\beta_1 + \beta_2 + \beta_3$, which identifies the effect of the full package relative to control.

3.2 Treatment on the Treated Effects

If we indeed find increases in session delivery and TaRL attendance, the randomization of our intervention allows us to estimate the effect of TaRL instruction on children's foundational literacy and numeracy within a 2SLS framework. Our preferred specification uses the incentive intervention alone as an excluded instrument. Let a_{ij1} denote child i's exposure to TaRL at endline, measured either as an indicator for having attended any session in the reference week or, in a dosage specification, as the number of sessions their school held. The first stage is:

$$a_{ij1} = \alpha + \phi I_j + \psi M_j + \gamma a_{ij0} + \delta_d + \nu_{ij} \tag{2}$$

and the second stage is

$$y_{ij1} = \alpha + \pi \widehat{a}_{ij1} + \psi M_j + \gamma y_{ij0} + \delta_d + \varepsilon_{ij} \tag{3}$$

where $\widehat{a}_{ij1}$ is predicted from Equation 2. Since M_j is orthogonal to I_j, allowing it to enter as a control allows us to use assignment to our incentive intervention as an instrument across our entire sample. π then identifies the effect of TaRL on children who receive instruction as a result of our incentive intervention. Since we expect this intervention to primarily impact TaRL attendance on the margin of whether sessions are held, we do not expect this population of "compliers" to be selected *within* school. Thus, π recovers the treatment on the treated effect of TaRL instruction across the population of children who attend school. This creates a comparable parameter to our motivating academic studies in the spirit of Angrist and Meager (2023).

The exclusion restriction for this estimation requires that the incentive affects learning only by changing whether TaRL sessions occur. The most plausible violation runs through teacher motivation, as the payment may improve pedagogical quality even in classes that would have been held regardless. If this were the case, the interpretation of π would diverge from the treatment-on-the-treated effect to a composite effect of the ToT and the effects of improving pedagogical quality in the classroom. Such a parameter would still be of value but less comparable to the academic literature on TaRL. While this is fundamentally untestable, our classroom observation, while exploratory, will allow us to construct measures of pedagogical quality within the classroom to test for improvements in teaching as a result of the incentives.

In addition to this primary 2SLS specification, we will report the results of a specification that adds the school meals treatment as a second excluded instrument. This exercise is different in two respects. First, meals shift compliance at the level of the child rather than

the class, so the set of "compliers" becomes the set of marginal children who would not have received TaRL instruction if not for our intervention. It is reasonable to expect that these children are, on average, of lower socio-economic status than those who attend school and remedial instruction with high fidelity. This means that the predicted sign and magnitude of π will be ambiguous. On the positive side, the "marginal" child may have greater need but also face greater learning impediments, meaning the effects of TaRL may be larger or smaller than many academic estimates for this set of children. On the negative side, some have hypothesized that the global learning crisis is the result of overcrowding in school. To the extent that our interventions have the potential to increase crowding, it is possible that these estimates of π are negative. The estimates of π are relevant to policy in that they can tell us if we expect the gains to the marginal dollar spent on attendance to be positive or negative (locally). But also for academia, because they allow us to build a heuristic for whether we expect the non-attendance observed in studies like de Barros and Lubozha (2026) to bias the internally valid RCT estimates above or below an externally valid population treatment effect.

Second, however, the inclusion of school meals as an instrument means that the exclusion restriction is considerably more fragile. As we've discussed throughout this document, meals plausibly affect learning through food security and attentiveness whether or not the child attends a TaRL session. For this reason, we consider this estimate exploratory and secondary to the strategy reported in equations 2 and 3, especially if the hypothesized effects on food security and nutritional status materialize.

3.3 Primary Outcomes

For our analysis, we pre-specify four primary outcomes, all of which are measured in the endline survey. Three of our outcomes (school attendance, food security, and TaRL session attendance) are measured at the child level. One (Sessions held) is measured at the school level.

1. **School attendance.** We will measure children's attendance as the number of days in the seven prior to our endline survey that a child has been absent from school. Here, an indicator variable that takes the value one if a child is present on the day we attempt to survey serves as both a measure of differential attrition and the effects of the intervention on attendance.
2. **Food security.** We measure children's food security using the child food insecurity experience scale and a dietary diversity score, both collected at baseline and endline in the child survey. We report effects on each measure and on a standardized summary

index of the two (Anderson, 2008).

3. **TaRL session attendance.** An indicator equal to one if the child reports attending at least one TaRL session, in either subject, in the seven days preceding the endline survey. This is the same question used at baseline and in de Barros and Lubozha (2026). We will also disaggregate this indicator across literacy and numeracy sessions.
4. **Session fidelity.** We will measure session fidelity in two ways. The first will be an indicator variable that takes the value one if a school held *any* TaRL sessions in the seven days prior to our survey. The second will be a continuous variable that measures the number of TaRL sessions a school held in the seven days prior to our endline survey. TaRL sessions are reported in the head teacher survey and, as with TaRL attendance, we will report the effects on these outcomes aggregated across literacy and numeracy sessions, as well as disaggregated across subject.

3.4 Secondary Outcomes

In addition to our primary outcomes, we pre-specify several secondary outcomes which are causally downstream from our main outcomes.

- **Literacy.** Literacy in our survey is measured using the Early Grade Reading Assessment. Following Chimbutane et al. (2026), we test the effects of our intervention on the raw scores on each of the EGRA sub-domains: Letter identification, familiar words, listening comprehension, reading fluency, and reading comprehension. As in Chimbutane et al. (2026) we will also report the effects on a set of dummy variables that take the value one if a child records a positive score within a given subdomain. Finally, we will deviate from Chimbutane et al. (2026) by also reporting the effect on children's total EGRA raw score, and a summary weighted index (Anderson, 2008) by domain and collectively across the items in the instrument.

- **Numeracy.** Our measurement of children's numeracy follows a symmetric structure as for literacy. Here we use the Early Grade Mathematics Assessment, where the domains are: Number Identification, Number Discrimination, missing number, addition and subtraction, and word problems.

- **Nutritional status and wellbeing.** Beyond the food security measures we treat as primary, we will report several further indicators of children's nutritional status and wellbeing, where we may expect the HGSM program to confer non-learning benefits. These outcomes are height-for-age and BMI-for-age z-scores and the psychosocial wellbeing index.

3.5 Multiple Hypothesis Testing

For each secondary outcome family, we construct a generalized-least-squares-weighted standardized summary index and report sharpened q-values controlling the false discovery rate within family, both following Anderson (2008). The four primary outcomes are treated as a single family and corrected jointly.

3.6 Heterogeneity

We will estimate heterogeneity along four pre-specified dimensions, each interacted fully with the treatment indicators:

1. Below-median baseline foundational skills, measured as a summary index of the raw sub-domain scores of the EGRA and EGMA;
2. Child sex;
3. Baseline household food insecurity, measured by an above-median child food insecurity experience scale score;
4. Baseline school-level TaRL session delivery above versus below the median.

Gender disaggregation of all primary and secondary outcomes is required by J-PAL's Learning for All Initiative and will be reported as well.

3.7 Attrition

Because the interventions aim to boost attendance and outcomes are measured in a school-based survey, differential attrition is a threat by design. We will report attrition rates by cell, test their equality, and compare the baseline characteristics of attritors across cells. Regardless of the results of these tests, we will report our main treatment effects adjusted through inverse propensity score weighting for a child's propensity to be absent at the time of our survey, and will place bounds on our estimates in the Lee and Kling-Liebman-Katz traditions.

References

Anderson, Michael L, "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*, 2008, *103* (484), 1481–1495. 6, 7

Angrist, Noam and Rachael Meager, "Implementation matters: Generalising treatment effects in education," Technical Report WP_2023/001, What Works Hub for Global Education November 2023. Publication Title: What Works Hub for Global Education. 4

Chimbutane, Feliciano, Naureen Karachiwalla, Catalina Herrera-Almanza, Jessica Leight, and Carlos Lauchande, "The effect of teacher training and community literacy programming on teacher and student outcomes," *Journal of Development Economics*, January 2026, *178*, 103578. 6

de Barros, Andreas and Theresa Lubozha, "Targeting Foundational Skills at Scale: Skill Specificity and Transfer," March 2026. 5, 6