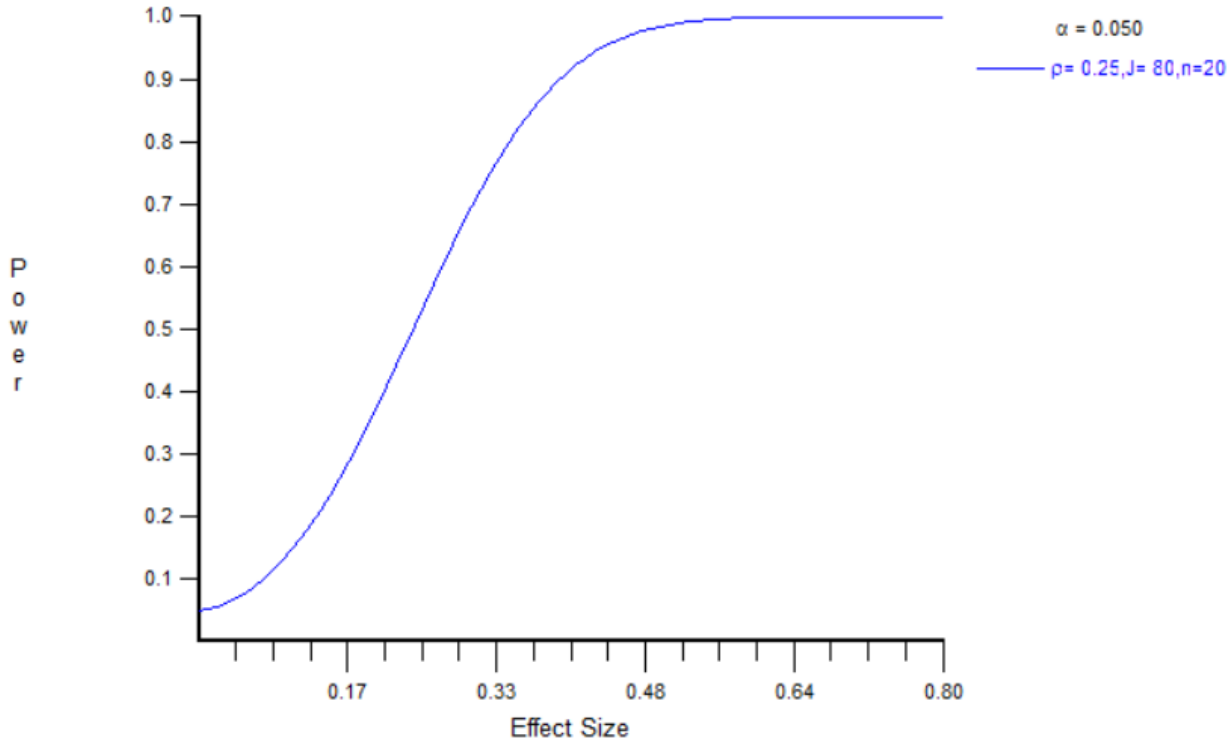
Power Calculations

Main Intervention

Our power calculations are based on an 8-school pilot RCT that we began in early 2024, toward the end of when students were in Grade 1. The ICC of overall reading test scores was 0.25. To be conservative, we zero predictive power from baseline covariates for endline test scores. We set a significance level (α) of 0.05, and assess the minimum detectable effect size at 90% power. We plan to sample 22 students per school and assume an attrition rate of 10% per year, so that we have 20 students/school at endline. We focus here on power to detect effects at the end of grade 1, so in mid-2025 for Phase 1 and mid-2026 for Phase 2. Since the intervention will run through grade 3 in Phase 2, the effects will be even larger by the end of that Phase, making our MDE estimates here a lower bound.

In Phase 1 of the study, when we have 80 total schools (40 per study arm), we will have an MDE of 0.39 SDs at 90% power:



While these effect sizes are large, they are realistic for two reasons. First, our intervention builds off of a family of related interventions with typical effect sizes that are comparable or much larger. For example, the NULP (Buhl-Wiggers et al. 2024) raises test scores by 1.4 SDs by the end of Grade 3, an average gain of 0.47 SD/year. After just one year of the program, it had boosted test

scores by 0.64 SD (Kerwin and Thornton 2021). Similarly, PRIMR increased reading scores after one year by 0.73 SD in Kiswahili and 0.93 in English (Piper et al. 2018).

Second, we have preliminary evidence that the TFLI intervention has large impacts comparable to these other programs. Our 8-school pilot RCT ran for just four months, from March to July 2024, and we found that the intervention increased test scores by 0.26 SDs in Grade 1 and 0.22 SDs in Grade 2, with the latter effect being significant at the 0.05 level. If we scale up the intervention's effects linearly over the whole school year, both of these impacts are larger than 0.5SDs, easily giving us enough statistical power to detect the intervention's impacts.

Enumerator Demand Effects

Our power calculations for the enumerator demand effects are based on assuming the following error structure where student test scores are affected not only by their school, but also the enumerator who collects their test score data. We calibrated these values using data from Rodriguesz-Segura and Schueler (2023) to correctly reflect the true shares of test score variance.

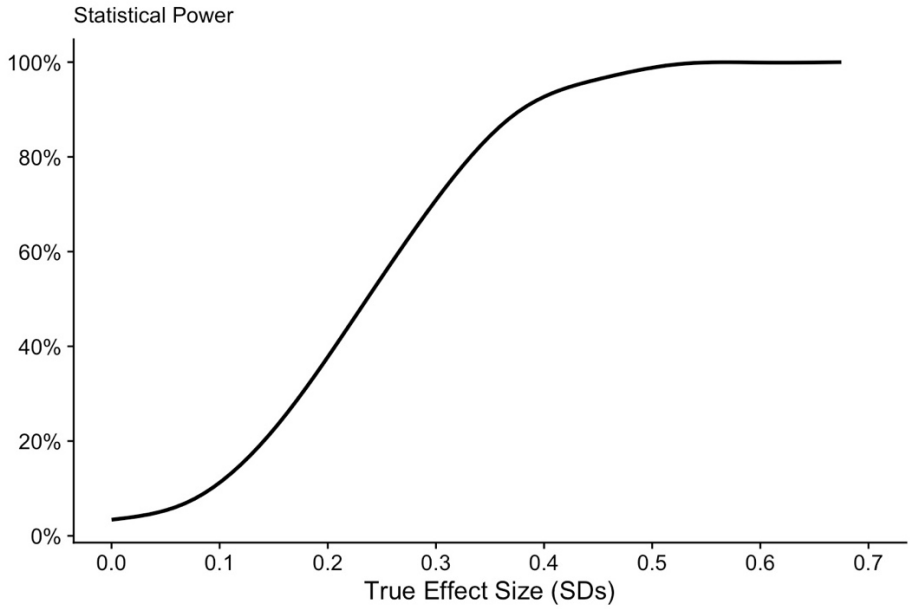
$$\epsilon = \sigma_{school} + \sigma_{enumerator} + \sigma_{student}$$

We plan to have 80 total schools (40 treatment and 40 control), and 24 enumerators (12 treatment trained teachers, and 12 non-teachers) in the study. We again assume we will have on average 20 students per school at endline. We set $\alpha = 0.05$, and calculate an MDE at 80 and 90% power.

We assume that exactly one treatment and one control enumerator go to each school with the pairing randomly varying across schools. We also assume zero predictive power for baseline covariates at endline. Both of these assumptions imply that our estimated MDEs are conservative estimates of the true MDEs, and so are a lower bound.

To find our statistical power for enumerator effects, we simulate the data under this DGP 1000 times and compute proportion of simulations where the estimated coefficient is significant. We use two-way clustering of the standard errors by the treatment and control enumerator at a school. The treatment enumerator is the trained teacher, and the control is the non-teacher assigned to each school. These are constant for each student in a school. The randomization procedure creates correlation at the level of each enumerator, but also at the level of each pair of enumerators, and this clustering scheme accounts for that.

For the main effects of the enumerators, we will have an MDE of 0.35 SDs at 80% power, and 0.4 SDs at 90% power.



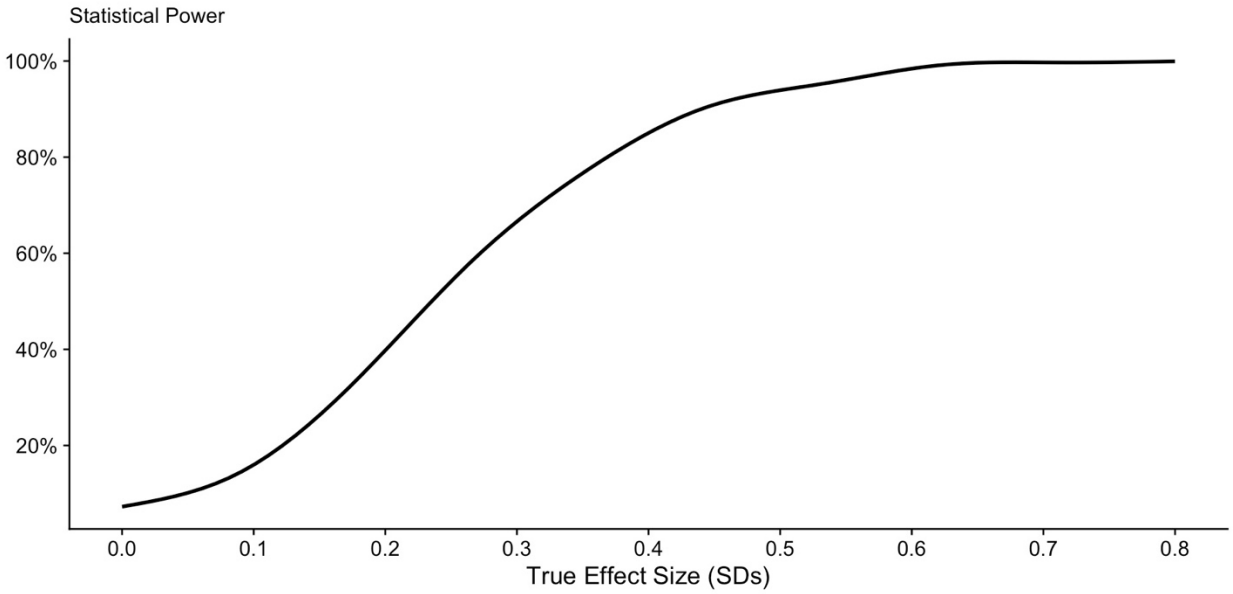
Power analysis assumes 12 treatment and 12 control, 80 schools, and 20 children per school.
: and one control enumerator per school. The MDE at 80% power is 0.325, and at 90% power it is 0.375.

We will also analyze the interaction between the main intervention and the enumerator demand effects to determine if enumerators differentially affect the treatment and control group. The coefficient of interest is β_3 in equation (1), below. The error structure we assume is the same as for enumerator demand effects, and values are calibrated in the same way.

$$y_{ise} = \alpha + \beta_1 MainTreatment_{is} + \beta_2 EnumeratorTreatment_{ise} + \beta_3 MainTreatment * EnumeratorTreatment_{ise} + \epsilon_{ise} \quad (1)$$

We plan to have 80 total schools (40 treatment and 40 control), and 24 enumerators (12 treatment trained teachers, and 12 non-teachers) in the study. We again assume we will have on average 20 students per school at endline. We set $\alpha = 0.05$, and calculate an MDE at 80 and 90% power.

For these power calculations we use the same simulations, but use two-way clustering of the standard errors by the treatment enumerator and stratification cell. The treatment enumerator is the teacher assigned to a given school, which is constant for each student at a given school. Our simulations show that this is the level of clustering at which there is a 5% rejection rate when there is a true effect of 0 for the interaction term. We have an MDE of 0.36 SDs at 80% power, and 0.43 at 90% power for the interaction coefficient.



Power analysis assumes 12 treatment and 12 control enumerators, 80 schools, and 20 children per school. Randomization is done with 1 treament and one control enumerator per school. The MDE at 80% power is 0.365, and at 90% power it is 0.435

There is very little power loss relative to the main effects of enumerator type, and higher power than the main effects of the intervention. Intuitively, this is possible because this analysis exploits two independently-randomized sources of variation. In our simulations, the true standard deviation of the interaction coefficient has a smaller standard deviation than that on the main effect of enumerator type.

While these power calculations are contingent on the assumptions we made in our simulations, they indicate that we will be powered to detect enumerator type-treatment interactions that are comparable in magnitude to the main effects of the intervention. Thus we should be able to rule out the entirety of the treatment effect being explained by enumerator type.