Integrating Educational Technology with Structured Pedagogy to Improve Learning Outcomes for Every Student

Analysis Plan

Erik Andersen, Simon Graffy, Jason T. Kerwin, and Monica Lambon-Quayefio

June 14, 2025

In this document, we outline the main analyses to be undertaken in the evaluation of the Tools for Foundational Learning Improvement (TFLI) intervention. TFLI is a structured pedagogy program which uses a 'science of reading' approach, combined with assessment-informed targeted instruction for early-grade reading. The current study will be implemented in low-fee private primary schools in Cape Coast in Ghana in an initial ("Phase 1") RCT with follow up assessments and surveys for grade 1 pupils over the next academic year. In addition to the prespecified analyses indicated in this document, we also expect to conduct further exploratory analyses beyond the ones described here. Those analyses will flow from one or more of the following: (i) further reflection on our part, (ii) developments in the related literature, or (iii) unexpected patterns in the data.

Sample

The planned analyses in this document will be implemented using students who enter grade one (BS1 in the Ghanaian education system) in the 2024-2025 academic year. This is Student Cohort 1. Phase 1 will include 80 schools (40 treatment, 40 control).

We plan to collect data on Student Cohort 1 at the end of the 2024-25 school year. If the budget allows, we will also collect follow-up data on Student Cohort 1 in the following school years. We will conduct our own exams for students at each follow-up wave; these will include the Early Grade Reading Assessment (EGRA) as well as other tests.

This analysis plan only describes the analysis from the end of 2024-25. There is no baseline data; instead, we collected student lists for every school at the beginning of the 2024-25 school year.

Our primary outcomes for Student Cohort 1 will use data from the end of Grade 1. All other analyses will be considered secondary outcomes.

We will run a separate analysis studying whether the type of enumerator who collects the data affects measured test scores. Specifically, we have two groups of enumerators: 1) people with teaching certifications (which we call "teachers", although they are not currently working as teachers) and 2) outside contractors who have no other connection with the education sector ("non-teachers"). There will be 12 teachers and 11 non-teachers.[1] We will randomly assign an equal number of teachers and non-teachers to each school and randomize at the student level which

[1] We intended to have 12 of each but one non-teacher resigned right after the enumerator training ended.

enumerator collects each student's data. This analysis will use the same data as our main analysis of treatment effects.

Obtaining impact estimates

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

$$Y_{ij} = \beta_0 + \beta_1 TFLI_j + Z'_j\tau + X'_i\gamma + \epsilon_{ij} \quad (1)$$

where *i* indexes students, which are nested within their original schools indexed by *j*.$TFLI_i$ is the indicator for a school being randomly assigned to receive the TFLI program. Z_{ij} 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, indicators for each value of age in years (as of the beginning of the academic year)), and the interactions between the two.[2] For students with missing values of any baseline variable, we will replace the missing values with zero and include a separate indicator variable for the original value being zero, along with any appropriate interaction terms.

For the enumerator type effects section of the intervention, we will still use equation (1), but $TFLI_j$ will be replaced with $Teacher_i$ which is an indicator for a student randomly being assigned to be assessed by a teacher. This is given in equation (2).

$$Y_{ij} = \beta_0 + \beta_1 Teacher_i + Z'_j\tau + X'_i\gamma + \epsilon_{ij} \quad (2)$$

We will also estimate treatment effect heterogeneity by teachers versus non-teachers. These will be estimated using equation (3) which we will estimate via ordinary least squares.

$$Y_{ij} = \beta_0 + \beta_1 TFLI_j + \beta_2 Teacher_i + \beta_3 TFLI_j * Teacher_i + Z'_j\tau + X'_i\gamma + \epsilon_{ij} \quad (3)$$

Inference

We will conduct inference on our main 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.

[2] We will winsorize age at the 5th and 95th percentiles of the distribution.

We will use multi-way clustered standard errors for inference for our enumerator type effect estimates. We will cluster standard errors at the level of each enumerator who attends a school; that is, if there is one teacher and one non-teacher at a school, we would cluster by each enumerator separately.

For the interaction between the two treatments, given by β_3 i in (3), we will cluster standard errors at the level of the stratification cell and also the teacher enumerator.

These inference schemes were validated via simulation.

Null hypotheses

We plan to consider one null hypothesis for each outcome we study except treatment heterogeneity:

$$H_0: \beta_1 = 0$$

For our estimate of treatment heterogeneity by SISO versus contractor we will consider the following null hypothesis:

$$H_0: \beta_3 = 0$$

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.

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. Since we plan to have just a single confirmatory hypothesis test, this approach will yield the original *p*-value, and thus we will not have to actually do the adjustment. However, if we analyze multiple confirmatory hypotheses in the future, we will use this method.

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

Main (Confirmatory) Outcomes

Following common practice, we divide our planned analyses into confirmatory and exploratory analyses. We have just one confirmatory analysis:

1. English EGRA score (in SDs)
    - Score is the weighted average of the subtest scores, where the weights are 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 the relevant cohort. We will standardize each subtest score by the control-group mean and SD before running PCA.
        - The specific subtests are:
            - Listening comprehension
                - Scored as the number of correct answers marked correct by the enumerator out of 3.
            - Letter Names
                - Scored as the number out of 100 letters marked as correctly read by the enumerator divided by the number of seconds it took to finish the letter grid.
                - If the student gets none of the first 10 letters correct, the test will end early, and they will get a 0 out of 100.
                - If the student does not finish reading the letters at the end of the allotted 60 seconds, the test will end, and they will still be scored out of 100.
            - Letter sound identification
                - Scored as the number out of 100 letter sounds marked as correctly read by the enumerator divided by the number of seconds it took to finish the letter grid.
                - If the student gets none of the first 10 letters correct, the test will end early, and they will get a 0 out of 100.If the student does not finish reading the letters at the end of the allotted 60 seconds, the test will end, and they will still be scored out of 100.
                - Students have 60 seconds to read the letter grid. If they don't finish in that time, their time will be marked as 60 seconds.
            - Initial sound identification
                - Scored as the number out of 10 letter sounds marked as correctly identified by the enumerator.
            - Familiar word reading
                - Scored as the number out of 50 words marked as correct by the enumerator divided by the number of seconds it took to read the word grid.
                - If the student gets none of the first 5 words correct, the test will end early, and they will get a 0 out of 50.

        - If the student does not finish reading the letters at the end of the allotted 60 seconds, the test will end, and they will still be scored out of 50.
    - Non-word reading
        - Scored as the number out of 50 words marked correct by the enumerator divided by the number of seconds it took to read the word grid.
        - If the student gets none of the first 5 words correct, the test will end early, and they will get a 0 out of 50.
        - If the student does not finish reading the letters at the end of the allotted 60 seconds, the test will end, and they will still be scored out of 50.
    - Oral reading passage
        - Scored as the number of words read correctly by the student divided by the number of seconds it took to finish the passage.
        - If the student gets none of the first few words correct (up to a word indicated in a box), the test will end early, and they will get a score of.
        - Students have 60 seconds to read the passage. If they don't finish in that time, their time will be marked as 60 seconds.
    - Reading comprehension
        - Scored as the number of correct answers out of 5.

  We will standardize the overall PCA index by the control-group mean and SD, so that it has units of SDs of the control-group distribution

Our only confirmatory hypothesis test will be a test of the null that $\beta_1 = 0$ in equation 1.

<u>Secondary (exploratory) analyses</u>

Our exploratory analyses will include the following:

First, we will estimate equations (2) and (3) to study enumerator type effects. We will also run versions of these two regressions where $Teacher_i$ is replaced with $SISO_i$, which is an indicator for the enumerator worked as a School Improvement Support Officer, which is a subset of the teacher enumerators. This is not explicitly targeted for randomization in our experiment, but due to the random assignment and pairing of enumerators it will be random as well.

Second, we will consider the distributional effects of the intervention with three methods. We will run quantile regressions using equation (1) at each of the percentiles of the distribution of scores.

Next, we will run a distributional regression using the Chernozhukov, Fernandez-Val, and Melly (2013) method. Finally, we will run a K–S test to determine if the treated and control distribution of outcomes are different.

Third, we will also estimate the effects of TFLI on individual subtests of the EGRA evaluation listed in the main outcomes section by running equation (1) with each subtest individually as the outcome variable. We will also look at distributional effects on each subtest.

Fourth, we will validate the data collected by enumerators by recording the audio of the EGRA exams, and re-score the exams using an AI tool. We will compare both the underlying scores and the estimated treatment effects across methods.

Fifth, we will look at treatment effect heterogeneity by student gender, teacher gender, and teacher-student gender match.

Finally, we will also express our treatment effects as Equivalent Years of Schooling (EYS). Since we have no baseline test scores, we will explore other ways of measuring typical progress on the EGRA each year.

Besides EGRA scores, we will also study the following outcomes from data collected during school visits.

First, we will study other questions from the student survey not named in the main outcomes section above. Specifically, we plan to analyze how the intervention affects students' perceived class ranks, career, and academic aspirations. We will also test if the intervention affects whether students practice reading and writing outside of school.

Second, using data from BS1, we will analyze the difference of quality of teaching instruction between treatment and control schools by recording two lessons and analyzing them in the following way. We will construct a quality score where the score is the weighted average of the different metrics, and the weights are the first principle of the control group data across all classroom observations for BS1. We will standardize each of the following metrics by the control school mean and SD before running PCA.

1. Oral language
    a. Teacher gets learners using new vocabulary in conversation (1-3 scale, divided by 3)
    b. Teacher asks questions before, during, and after read aloud (1-3 scale, divided by 3)
    c. Teacher gets learners practicing the new language structure (1-3 scale, divided by 3)
2. Phonics

    a. Teacher engages learners in phonological skill drills (1-3 scale, divided by 3)
    b. Teacher gets learners to say the sound and write the letters (1-3 scale, divided by 3)
    c. Teacher says the right sound and gets learners to write and say it (1-3 scale, divided by 3)
3. Reading
    a. Teacher gets learners blending sounds to make words (1-3 scale, divided by 3)
    b. Teacher displays sight words and gets learners to say them (1-3 scale, divided by 3)
    c. Teacher gets learners to read aloud during structured reading (1-3 scale, divided by 3)
4. Writing
    a. Teacher clearly demonstrates the writing skill using examples (1-3 scale, divided by 3)
    b. Teacher uses a whole class activity to get learners thinking (1-3 scale, divided by 3)
    c. Teacher sets learners a task and gives them time to practice (1-3 scale, divided by 3)
5. The teacher progresses at an appropriate pace (1/0)
6. The teacher has good presence and speaks clearly (1/0)
7. The teacher proactively manages behavior (1/0)
8. The teacher moves around to check understanding regularly (1/0)
9. The teacher supports learners who need additional help (1/0)
10. The learners appear familiar with the routines and know what to do (1/0)
11. The learners engage with the workbooks as expected (1/0)
12. The learners talk to each other when directed by the teacher (1/0)
13. The learners get involved in activities and ask appropriate questions (1/0)
14. The learners stay on task during independent practice (1/0)

We will also analyze compliance with the program using BS1 data. We will construct a compliance score that is an evenly weighted average of the following components:

- Number of assessments run divided by number they were supposed to run*
- (Lesson number the teacher is on)/ (lesson number the teacher should be on for this day)*
- Number of peer coaching meetings run divided by the number they were supposed to run*
- Did class use workbooks (1/0)
- Did class use teacher guides (1/0)
- What fraction of students did the teacher assess one-on-one last term? (0-3 scale, divided by 3)

- Did the teacher send report cards home last term? (1/0)

  * question will be scored zero for all control-group teachers

Some of these variables have only one-sided non-compliance: the control-group will be mechanically scored zero for anything where compliance is only possible for the control group, as indicated by an * above. We will construct two versions of the compliance score. In the first we will use the true values of each variable, other than the ones marked with an *. In the second we will assign all control-group teachers an overall compliance score of zero.

We will analyze this compliance score in two ways. First, we will put it on the left-hand side in equation (1). Second, we will estimate

$$Y_{ij} = \beta_0 + \beta_1 Compliance_i + Z'_j\tau + X'_i\gamma + \epsilon_{ij} \quad (4)$$

using $TFLI_i$ as an instrument for $Compliance_i$. This will let us estimate how well the program would have worked with full compliance, under the assumption that the effects of the randomized intervention operate only through compliance with the program.

In addition, we will run the sequential g-estimator of Acharya et al. (2016) to see how much of the treatment effect is mediated by compliance measures and teaching quality measures. We will follow the implementation from Kerwin and Thornton (2021), using all the individual components of the teaching quality score and compliance score at the same time, as they are all both potential mediators and potential omitted intermediate variables for one another.

To further test program implementation, we will non-causally test how well 1:1 oral reading fluency (ORF) test scores run in treatment schools as part of the program correlate with ORF scores collected by enumerators at the endline data collection. We will do this by estimating equation (6), where $OneOnOneORFi$ is the final ORF assessment run by the teachers as part of the TFLI program before the endline survey.

$$Endline\ ORF_i = \beta_0 + \beta_1 OneOnOneORF_i + X'_i\gamma + \epsilon_i \quad (6)$$

We will likely also conduct other exploratory analyses of the data, motivated by our initial findings.

Attrition

We will use Lee (2009) bounds to deal with potential differential attrition between the treated and control groups. Specifically, we will estimate the trimming proportion by taking the difference in the proportion of non-missing outcomes between the treatment and control group, $\hat{p}$. We will use this proportion to estimate the $\hat{p}$, and $(1-\hat{p})$ quantiles of the distribution of outcomes in the treated group. Using these, we will trim the top and bottom of the data for the treated group outcomes and calculate upper and lower bounds for β_1. We will implement this in Stata using the leebounds command.

References

Acharya, A., Blackwell, M., & Sen, M. (2016). Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects. *The American Political Science Review*, *110*(3), 512.

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.

Chernozhukov, V., Fernández-Val, I. and Melly, B. (2013), Inference on Counterfactual Distributions. Econometrica, 81: 2205-2268.

Kerwin, J. T., & Thornton, R. L. (2021). Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures. *The Review of Economics and Statistics*, *103*(2), 251–264. https://doi.org/10.1162/rest_a_00911

Lee, David S. Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects, *The Review of Economic Studies*, Volume 76, Issue 3, July 2009, Pages 1071–1102.