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

Phase 1, Wave 2 Analysis Plan

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

June 19, 2026

This document extends the original Analysis Plan document dated June 15, 2025 to account for our planned analyses of the second wave of data collection that will take place at the end of the 2025-2026 school year. Most details remain identical to that earlier document. We note only the central points in this document, and any key differences in this new document.

Sample

The sample is the same set of students as specified in the earlier document: Student Cohort 1, from Phase 1 of the study. The data we will analyze is from Wave 2 of data collection on this cohort of students, which will take place when Student Cohort 1 has been in school for two years.

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:

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

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] (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 study), and the interactions between the two.[2] For students with missing values of any baseline variable, we will replace the missing values

[1] If too many entire schools attrit from the study, observations from their whole stratification cell will be dropped from the regression because there will be no variation in the treatment indicator within the cell. In that case, we will reassign schools from the stratification cell that would be dropped to the previous stratification cell on the list, which will have similar average school sizes (because the cells were created by sorting the schools by size and then breaking them up into groups). So if stratification cell 5 loses all its treatment schools, we will merge it with cell 4. If this happens for stratification cell 1, we will merge it with cell 2 rather than the highest cell value to maintain similarity. We will also report our results without applying this rule as a robustness check.

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

with zero and include a separate indicator variable for the original value being zero, along with any appropriate interaction terms.

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.

This inference scheme was validated via simulation.

We will use the seed 3935890 for all instances of randomization inference, bootstrap inference, or other processes which involve which involve pseudorandom number generators. This number came from Random.org, which generates true random numbers using atmospheric noise.

Null hypotheses

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

$$H_0: \beta_1 = 0$$

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>

In addition to the prespecified analyses indicated in this document, we also expect to conduct a variety of exploratory analyses. These will include the secondary analyses described in the original analysis plan document. Beyond these our exploratory analyses will include the following:

First, we will estimate treatment effect heterogeneity using a causal forest (Athey, Wager; 2019) using the same set of covariates as in Andersen et al. (2026).

Second, we will estimate the model of skill formation in Andersen et al. (2026) using the method of simulated moments.

Third, in addition to the EGRA tests we will estimate the effect of the treatment on questions from previous national standardized tests for second grade students, contingent on being able to access those test items and add them to our assessments.

Fourth, we will assess the difference in the quality of teaching instruction between treatment and control schools by recording literacy lessons and coding them with the World Bank's Teach Primary classroom observation tool.

Fifth, since several schools have closed, we will analyse the treatment effect on treated schools in the following way. We will estimate:

$$Y_{ij} = \beta_0 + \beta_1 Years\ of\ \ treatment_j + Z'_j\tau + X'_i\gamma + \epsilon_{ij} \qquad (2)$$

Where $Years\ of\ treatment_j$ is the number of years each school has received the TFLI program. For schools that have closed at the start of the 2025-26 academic year, this will be equal to 1. For schools that closed at the start of the 2024-25 academic year, this will be equal to 0. For all other treatment schools, this will be equal to 2. Due to an administrative error, two schools had their treatment assignment swapped during implementation: a school assigned to control received the TFLI program and vice versa. For these schools, we will set the years of treatment variable for the control school that received the treatment to 2 and we will set it to 0 for the treatment school that did not receive TFLI. Because years of treatment is an endogenous variable, we will use $TFLI_j$ as an instrument for $Years\ of\ treatment_j$. This will let us estimate how well the program would have worked if all schools had received the full two-year dose of TFLI under the assumption that the effects of the randomized intervention operate only through length of exposure to the program.

We have improved our method for determining the lesson the teacher is on for the compliance index, so we will update the definition of that variable. We will determine the lesson the teacher is on as an equally weighted average of the lesson number the teacher is on in their teacher guide and the lesson number three randomly selected students are on in their workbook. We handle this one-sided non-compliance in the same way as described in the previous version of this document.

We may also conduct additional analyses flowing 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.

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 manually in Stata to allow for cluster robust inference.

References

Andersen, Erik T. J., Simon Graffy, Monica Lambon-Quayefio, and Jason Kerwin. 2026. "How to Build a Reader: Evidence from a Scalable Literacy Intervention in Ghana." Working Paper.

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.

Athey, Susan and Stefan Wager. "Estimating Treatment Effects with Causal Forests: An Application." Observational Studies, vol. 5 no. 2, 2019, p. 37-51. Project MUSE, https://dx.doi.org/10.1353/obs.2019.0001.

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

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.

McFadden, Daniel. 1989. "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration." Econometrica, 57(5): 995–1026.