Interactive Phone Calls to Encourage Home Reading

Pre-Analysis Plan

Anja Sautmann[*]

June 11, 2021

1 Overview

This document outlines an adaptive experiment with exploration sampling (Kasy and Sautmann (2021)) to select one out of six treatment arms of biweekly phone calls to parents that deliver reading exercises for first graders. The experiment is carried out in two waves during the last trimester (term 3) of the first grade in a sample of 108 private primary schools in Kenya. In what follows, we use "implementer" for the organization running these schools. The method uses Bayesian econometrics to estimate average treatment effects, and exploration sampling to assign parents to treatment arms in waves other than the first wave.

The objective of carrying out this trial was to identify the treatment arm that in expectation leads to the greatest improvements in reading fluency, through engaging parents in reading regularly with their children at home. Exploration sampling is a sampling method designed for the objective of selecting the best policy option for implementation out of a larger set of possible options. It is particularly suited to informing policy choice under constraints on the sample size, cost, or time to run the experiment.

This experiment was subject to several such constraints. The implementer wanted to make a decision about the type of interactive voice response (IVR) call with early literacy exercises they should make, if at all, after one school term of testing, with two rounds of data collection at the midterm and endterm exams. We designed and implemented the entire experiment from January 28 to May 11, 2021, with a technical pilot mid April, focus group interviews with pilot participants end of April, and internal approval by the implementer for the full experiment on April 30. Oral reading fluency (ORF) measurements, in the form of student-level correct words per minute (cwpm) assessments, had been introduced to the schools only in the term before and carried out twice at midterm and endterm of term 2 (see also below).

[*] Development Economics Research Group, World Bank, email: asautmann@worldbank.org. The findings, interpretations, and conclusions expressed in this document are entirely those of the author. They do not necessarily represent the views of the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

This pre-analysis plan was written after pre-registering the experiment on May 11 (roll-out of wave 1), but before roll-out of wave 2 (June 12). The pre-registration names reading fluency (measured as number of correct words per minute) as the outcome variable, which reflects the initial plan for the experiment. It was an intentional choice to post the pre-analysis plan later, after fully testing and developing the empirical model used to decide on the exploration sampling shares. However, additional adjustments to the study plan were made after observing engagement data from the first study wave, and these are described below. This corresponds with viewing pre-specification as a record of intentions rather than a final, binding plan for the experiment (Banerjee et al. (2020)). This experiment provides a test case for using exploration sampling to inform policy decisions under real-world time constraints and uncertainties.

2 Treatments

All treatment arms consist of twice weekly calls to the phone number on record for a child's parents using interactive voice response (IVR) technology, which deliver specific reading exercises over the phone. The more advanced exercises are based on reading passages from the children's homework book, and the more emergent exercises were based on reading letter combinations or words that the parent is asked to note down during the call. These exercises were developed by the implementer.

The treatment arms vary the content of the exercises as well as the delivery format. The design of the interventions was also chosen by the implementer in collaboration with the research team, comparing treatment variations that were genuine "contenders" for having the greatest impact on reading fluency. The objective of the experiment is to select one of the treatment arms for implementation for all future first graders. The implementer had never used automated voice calls before to contact parents.

Varying exercise content. The baseline data from term 2 showed high variance in fluency levels, in line with other comparable data in developing contexts (for instance, see Muralidharan et al. (2019)). Prior evidence has suggested that there can be benefits on average to leveling remedial programs (see e.g. Banerjee et al. (2007), Banerjee et al. (2017)). Recent work has suggested that proper customization of "edtech" interventions could benefit the lowest achieving students the most (de Barros and Ganimian (de Barros and Ganimian)).

However, the reading fluency data are fairly noisy, as indicated by the comparison of midterm and endterm scores from term 2, seen in figure 1. This might make leveling ineffective or even counterproductive. Other treatment variants therefore give all children the same fixed exercise sequence matched with the progression of the term, or allow parents to choose between different exercises.

From a variety of possible designs, the implementer selected the three content intervention variants A, B,

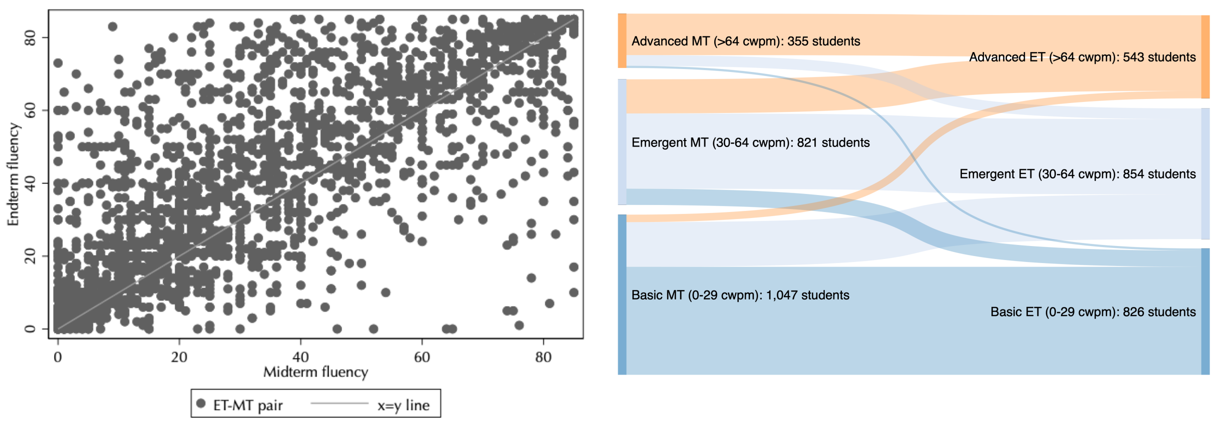
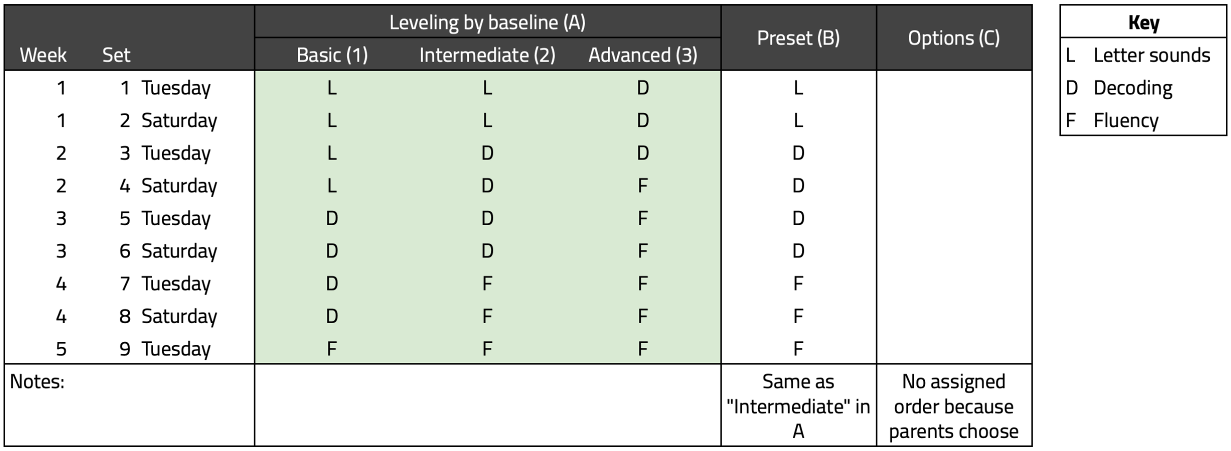


Figure 1: Term 2 midterm and endterm oral reading fluency scores, in units of correct words per minute. The left panel maps individual student scores and shows that they are only noisily correlated. The right panel shows that there is some movement from higher leveling categories to lower ones, as well as small but significant numbers of students "skipping" from basic to advanced level.



		Leveling by baseline (A)			Preset (B)	Options (C)
Week	Set	Basic (1)	Intermediate (2)	Advanced (3)		
1	1 Tuesday	L	L	D	L	
1	2 Saturday	L	L	D	L	
2	3 Tuesday	L	D	D	D	
2	4 Saturday	L	D	F	D	
3	5 Tuesday	D	D	F	D	
3	6 Saturday	D	D	F	D	
4	7 Tuesday	D	F	F	F	
4	8 Saturday	D	F	F	F	
5	9 Tuesday	F	F	F	F	
Notes:					Same as "Intermediate" in A	No assigned order because parents choose

Key	
L	Letter sounds
D	Decoding
F	Fluency

Figure 2: Exercise content variations.

and C as shown in figure 2:

A. Leveling by baseline: assign students to a "basic", "intermediate", or "advanced" arm;

B. Preset: assign all students to an "intermediate" exercise sequence;

C. Options: allow parents to select the exercise from a menu.

The leveling by baseline uses observed fluency scores from the end of term 2 and assign students with fluency scores of 0-29 into the"basic" arm, 30-64 into the "intermediate" arm, and 65+ into the "advanced" arm. These cutoffs were used previously in a similar context (the external TUSOME evaluation in Kenya, see Piper et al. (2018)). For students with missing scores, we assign them their class median. For classes with missing scores, we assign the intermediate level (which also happens to be the sample median).

Varying delivery format. We also test two delivery mechanisms that use the IVR functionality slightly differently.

- T1 – Engaging parents: The IVR explains to the parent how to do the exercises, then asks them to carry them out with their child after the call;
- T2 – IVR as a tutor: The IVR asks parents to put their phone on speaker phone and then goes through the exercises with the parent and child on the call.

Treatment arms and call design. We cross-combine the 3x2 interventions to create 6 treatment arms. A wave contains 9 sets of calls, and each call contains 4 different exercises. In other words, if a parent were to engage in all calls during a week, they would be doing 8 individual exercises per week. Calls are designed by recording a set of modular text snippets and jingles that are sequenced in response to listener input. The recordings were created by a female Kenyan voice artist and edited by the voice call provider, Uliza. The IVR system makes multiple call attempts and also allows the parent to "flash" Uliza's number, meaning that they can call the number at a convenient time, and the system hangs up and immediately calls back. This is a common method to avoid charges to one party. All these interactions with Uliza are recorded. Before starting the biweekly calls, we conducted a phone based enrollment and consent procedure and also allowed parents to change the enrolled number.

3 Enrollment and randomization

Sample. The phone on record for the parent is used to select children and identify siblings/members of the same family. All phone numbers are de-identified by the implementer before sharing with the researchers.

We split the sample of students enrolled in term 3 into two equal waves. In addition to the 6 treatment arms, we hold back 1/7 of the sample in each wave as a control group to be able to estimate absolute effects of treatment on reading performance.

We are interested in the treatment arm with the greatest effect on reading fluency. The implementing partner measures oral reading fluency using "correct words per minute" assessments (cwpm, see also below). ORF is also used to level the exercises in treatment variant A. Term 2 data shows that average ORF varies widely by school. The (student-level) randomization is therefore stratified at the school level. We dropped 2 schools from the grade-1 sample that had fewer than 5 students, and 2 schools in which average midterm (endterm) scores were 83 and 81 (81 and 75), respectively, suggesting that ORF was mis-measured by not correctly timing the child (and a risk that the same error would happen in the future). We also randomly selected one student ID in the few cases where several student IDs were associated with the same parental phone number (likely siblings), leaving us with 108 schools with 3163 unique student-phone number combinations.

We assigned student IDs with equal probability to wave 1 and wave 2, stratified by school ID, and selected 1581 IDs for wave 1. We then conducted an enrollment call to wave-1 phone numbers to obtain consent for participation in the IVR ”Reading at Home” project, followed by a text message confirming enrollment and explaining procedures for opt out and for switching to a different phone number.

In this process, for wave 1, 33 parents opted out by text message or by selecting opt-out during the enrollment call, and 39 phone numbers were identified as unreachable or invalid.

796 student IDs never responded to the call and therefore did not explicitly opt out. It is possible that their phone numbers are invalid. For the random treatment assignment, we therefore stratified the final sample by school as well as by whether the enrollment call was successfully completed. We then allocated approximately 1/7th of each stratum to the six treatment arms and the control group. The final sample allocation in each treatment arm in wave 1 is:

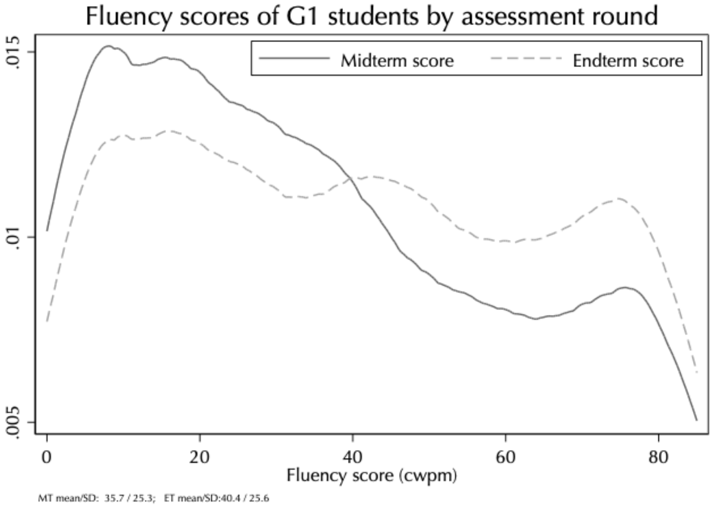
- T1A: 211 (14.12%)
- T1B: 220 (14.73%)
- T1C: 207 (13.86%)
- T2A: 214 (14.32%)
- T2B: 205 (13.72%)
- T2C: 226 (15.13%)
- Control: 211 (14.12%)

For wave 2, we will carry out the same procedure: conduct the enrollment process, remove opt-outs and invalid numbers, and stratify the remaining wave 2 sample into the treatment arms according to the exploration sampling shares, allocating 1/7th to the control group. How the exploration sampling shares are determined is described in section 6.

4 Outcome Measurement

We will estimate treatment arm averages for two outcomes; oral reading fluency, and parental engagement.

Oral Reading fluency. ORF is measured as the number of words a child can read correctly in one minute of time (see Rodriguez-Segura et al. (2021) for the use of this measure to assess reading and literacy). The implementer collects an ORF score by asking the child to read from a list of words or passage for one minute, while the teacher counts the number of words read correctly. The measure can in principle range from zero to over 200, but for first graders it is typically not above 120. The observed reading fluency level may be truncated above based on the length of the provided word list (for example, in our baseline data



MT mean/SD: 35.7 / 25.3; ET mean/SD:40.4 / 25.6

Figure 3: Kernel plot of midterm and endterm fluency (cwpm) scores of grade 1 students, term 2 of 2020/21.

it is truncated at 85). In the schools in our sample, ORF is collected twice a term, during midterm and endterm exams. Figure 3 shows the midterm and endterm distributions of scores in term 2 in the schools in our sample. Correct words per minute are measured by teachers in the classroom during the examination periods, and submitted to the school's grade record system. We receive those records in de-identified form after they are submitted. The implementer chooses a standardized, and grade-appropriate word list and trains teachers to administer the cwpm measurement.

Engagement. We measure engagement using administrative records of the IVR provider. Uliza's records show every contact with the parent's registered phone number, along with the length of each call in seconds.

We define a call as successful if the parent started the first exercise. We define a parent as having engaged in an exercise set if they had at least one successful call in that set, i.e. they started the first exercise of the exercise set at least once. Getting to the first exercise requires tapping phone keys to confirm. Since there are 9 exercise sets per wave, a given phone number may have an engagement level between 0 and 9, indicating the number of times the parent engaged in an exercise set.

5 Estimation

We begin by defining the estimation approach for oral reading fluency and engagement.

In what follows, we denote ORF by y_{it}^{ks}, where t is the wave, i stands for the student, s is the school, and $k \in \{0, 1, \ldots, 6\}$ denotes the treatment arm. "Treatment" $k = 0$ indicates the control group. We denote engagement by Z_i^{ks}, where we suppress that the student is observed in a specific wave t, under the assumption that engagement is not subject to a common time trend between the two waves.

For both outcomes, we use a hierarchical Bayesian linear model to estimate average treatment effects, and allow average outcomes to vary across strata (schools).

The model for parental engagement. Define $\theta^{sk} \in [0, 1]$ as the average probability of engagement in a given exercise set in school s in treatment arm $k = \{1, \ldots, 6\}$. Engagement is by definition 0 in the control group, so we restrict the sample to enrolled phone numbers in the 6 treatment arms.

We model the average engagement probability with a hierarchical logistic regression model. Therefore, for a student's parental engagement, conditional on θ^{sk}, we have

$$Z_i^{sk} \mid \theta^{sk} \sim \mathit{Binomial}(9, \theta^{sk}) \ , \tag{1}$$

and we model the success probability of students in s, k as

$$\theta^{sk} = \text{logit}^{-1}(\beta^E x^k + \kappa^E \eta_s^E) \ . \tag{2}$$

The vector x^k is a unit vector indicating the treatment arm k, β^E is a 1×6 vector of average treatment effects, and $\kappa^E \eta_s^E$ is the school-level realization of the random effect. Note that we do not use any individual-level covariates since we have no information on parental background.

We use a non-informative improper prior on $\{\beta_k^E\}_{k=1}^6$, a Half-Normal prior distribution for κ^E (the standard Normal on $[0, +\infty)$), and a Standard Normal prior distribution for the school random effects.[1]

$$\begin{aligned} p(\beta_k^E) &\propto 1 \quad \forall k = 1, \ldots, 6 \ , && (3) \\ \kappa^E &\sim \text{Half-Normal}(0, 1) \ , && (4) \\ \eta_s^E &\sim N(0, 1) \ . && (5) \end{aligned}$$

The hyperparameters describe our beliefs about the mean outcome within each school and treatment arm and the variance across schools. This model was chosen using engagement data from the first three weeks of the experiment. Model choice was based on a variety of possible parameterizations of the school and treatment effects, judging model performance using posterior predictive checks for the specific schools observed in our sample.

The model for oral reading fluency. We will define the exact model for ORF measured in "correct words per minute" after receiving information from the mid-term exams. We plan to use school-specific

[1] We use this parameterization of the random effects to avoid what is known as "Neal's funnel" when sampling from the joint distribution of the treatment effects and random effect variance (Neal (2003)).

effects as well as to allow for a common time trend from wave 1 to wave 2, to account for normal average reading progression. We are currently considering a censored linear normal model or a beta generalized linear model at the student level to account for the truncation of z at 0 and 85. We will quantify out-of-sample prediction accuracy to select the optimal model (see below, model comparison metrics).

Model estimation and convergence checks. We sample from the posterior of the relevant variables using Hamiltonian Monte Carlo (HMC). We implement HMC in Stan and rely in particular on three convergence checks to assess the performance of the sampler: split-$\hat{R}$, tree depth, and divergent transitions.

Posterior predictive checks. We use posterior predictive checks to explore systematic differences between our model and the data. For this purpose we replicate outcomes based on the parameters estimated off of the existing data and then compare the simulated and real data. For example, for the parent engagement model, we first draw the hyperparameters β^E, κ^E, and η_s^E from their respective posteriors. This provides a draw of $\theta^{sk} = \text{logit}^{-1}(\beta^E x^k + \kappa^E \eta_s^E)$. Then we draw a set of replicated student level observations from

$$Z_{j,\text{rep}}^{sk} \sim Binomial(9, \theta^{sk}) \ ,$$

and compare the means of the original and replicated data. For this comparison, we calculate a Bayesian p-value, based on the mean,

$$p_B = Pr\left(\frac{1}{N}\sum z_{j,\text{rep}}^{sk} \geq \frac{1}{N}\sum z_i^{sk}\right)$$

from the proportion of simulations in the Markov Chains for which $\frac{1}{N}\sum z_{j,\text{rep}}^{sk} \geq \frac{1}{N}\sum z_i^{sk}$. We expect to see values of p_B greater than 0.05 and smaller than 0.95 if our model successfully replicates the data.

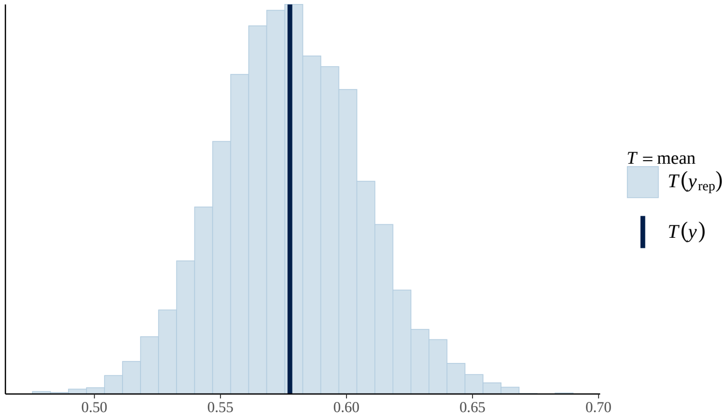


Figure 4: Distribution of the means of replicated outcomes (histogram) and mean of observed outcome (vertical line).

Using wave 1 data, the model proposed in Eq. 1 and 2 generates $p_B = 0.5$, which suggests that our model is successfully replicating the mean of the observed data. Figure 4 displays the distribution of $\frac{1}{N}\sum z^{sk}_{j,\text{rep}}$ (recall that each data point is the mean of the replicated outcome in one step of the simulation). The vertical line at 0.58 denotes the mean of the observed outcome.

Sensitivity analysis We use sensitivity analysis to measure the consequences of our prior selections on the posterior distributions of the parameters. We often use non-informative priors for location parameters (i.e. $\{\beta^E, \beta^C, \tau, \mu\}$). We test the impact on the posterior of selecting (i) a more constrained, weakly-informative normal prior and (ii) a weakly-informative t-student prior. We use a Half-Normal(0,1) prior for parameters describing dispersion (i.e. $\{\sigma, \kappa^E, \kappa^C, \nu\}$). Here we test the impact of using (i) a weakly informative χ^2 distribution and (ii) a weakly informative Half-t-student distribution. Importantly, we also explore the sensitivity of the exploration sampling shares under the different prior selections (see section 6).

Model comparison metrics During the modelling process, we may consider different model specifications in an effort to improve performance. We compare models using two metrics that estimate a point-wise out-of-sample prediction accuracy (Vehtari et al. (2017)): (i) Watanabe-Akaike or widely available information criterion (WAIC), and (ii) Leave-one-out Cross-Validation (LOO-CV).

6 Adaptive Sampling in Wave 2

We adapt the sample size in each treatment arm in wave 2 based on the Exploration Sampling algorithm described in Kasy and Sautmann (2021) in order to be able to pick the most successful arm.

We originally intended to target average oral reading fluency as the measure of success. However, we decided ultimately to use engagement; the considerations that entered this decision are outlined below.

Measuring engagement vs. ORF. During the pilot and wave 1, we learned that an issue with school records is that teachers sometimes fail to submit exam scores for the class to the recording system. As a result, the scores of an entire class room may be missing from the system for some time. Absenteeism causes additional missing records for individual students. We also learned during wave 1 that midterm exam grading was pushed back to June 12, or the first planned roll-out day of the second wave of the experiment. It takes typically a week or longer for teachers to submit most of the grades. Engagement, by contrast, is based on administrative data from the phone provider and therefore is complete and instantly available.

Reading ability and engagement. Any increases in reading ability as a result of one of the IVR treatments is a combination of the child's exposure to the exercises, and how effectively the delivery and content of the exercises in this arm improve reading (efficacy of the arm for short). Data from wave 1 show that overall parental engagement differed between treatment arms, but was overall low (under 10% for all arms). This is a good level of engagement for phone based contacts, but low for a remedial fluency program.

Engagement, that is, parents actually listening to the exercises, is a necessary condition for exposure and ultimately fluency improvements. Moreover, we conjecture that parents are likely to engage more with an IVR arm if they feel that the child learns more, suggesting that engagement is (weakly) positively correlated with both exposure and treatment arm efficacy. However, we cannot a priori rule out a *negative* correlation. If this negative correlation is strong enough, the arms with the highest engagement may not be the arm that generates the highest fluency increases.

Reading ability and midterm/endterm cwpm assessment. Assessments of ORF through cwpm have been found to provide a useful measure of a child's ability to read. However, the small sample of parents who were actually taking up treatment, combined with the short treatment duration (half or less than half of a term) and the noise and missingness in the implementer's data, mean that ORF scores at midterm and endterm are unlikely to provide a precise measure of the treatment effects on reading ability.

After observing the first set of engagement data from wave 1 and learning about the issues with ORF data, we had to choose whether (i) to proceed as planned with two waves of nine exercise sets, but base the adaptive sampling on engagement; (ii) to base adaptive sampling on ORF from wave 1, but to delay the start of wave 2 by 2-3 exercise sets and shorten it in the process; (iii) to use some combination of the two outcomes (with the same consequences for wave 2 implementation), or (iv) to follow some different experimental protocol.

None of these options is optimal. But based on all the information available, and with the objective of choosing one IVR arm by end of term 3, we chose (i). In addition, we will use ORF information to estimate treatment effects on ORF, and in particular, we will test our conjecture above – that treatment efficacy is weakly positively correlated with engagement. Besides the reasons already outlined, an important influence on our final decision was that the implementer weighed in strongly in favor of that option, not least because they see parental engagement as an important objective in its own right.

Drawing Thompson shares and calculating Exploration Sampling shares. Since the school random effects are assumed to be additive in our model, average treatment effects $\{\beta_k^E\}_{k=1}^6$ do not depend on the random effect realizations in individual schools. We can therefore simulate the Thompson shares p^k using the posterior distributions of the parameters $\{\beta_k^E\}_{k=1}^6$ from the parent engagement model, in order to calculate the probability that each treatment arm k is optimal,

$$p^k = Pr(k = \underset{k'}{\operatorname{argmax}}\, \beta^E_{k'}) \ . \tag{6}$$

Then, we calculate Exploration Sampling shares q^k based on the Thompson shares p^k:

$$q^k = \frac{p^k(1-p^k)}{\sum_{k=1}^{6} p^k(1-p^k)} \ . \tag{7}$$

We stratify the wave-2 sample (i) at school level and (ii) by parent consent (see above) and allocate treatments within each stratum based on the Exploration Sampling shares q^k. Note that this adaptive sampling strategy assumes that on average, higher engagement translates 1:1 into greater improvements in the desired outcome, fluency.

References

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton (2017, November). From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. *Journal of Economic Perspectives 31*(4), 73–102.

Banerjee, A., E. Duflo, A. Finkelstein, L. F. Katz, B. A. Olken, and A. Sautmann (2020, April). In praise of moderation: Suggestions for the scope and use of pre-analysis plans for RCTs in Economics. *NBER Working Paper 26993*.

Banerjee, A. V., S. Cole, E. Duflo, and L. Linden (2007, 08). Remedying Education: Evidence from Two Randomized Experiments in India. *The Quarterly Journal of Economics 122*(3), 1235–1264.

de Barros, A. and A. J. Ganimian. Which Students Benefit from Personalized Learning? Experimental Evidence from a Math Software in Public Schools in India. *Working Paper*.

Kasy, M. and A. Sautmann (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica 89*(1), 113–132.

Muralidharan, K., A. Singh, and A. J. Ganimian (2019, April). Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India. *American Economic Review 109*(4), 1426–1460.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, 705–741.

Piper, B., J. Destefano, E. M. Kinyanjui, and S. Ong'ele (2018). Scaling up successfully: Lessons from kenya's tusome national literacy program. *Journal of Educational Change 19*(3), 293–321.

Rodriguez-Segura, D., C. Campton, L. Crouch, and T. S. Slade (2021). Looking beyond changes in averages in evaluating foundational learning: Some inequality measures. *International Journal of Educational Development 84*, 102411.

Vehtari, A., A. Gelman, and J. Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing 27*(5), 1413–1432.