Policy Research Working Paper 10098

Adaptive Experiments for Policy Choice

Phone Calls for Home Reading in Kenya

Bruno Esposito
Anja Sautmann

WORLD BANK GROUP
Development Economics
Development Research Group
June 2022

Policy Research Working Paper 10098

Abstract

Adaptive sampling in experiments with multiple waves can improve learning for “policy choice problems” where the goal is to select the optimal intervention or treatment among several options. This paper uses a real-world policy choice problem to demonstrate the advantages of adaptive sampling and propose solutions to common issues in applying the method. The application is a test of six formats for automated calls to parents in Kenya that encourage reading with children at home. The adaptive ‘exploration sampling’ algorithm is used to efficiently identify the call with the highest rate of engagement. Simulations show that adaptive sampling increased the posterior probability of the chosen arm being optimal from 86 to 93 percent and more than halved the posterior expected regret. The paper discusses a range of implementation aspects, including how to decide about research design parameters such as the number of experimental waves.

This paper is a product of the Development Research Group, Development Economics. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at asautmann@worldbank.org.

Produced by the Research Support Team

Adaptive Experiments for Policy Choice: Phone Calls for Home Reading in Kenya

Bruno Esposito[*] Anja Sautmann[†][‡]

Latest version here.

Keywords: adaptive experiments; multi-armed bandits; education technology; early literacy; Kenya

JEL codes: C11, C93, I25, O15

[*] Development Economics Research Group, World Bank, email: bespositoacosta@worldbank.org

[†] Corresponding author. Development Economics Research Group, World Bank, email: asautmann@worldbank.org.

[‡] The authors thank Tim Sullivan and Clotilde de Maricourt at New Globe and Peter Bergman for connecting us with them, as well as Grant Bridgman, Christine Vorster, and Faith Kibuswa at Uliza. Isaiah Andrews, Jiafeng Kevin Chen, Daniel Rodriguez-Segura, Adam McCloskey, and Kelly W. Zhang were extremely generous in sharing their expertise. We thank Kathleen Beegle, Daniel Björkegren, Esther Gehrke, Andrew Foster, Maximilian Kasy, David McKenzie, Robert Pless, and seminar participants at the World Bank and Rochester University for helpful comments and feedback. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. 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. All errors are ours.

1 Introduction

The use of experiments in research on economic development and policy represents one of the biggest methodological innovations in economics in the last few decades. For research that tests policy interventions and programs, the hope has been that rigorous experiments can lead to better policy decisions. In this context, the learning goal of a policy maker such as a government or NGO may be characterized as follows: they would like to improve a certain outcome, say in education or health, and are looking to identify the best (least expensive, most effective) policy to affect this outcome. We call this a "policy choice problem" for short.

Randomized control trials (RCT) as they are found in economics and related fields are typically aimed at identifying causal treatment effects and estimating these effects as precisely as possible, and design choices like equal-sized treatment groups and re-randomization or stratification support this goal. But this approach to experimental design is not ideally suited to inform a policy choice problem. To see why, note that sample sizes that deliver the power to statistically distinguish the effect sizes in multiple treatment arms from zero (and each other) quickly grow large. Yet for the objective of choosing and implementing only one of the tested policies, precise treatment effect estimates for low-performing options are not actually needed; ex post, some of the sample assigned to these arms could have been put to better use to distinguish the treatment effects in the highest-performing arms. This is particularly detrimental if the sample is small or there are budget and time constraints that prevent prolonged experimentation.

When the experiment can be carried out in two or more waves, the research design for policy choice can, in many cases, be improved by using adaptive sampling. The objective in the policy choice problem is to maximize the average outcome at the end of the experiment, or equivalently, minimize expected policy regret, that is, the expected loss from selecting a suboptimal arm. Choosing the arm with the highest outcome after repeated experimentation is a special case of the multi-armed bandit problem with a pure "exploration" motive, but no "exploitation" motive (Bubeck et al., 2009; Audibert et al., 2010). Efficient learning means adapting the assignment of experimental units to treatment arms based on what was learned in earlier waves.[1] The best adaptive learning strategy for policy choice tends to assign a larger share of the sample to higher-performing arms. This helps to distinguish these arms from each other while spending less effort on low-performing arms. In practice, researchers use sampling algorithms that approximate the optimal strategy to reduce computational burden.

This paper puts adaptive sampling for policy choice to the test by applying it to a real-world policy choice

[1] This is true for many different learning objectives, including but not limited to policy choice. It holds also for learning goals that usually motivate standard RCTs, such as efficient hypothesis testing: it is typically not optimal to randomly assign equal sample shares to all treatment arms in later waves, see e.g. Tabord-Meehan (2018).

problem in education technology. The goal is threefold: addressing conceptual and practical challenges in implementing adaptive experiments in policy settings, studying performance of an adaptive research design in a short time horizon where asymptotic performance guarantees may not apply, and, last but not least, informing the actual policy choice problem at hand.

The example we use is an experiment on using phone calls with interactive voice response (IVR) technology to deliver regular short reading exercises directly to parents in Kenya. The calls are intended to encourage parents to read with their children at home, a practice known to improve language acquisition and fluency (Mayer et al., 2019; York et al., 2019; Knauer et al., 2020). The implementer was NewGlobe, an organization that both supports public schools and operates its own community schools in several countries, including the Bridge Kenya primary schools in our sample. Faced with many options for IVR call and exercise designs, NewGlobe was looking to decide which call format (if any) they should roll out to all parents.

We use the IVR experiment to discuss in detail how to approach designing and conducting an adaptive experiment for policy choice – from estimation approaches, to the algorithm used, to research design decisions, for example about sampling and sampling size. In our example, we tested six different IVR call options during the third term of the 2020 school year. The experiment was designed to identify the call format with the highest level of engagement, measured as the number of IVR calls in which the respondent started the reading exercises, and to test whether IVR calls can increase reading fluency. The calls cross-combine two delivery formats for the exercises – parent-led vs. IVR-led reading – with three different ways of matching exercise contents to the child's reading level, motivated by evidence that targeted instruction can improve outcomes especially in the tails of the distribution (Banerjee et al., 2007; Muralidharan et al., 2019; Doss et al., 2019).

In order to efficiently identify the arm with highest call engagement, the experiment uses a version of the exploration sampling algorithm proposed by Kasy and Sautmann (2021a) to assign experimental units to treatment arms. Exploration sampling is a Bayesian bandit algorithm that was shown to perform well in both real and simulated experiments for policy choice and shares attractive asymptotic efficiency properties for best-arm identification of a set of similar algorithms (Russo, 2020; Qin et al., 2017; Shang et al., 2020; Kasy and Sautmann, 2021b). To our knowledge, there are to date only three policy choice experiments that have used it: an application in the original paper, a test of an SMS-based information campaign in India to reduce the spread of Covid-19 (Bahety et al., 2021), and a trial on contraceptive uptake in Cameroon that is ongoing at the time of writing (Athey et al., 2021). Instead of a Bernoulli outcome with a Beta prior as in the original paper (and many multi-armed bandit settings), we use a hierarchical binomial model to obtain assignment shares and parameter estimates.

The IVR experiment has only two experimental waves and the outcome distribution is more complex than assumed in theoretical treatments of Bayesian best-arm algorithms. We are therefore particularly interested how adaptive sampling influences treatment assignment, performance, and estimation results. We find that after just one wave, there is sufficient learning so that adaptive sampling leads to a substantial shift in the assignment shares: based on the different call success rates in each arm after wave 1, we obtained sample allocation shares for wave 2 that varied between 0% and 39%. After wave 2, we estimate a 93% probability that the best call design for engagement uses parent-led reading exercises and delivers the same intermediate-level exercises to all students. In this arm, parents engage – meaning, they start the reading exercises – with 8.40% probability per call, compared to 3.93% in the least successful arm. The arm with second highest engagement (7.43%) has only a 5% probability of being optimal. Expressing the expected policy regret in terms of average engagement probability, the selected treatment arm has an estimated 0.02% expected loss from potentially making a suboptimal choice, compared to 1.27%-4.49% for the other arms.

Even though the experiment targeted call engagement, we also estimate the treatment effects on oral reading fluency (ORF), using exam scores collected by the implementer. Although the data is noisy, we find that the arm with the highest level of engagement leads to estimated increases in ORF of 1.68 correct words per minute, equivalent to 0.065 standard deviations of the baseline data, with a credible interval between 0.13 and 3.21. The precision of this estimate is partly due to the large sample assigned to the best arm.

The results of the IVR experiment speak to an important policy question: whether there are low-cost, automated methods of increasing the probability that parents read with their young children, and what their best design might be. Especially during the Covid-19 pandemic, it became clear that there is an unfilled need for sustained learning at home and reaching children in families with limited educational and technological resources. Personal calls have been shown to be highly effective(Angrist et al., 2020b), but may require significant of resources. The experiment shows that mass-deployed IVR calls can increase parental involvement in the child's schooling but that the call design matters significantly for uptake.

Beyond these findings, the main contributions of this paper are an evaluation of the merits of using adaptive sampling for policy choice, and a detailed guide to implementation.[2] In particular, we use simulation approaches to examine the performance of the experiment and understand the impact of different design choices. In a first exercise, we compare 'ex post' the exploration sampling design with an alternative design with equal-sized (stratified) treatment arms, akin to a "standard" RCT, using simulated samples drawn from the experimental observations. This shows that adaptive assignment in only one wave achieved meaningful reductions in uncertainty – from on average 86% probability that the chosen arm is optimal in the RCT to 93% probability with exploration sampling – and reduced posterior expected policy regret by more than half from 0.05 percent to 0.02 percent engagement probability.

The next two exercises carry out 'ex ante' simulations based on the outcome model in order to determine

[2] This complements the excellent practitioner's guide on adaptive experiments by Hadad et al. (2021).

the gains from (a) conducting two experimental waves instead of one (non-adaptive) wave with the full sample, and (b) adding a second wave after having observed the outcomes of the first. These are examples of simulations a researcher might conduct to determine the research design, akin to power calculations. In case (a), the predicted reductions in expected regret seem plausible for specific parameter vector, but the flat prior distributions of the treatment effect parameters do not provide a good basis for simulating the gains from adaptivity; researchers may instead choose to focus on specific parameter values, not least to reduce computational burden (akin to power calculations where a minimum detectable effect size is imposed). In case (b), where the wave-1 posteriors can be used to simulate parameter draws, "agnostic" simulations do better, but they appear to still somewhat under-predict the gains.

As we go along, we discuss many details of implementation and experimental design, such as formulating and validating the Bayesian models for treatment effect estimation, calculating the expected posterior policy regret of each arm, and writing a pre-analysis plan. We address questions such as when an adaptive experiment is possible and when it may be most valuable, and approaches to correcting estimated treatment effects and confidence intervals for sampling bias and the "winner's curse" that affects the treatment effect estimate of the best arm (e.g. Melfi and Page, 2000; Andrews et al., 2021). We also spend some time discussing the trade-offs that were involved in choosing the targeted outcome.

The constraints on this experiment are representative of the decision contexts in which policy makers work day-to-day. In NewGlobe's situation, with a limited budget and only one school term available to test IVR, many organizations might decide against an experiment entirely—but our trial shows that adaptive sampling methods can enable rigorous learning even when the parameters of experimental design are severely constrained. The solutions we propose can help inform future adaptive experiments for policy choice, in EdTech as well as many other contexts.

The next section introduces the concepts behind adaptive sampling for policy choice, showing how the sampling algorithm used is determined by the objective of the experiment, describing the exploration sampling algorithm, and discussing the use of Bayesian estimation. It also lays out some considerations for choosing parameters of the research design such as the number of waves. Section 3 discusses the policy background, interventions, and experimental design of the IVR experiment, including the choice of targeted outcome, highlighting lessons for adaptive experiments in general. Section 4 discusses the data and details the models used for estimation, including how to derive the probability optimal and the expected policy regret, quantities used in the exploration sampling algorithm. Section 5 presents treatment effect estimates for parental engagement, shows the assignment shares based on these estimates, and discusses the impact on reading fluency. Finally, section 6 picks up the question of research design again. In the concrete context of the IVR experiment, we first show how the adaptive and a non-adaptive design compare 'ex post' in simulated samples from the experimental data, and then demonstrate how 'ex ante' simulations can be used to decide, for example, on the number of experimental waves. Section 7 concludes with a short discussion.

2 Using Adaptive Sampling in Experiments for Policy Choice

This section gives an overview over the use of adaptive sampling for policy choice and the exploration sampling algorithm proposed by Kasy and Sautmann (2021a). We start with the "basic ingredients" for an adaptive experiment: the objective, an algorithm that builds on the data from each wave to adaptively allocate units to treatment arms, the estimation approach, and constraints that determine whether an adaptive experiment is feasible. The corresponding features of the IVR experiment are described in detail in sections 3 and 4.

This section also discusses the gains from adaptive vs. non-adaptive sampling or adding adaptive waves, and how these gains can be calculated in simulations to choose the number of experimental waves. We return to this in section 6, building on the data collected in Kenya and the estimation results in section 5. To begin with, however, we assume that there are $t = 1, ..., T$ exogenously given consecutive sample draws (waves) of size N_t available for testing.

Objective. In the canonical policy choice problem, there are $K > 2$ policy options – or treatment arms – labeled $k = 1, 2, \ldots, K$. Each arm has unobserved (stationary) average outcome θ^k, and the policy maker wants to implement the arm with the highest average outcome. Formally, let $k^{(1)} = \operatorname{argmax}_k \theta^k$ be the true best arm, and k^* the arm that is chosen. We call the loss (per unit) from implementing a suboptimal arm k the policy regret, $\Delta^k = \theta^{k^{(1)}} - \theta^k$. Ex post, the policy maker will select the arm k^* that has the highest average outcome, or lowest policy regret, based on the observed data.

It is assumed that the outcomes of the experimental units are observed at the end of each period t. This means we can learn from wave t and adjust the allocation of units to treatment arms in wave $t + 1$, i.e. use adaptive sampling. In the policy choice problem, the policymaker's wants to implement an adaptive sampling strategy that maximizes welfare, that is, minimizes the expected policy regret from the final choice given the true (unobserved) vector of average outcomes: $E[\Delta^{k^*}|\theta]$. Adaptivity increases the efficiency of learning for a given objective by over-sampling some arms based on what was learned, at the expense of other arms (and other objectives).

Remark: Other Objectives. Large literatures consider sampling for specific learning goals.

The classical multi-arm bandit problem (MAB) considers the objective to maximize average outcomes during the ongoing experiment, or equivalently to minimize in-sample regret, which introduces the well-known exploration-exploitation trade-off (e.g Lai and Robbins, 1985; Bubeck and Cesa-Bianchi, 2012). The policy choice problem of choosing the arm with the highest average outcome can be seen as a special case of the MAB problem, where the experimenter has no "exploitation" motive (Bubeck et al., 2009; Audibert et al., 2010). Closely related to the "pure exploration" problem of policy choice is the problem of "best arm identification" (BAI), to the point that they are often treated as interchangeable. Here, the experimental design aims to either minimize the probability of choosing a sub-optimal arm after a given number of waves (the "fixed budget" setting), or minimize the expected number of waves to achieve a given level of certainty about which arm is optimal (the "fixed confidence" setting (Garivier and Kaufmann, 2016); see e.g. Lattimore and Szepesvári (2020) for an excellent and in-depth overview).

Even in non-adaptive experiments, common sampling techniques such as stratification and re-randomization aim to maximize power to detect a difference between treatment and control group (Athey and Imbens, 2017).[3] Adaptive strategies can further increase power for particular tests (Robbins, 1952). For example, Tabord-Meehan (2018) proposes an adaptive stratification procedure for a two-stage experiment with the objective of minimizing the variance of the estimator for the average treatment effect.

A Bayesian Bandit Algorithm for Policy Choice. Although the allocation of experimental units to arms for a given experiment of length T is a finite decision problem, determining the optimal allocation exactly is computationally prohibitively costly.[4] In the IVR experiment this is the case even with just one adaptive wave (see also simulations in section 6). This has led to the development of various heuristics for treatment assignment.

The exploration sampling algorithm used here is a Bayesian bandit algorithm: it starts from a prior over the model parameters – with identical priors for the k treatment effects – and updates the parameter distributions as the outcomes of each wave t are observed. The posterior distribution for the arm-specific θ^k is used to calculate the posterior probability that k is the best arm, $p_t^k = \Pr_t(k = k^{(1)})$ and the (posterior) expected policy regret $E_t(\Delta^k)$. In $t+1$, the algorithm assigns experimental units to arm k with sampling shares

$$q_t^{k'} = \frac{p_t^{k'}(1-p_t^{k'})}{\sum_{k=1}^{K} p_t^k(1-p_t^k)} \ . \tag{1}$$

Exploration sampling is a modification of Thompson sampling, which directly uses the probabilities p_t^k as the assignment shares in the next wave. Thompson sampling is a MAB heuristic for minimizing in-sample

[3] The specific objective also matters for stratification. Kasy (2016) considers stratification with continuous covariates and shows in a statistical decision theory framework that a deterministic design delivers maximal power for a given prior or a minimax decision criterion. However, Banerjee et al. (2020) argue that (some) randomization improves the ability to convince diverse and potentially adversarial audiences with a range of priors. The argument is relevant for adaptive designs as well: reducing the sample size of some arms in favor of other arms is likely to be the wrong decision under at least some priors about the true θ, and therefore an adaptive experiment is less convincing to an adversarial audience than a non-adaptive experiment.

[4] The supplement to Kasy and Sautmann (2021a) shows some simple examples of the optimal treatment assignment.

regret (Thompson, 1933). Compared with Thompson sampling and other algorithms that target in-sample regret, exploration sampling shifts measurement effort away from the best arm, increasing exploration and decreasing exploitation. This is because we need to learn not only about the best arm but also its close competitors for efficient policy choice. At the same time, it shifts measurement effort towards the higher-performing arms compared to an experiment with uniform assignment (i.e., equal sampling shares $1/K$), because information about the low-performing arms is unlikely to be relevant.

For the case of Bernoulli distributed binary outcomes with a Beta prior, Kasy and Sautmann (2021a,b) show that exploration sampling balances the sampling allocation in the limit at $T \to \infty$ between the sub-optimal arms, yielding constrained optimal posterior convergence (subject to the sampling share of the best arm converging to a pre-selected proportion). In the Bernoulli case, posterior expected regret converges at the same rate because regret is bounded by 1. Several Bayesian best-arm algorithms – applied to specific outcome distributions – have been shown to have this property (Qin et al., 2017; Russo, 2020; Shang et al., 2020).[5] Each heuristic has its own merits, but exploration sampling is appealing for its simple form that does not require a tuning parameter, its convenience for batch settings (waves larger than 1 unit), and its motivation based on sampling the best arm from the posterior for θ with the restriction of never assigning the same arm twice for increased exploration.[6] The existing theoretical performance guarantees apply only asymptotically and for specific outcome distributions. However, Kasy and Sautmann (2021a) demonstrate the good performance of exploration sampling for expected policy regret in the Beta-Bernoulli case in simulations based on pre-existing data and for posterior convergence in an experiment testing different enrollment methods for an agricultural extension service. We also use simulations in section 6 to assess the gains from exploration sampling over uniform assignment.

In the IVR experiment, the primary objective was to identify the best arm measured by the parents' engagement with the IVR calls. We therefore used exploration sampling on 6/7th of the sample. A secondary goal was to understand whether the IVR calls have an effect on reading ability. The design therefore included a (fixed) control group of 1/7th of the sample for identifying the time trend in reading fluency and estimating treatment effects (see sections 3 and 4). Designs that combine adaptive treatment arms with a control group

[5] All are"top-two" algorithms based on expending greater measurement effort on the current best two arms, with a tuning parameter β determining the allocation between them as well as the limit sample share of the best arm. Russo first proposed three algorithms and establish constrained optimal posterior convergence for a family of outcome distributions: Top-Two Probability Sampling (TTPS), Top-Two Value Sampling (TTVS), and Top-Two Thompson Sampling (TTTS). Top-Two Expected Improvement (TTEI) by Qin et al. (2017) modifies the expected improvement algorithm for Gaussian outcomes. The authors also show that the algorithm is asymptotically optimal in the fixed-confidence setting, which requires that the limit allocation is attained in finite time. Shang et al. (2020) propose a version called Top-Two Transportation Cost (T3C) that is less computationally demanding than TTTS and applies to a larger set of outcome distributions than TTEI, and prove optimality of both TTTS and TTEI in the fixed confidence setting for Gaussian outcomes. Finally, they establish posterior convergence for TTTS for Normal and Bernoulli distributed outcomes.

[6] Thompson sampling is equivalent to taking simple draws from the posterior without prohibiting repeat assignments. TTPS and TTVS determine two "top" candidate arms in each wave and randomly select the first with probability β and the second otherwise, making them poorly suited for batch allocation. TTEI is specific to normally distributed outcomes. Exploration sampling is closest to TTTS and T3C and with $\beta = 0.5$ all three converge to the same limit allocation.

are also used by Bahety et al. (2021) and Athey et al. (2021).

Estimation. In the IVR application, we focus on Bayesian estimation to obtain final parameter estimates. In Kasy and Sautmann (2021a), the outcome of each arm k is Bernoulli distributed with a Beta prior, so that the posteriors after t have closed forms. In the IVR experiment, we generalize the approach and estimate Bayesian hierarchical models with school-specific effects and a Binomial outcome distribution (Normal for reading fluency), described in detail in section 4. The Bayesian approach with updating between waves is internally consistent[7] and naturally produces p_t^k that we need for exploration sampling. Bayesian inference is valid with adaptively collected data.

However, users may be interested in frequentist inference about the parameter estimates. Frequentist estimates that do not account for the data being generated by an experiment for policy choice are subject to potential biases (Melfi and Page, 2000; Xu et al., 2013). First, observations from an adaptive experiment cease to be iid draws – intuitively, adaptivity introduces sampling bias because random fluctuations in early observations in a given treatment arm k affect the weight of these observations in the overall sample assigned to k (by changing the assignment shares of this arm in future waves). Second, inference on the best arm out of a set, where the ranking is based on the treatment effect estimates, creates an upward bias and invalidates standard confidence intervals even with non-adaptive sampling (Andrews et al., 2021).

Inference from adaptively sampled data is an active field of research, with particular focus on algorithms targeting in-sample regret, which exacerbate selection bias by quickly focusing on high-performing arms. Adaptively weighted estimators can correct sampling bias and produce asymptotically normal estimators (Hadad et al., 2021; Zhang et al., 2021). Andrews et al. (2021) propose corrections for the "winner's curse" when estimating the average outcome of the highest-performing arm that apply to asymptotically normal estimators. To our knowledge, there are to date no approaches that can provide confidence intervals with correct coverage for the optimal arm in an adaptive experiment in a model with random effects as we used in the IVR experiment. However, in section 5 we estimate a frequentist Binomial model for engagement and illustrate how the estimates are affected when (a) applying the weights proposed in Zhang et al. (2021) to restore asymptotic normality and then applying the winner's curse correction by Andrews et al. (2021).

Remark: Hybrid Algorithms. Given the problems with inference in adaptive procedures where low-performing arms are under-sampled, recent applications have used modified algorithms for a hybrid goal of (frequentist) estimation as well as regret minimization. For example, the "tempered Thompson" algorithm in Caria et al. (2020) uses a convex combination of Thompson shares and $1/k$ equal-sized shares. Another common

[7] In principle, updating the posterior from any earlier wave with the data collected afterwards should lead to the same posterior outcome distribution at t, including re-estimating the model with all the data collected and the initial prior, which is in practice the method we use.

modification is to impose a lower bound on the sampling share in each arm ("clipping", applied e.g. in Athey et al. (2021) with exploration sampling). Such modifications can be combined with setting aside a sample share for one experimental arm and in particular a control group, see for example the "control-augmented" Thompson sampling algorithm in Offer-Westort et al. (2021).

An important decision for research designs in practice is the size of the experimental sample N. The MAB literature often assumes that experimental units arrive through an exogenous process and can be used costlessly for experimentation, often indefinitely. In practice, researchers using adaptive experiments need to decide how to split the sample into waves, or how many waves of fixed size to conduct. We approach these questions in two steps, by first discussing constraints that delineate the space of possible experimental designs, and then outlining how to assess alternatives within these constraints.

Constraints on Adaptive Experimental Designs. The use of multiple waves imposes some constraints on the set of possible adaptive experimental designs. We outline these here briefly, partly to illustrate when adaptive design are in practice feasible.

Total time D^{max} available. Due to external constraints, such as funding timelines or deadlines for operational deliverables, the maximal duration of an experiment is typically limited.[8]

Comparable waves. Most bandit algorithms assume some form of stationarity, e.g. that the observations in all waves represent iid draws of the potential outcomes in the population. For efficient learning across waves, the treatment effects must be stationary and any time trends must be common to all arms. Annual cohorts of students or batches of survey participants recruited at random may fulfill these conditions, but e.g. job seekers in a seasonal industry at different times of the year likely do not.

Length of a wave d. To complete a wave, the intervention must be administered in full, outcome changes in response to the treatments must have manifested, and post-intervention outcome measures must be collected before the start of the next wave. This determines wave duration d.

Together, these constraints typically impose a limit on the number of waves T^{max}. If the policy environment changes rapidly, data is collected in a time-consuming survey, or the available time does not include two comparable periods, only one "wave" may be possible, $T^{max} = 1$. On the other hand, if a wave takes only hours or days and data are automatically recorded, many waves may be possible, e.g. $T = 10$ in Bahety et al. (2021) or $T = 17$ in Kasy and Sautmann (2021a).

Other constraints may limit e.g. the maximum sample size per wave or the total sample N^{max}. In the IVR experiment, due to time and comparability constraints, the choice was effectively only between conducting

[8] Such a limit is a reason to use policy choice algorithms that minimize expected regret after the experiment, rather than an algorithm that simply continues indefinitely and targets in-sample regret.

one or two experimental waves. The available sample was the full population of first graders in the Bridge Kenya schools in term 3 of 2020, see section 3.

Choosing the Research Design. Even if constraints narrow down the design space, the experimenter may still need to decide whether to use adaptive sampling and choose sample size, wave size, and number of waves to run. An added consideration is that even for a given sample size N, there are some costs to conducting testing in waves.

Per-wave Implementation Costs. Maintaining the infrastructure for data collection and interventions for all treatment arms, including the human capital costs of managing the experiment, adds fixed costs c_t per wave, on top of any per-unit costs c_i^k (which may vary by treatment arm).

Cost of Delay. Each new wave adds delay until the gains from the experiment – the average estimated treatment effect of the best arm – are realized for all potential beneficiaries.

Balancing these costs are the efficiency gains from adaptivity. It is computationally involved to estimate these gains, and so the researcher can typically only consider a small number of designs. Here, we briefly discuss two situations that will frequently arise in practice. First, experimenters often have a fixed N available and have to decide whether and how to divide the sample into waves. Second, the experimenter may need to decide at time t whether to run an additional wave in $t+1$.

This could be set up as a simple optimization problem. For example, consider choosing the number of waves $T \in \{1, \ldots, T^{max}\}$ for given sample size N, so that the wave size is $N_t = N/T$ (assuming equal-sized waves for simplicity). We would expect more efficient learning with more waves and more chances to adapt, and indeed the simulations in Kasy and Sautmann (2021a) with data from three existing experiments show how splitting the sample into 2, 4, and 10 waves monotonically shrinks the expected policy regret. In practice, however, the marginal gains are likely decreasing in T.[9] Moreover, the gains must be weighed against the cost. The experimenter might solve

$$\max_{T \in \{1,\ldots,T^{max}\}} \left(\delta^{T+1} E(M\theta^{k^*}|T), -\sum_{t=1}^{T} c_t \right).$$

The second term penalizes the cost of increasing T. The first term is the term of interest: the expectation of the number of beneficiaries M times the per-person outcome in the chosen arm θ^{k^*}, discounted by δ^T due to the implementation delay.

[9] This is at least in part due to indivisibility issues. As T grows and the wave size shrinks, it becomes harder to implement the adaptive algorithm faithfully, and the actual assignment shares may differ substantially from the exploration sampling shares $q_t^k N_t^k$, especially if the sample is also stratified (see also section 5). With many treatment arms, in small waves some arms may not be assigned at all, updating about these arms will proceed slowly in terms of t, and the assignment shares may remain far from optimal for a long period of time.

In the second situation we defined, waves have fixed size N_t and the experimenter needs to decide when to end the experiment. In addition to the per-wave and delay costs given by c_t and δ^{T+1}, increasing T incurs $q_t^k N_t$ times the per unit cost c_i^k for each experimental arm.[10] In exchange, the experimenter observes additional N_t units in each wave t.

In each case above, the researcher needs to estimate $E(\theta^{k^*})$ as a function of the research design. Since closed forms are not typically available, these projected gains from adaptivity have to be obtained from simulations. This requires simulating not only experimental outcomes under different random sample draws, but also the different sampling paths that arise from adaptivity. The experimenter is typically restricted to comparing only a few hypothetical θ and a small number of possible research designs. We illustrate such simulations in the context of the IVR experiment in section 6.

3 IVR Calls for Reading in Kenya: Background and Experimental Design Choices

3.1 Background and Setting

Our application for adaptive sampling for policy choice is an EdTech intervention that uses interactive voice calling aimed at encouraging parents to read with their children. The implementing organization ("the implementer") is NewGlobe, the parent of Bridge International Academies. At the time of the study, NewGlobe operated 112 private primary schools all over Kenya.[11] The Kenyan school year usually has three terms that start just after New Year's and end late October. Due to Covid-19, the 2020 terms 2 and 3 took place 1/3 - 3/19 and 5/10 - 7/16 of 2021 (with the 2021 terms compressed into 7/26/21 - 4/2/22). All Kenyan schools at the implementing organization had introduced oral reading fluency (ORF) assessments for the first time in the midterm and endterm exams of term 2 of 2020.

The implementer wanted to make a decision about whether and how to use interactive voice response calls (IVR) to encourage parents to do reading exercises with their children. Reading with a child at home has benefits for language acquisition and fluency, even in contexts where parents themselves may have limited reading skills (Mayer et al., 2019; York et al., 2019; Knauer et al., 2020). Kenyan schools were closed for part of 2020 due to COVID-19, highlighting the benefits of developing effective home interventions targeting reading and numeracy.[12] More broadly, parental engagement is an important determinant of children's long-term success in school. Recent research has shown that relatively light-touch interventions such as personalized

[10] It may also reduce the number of beneficiaries by the additional experimental subjects.

[11] The schools follow Bridge's specific teaching model and charge fees; these fees are lower than typical private school fees and similar to the administrative costs of public schools.

[12] Prior research has shown that parental engagement interventions can counteract the detrimental effect of extended periods out of school (e.g. Kraft and Monti-Nussbaum (2017)). A combined text message and phone call intervention was able to reduce learning loss during COVID-19-related school closures in Botswana (Angrist et al., 2020a).

text messages increase parental engagement, which in turn improves early literacy outcomes (York et al., 2019; Doss et al., 2019). For older children, parental engagement also increases parents' information about attendance and performance at school and improves outcomes through this channel (Berlinski et al., 2021; Bergman and Chan, 2021; Bergman, 2021; Bettinger et al., 2021).

While many parent communication interventions rely on text messages, parental literacy barriers and length restrictions limit text messaging as a tool to deliver reading exercises (ICTworks, 2016). The implementer already routinely uses text messaging to contact parents with information about their kid's schooling, and collects phone numbers and consent for this purpose. However, to what degree these messages are received and read by parents, and whether they lead to behavior change, is only incompletely known. In an earlier trial with the same implementer in Nigeria, which used text messages to encourage parents to use a WhatsApp-based quiz platform, almost none of the message recipients engaged with the quizzes (Sautmann, 2021b).

Phone calls provide an alternative that may sustain higher rates of engagement and allows longer interactions and better instructions for home exercises. Personal calls have been shown to be effective for increasing parental engagement (Kraft and Monti-Nussbaum, 2017), but are costly and time consuming for teachers. IVR calls are pre-recorded and automated, designed by recording a set of modular text snippets and jingles that are sequenced in response to listener input through the keyboard or through spoken word. There is to date limited evidence on the effectiveness of IVR for improving early literacy. A small pilot with 38 families in rural Côte d'Ivoire reports encouraging qualitative results on the use of IVR to foster phonological awareness in low-literacy environments (Madaio et al., 2019).

3.2 IVR Intervention Design

During piloting and discussions prior to the experiment, it was decided to test six IVR call variants. All treatment arms consist of twice weekly calls to the parents' phone. The IVR delivers a sequence of reading exercises, either based on letter combinations or words that the parent notes down during the call, or based on passages from the children's term 3 homework book. An experimental wave contains 9 sets of calls (see below), and each call contains 4 different exercises. The exercises change from call to call. Before each wave, we conducted a phone based opt-out procedure that explained the calls and also allowed parents to change the enrolled number. The full intervention design, call logic trees, and sample recordings of two of the interactive calls can be found in an online supplement (Sautmann, 2021a).

All IVR 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

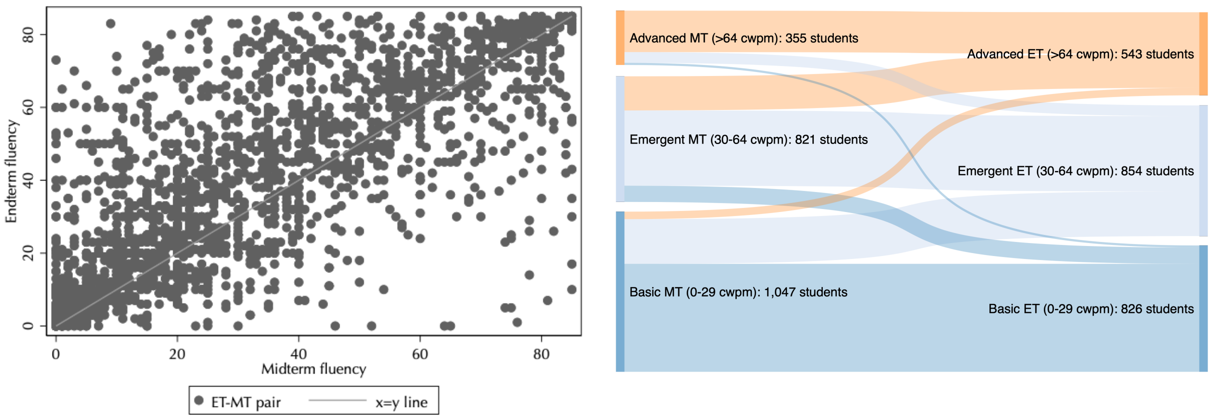


Figure 1: Term 2 midterm and endterm oral reading fluency scores, in units of correct words per minute, as used for exercise level assignment. The left panel shows that individual student scores are only noisily correlated. The right panel shows that there is some movement from higher leveling categories to lower ones, as well as small but significant numbers of students "skipping" from basic to advanced level. For 22.5% students in our sample, score information was missing.

back. This is a common method in Kenya that avoids calling charges to the parents.

We used three different ways of choosing a difficulty level for the exercises, and two different delivery formats, described in detail below. We cross-combined the 3x2 interventions to create the 6 treatment arms. In selecting the tested interventions, the aim was to create treatment variations that were genuine "contenders" for having the greatest impact on how often parents read with their children at home.

Varying exercise leveling. Baseline information on oral reading fluency (ORF) from term 2 showed high variation in reading scores, in line with other comparable data in developing-country contexts (for instance, see Muralidharan et al., 2019). In the presence of such variation, prior evidence has suggested that there can be benefits to leveling remedial programs (see, e.g., Banerjee, Cole, Duflo, and Linden, 2007; Banerjee, Banerji, Berry, Duflo, Kannan, Mukerji, Shotland, and Walton, 2017), and that customized EdTech interventions could benefit the lowest achieving students the most (de Barros and Ganimian, 2021; Doss et al., 2019).

However, our analysis of ORF scores showed that the available test data are very noisy (as seen in figure 1) and 22.5% percent of the sample were missing at the start of wave 1. There is reason to believe that there is selection bias in non-missing scores (see also below). This could make leveling based on observed or imputed past scores ineffective or even counterproductive. An alternative is to leverage parents' knowledge of their child's reading skills and let them choose the difficulty level during the call. But parents may be unable to accurately assess their child or may choose a poorly suited exercise, perhaps because they themselves are not secure readers or because their view of their child is too optimistic. A call that allows choice also takes longer, and parents may stop using the system if they find it fatiguing or challenging.

Based on these considerations, three intervention variants were chosen, (A) leveling on actual or imputed baseline scores, (B) providing the same sequence of intermediate-level exercises to all kids, and (C) giving parents a choice of exercises from a menu. Arm A uses observed fluency scores from the end of term 2 and assigns students with fluency scores of 0-29 into the "basic" group, 30-64 into the "intermediate" group, and 65+ into the "advanced" group. These cutoffs were used previously in a similar context (see Piper et al., 2018). Students with missing scores are assigned their class median. Whole classes with missing scores are assigned to the intermediate group (which also happens to be the full sample median). The exact exercise sequences in the basic, intermediate, and advanced groups are described in detail in Appendix A. Arm B assigns all students to the intermediate group, while Arm C allows parents to pick one exercise type (from basic letters, to letter combinations, to advanced text passages) out of a set of three.

Varying delivery format. We also test two formats that use the IVR functionality in different ways. In the first, the voice call explains to the parent how to do the reading exercises and asks them to carry them out with their child after the call (T1). In the second, the IVR asks parents to put the call on speaker phone, and then goes through the exercises with the parent and child on the call (T2).

A priori, either approach might work better for different reasons. In both call types, the parent is asked to take notes on the exercises during the call. The parent is instructed to point to the written letters or words while the IVR (or the parent) reads, and then again while the child reads. However, in T1, the parent may not pronounce letter combinations correctly from memory. She may also listen to the exercises during the call but then not carry them out with the child later. On the other hand, T2 may cause difficulty if the phone's speaker is poor or the IVR moves too fast for the child or is not responsive enough. All parties may be more motivated when the child and parent practice together, rather than following instructions from an unknown and disembodied voice.

3.3 The Research Design

A "standard" RCT of IVR for home reading would likely consist of extensive piloting, carrying out power calculations to determine sample size and number of tested intervention arms, randomizing at the cluster (school) level, and then administering an IVR program for at least a full school year, possibly accompanied by a home survey and independent tests of reading fluency. Based on the budget for delivering and deploying messages, the size of the sample, and the available staff time, such a comprehensive study was not feasible for NewGlobe. At the same time, at the outset, it was not even known whether parents would listen to the messages at all, and there is to our knowledge no existing guidance on how best to design such calls. In such a situation, NGOs and policy makers might resort to simply not using experimental methods. They might conduct an informal pilot, implement the program at scale, and then "tinker" with it after roll-out, or conversely, simply abandon the idea. Adaptive sampling could offer a solution that enables a rigorous experiment and makes the most of the limited sample and time available. The implementer saw as an attractive feature from an ethical perspective that even during the experiment, a larger share of participants benefit from the higher-performing treatment arms.

Objective. In conversations about the experiment, on the one hand, the implementer wanted to identify the "best" IVR call variant, and on the other, they wanted to verify that IVR calls with reading exercises actually have positive effects on reading fluency. This hybrid goal was a reason to keep a control group that received no intervention. At the same time, it suggested to use adaptive sampling to choose between the six call formats. We discuss how the notion of the "best" IVR call translated into the choice of targeted outcome below.

Constraints on the experimental design. The implementer was able to set aside only one first grade cohort and one term of the school year for testing the IVR calls, both due to other ongoing studies and due to the implementer's internal cost-benefit assessment.

The Kenyan school term is 10 weeks long, split equally into 5 weeks from start to midterm exams and from midterm to endterm exams. Reading tests are conducted as part of these exams, providing an administrative source of data. Moreover, the rhythm of the school term from start to midterm and from midterm to endterm is similar. For example, parents' attention to their child's schoolwork may increase closer to the exams.

Relative to the cost per call, the cost (in terms of both money and time) of developing sequences of reading exercises and recording them is high.[13] There was also concern that too many contact attempts from the school create fatigue in parents, especially with pilot programs that may not yet be optimally designed. For both reasons, an exercise sequence covering one half of the term was preferred to running the interventions for a full term.

Jointly, these constraints reduced the space of possible research designs to conducting one or two experimental waves in the first and second half of the term, with the total of first graders enrolled that year across all schools as the available sample.

Outcome measurement. The available outcome variables were take-up of the IVR calls, or call engagement for short, and oral reading fluency (ORF) scores collected by the school. Measures for both outcomes

[13] The exercises were developed by the implementer together with the research team. Dozens of sound snippets were recorded by a voice artist hired by Uliza. A first set of exercises was piloted with a small sample of parents in an older age group before completing all the exercises and recordings.

were provided by the implementer, with random ID numbers replacing parents' phone numbers and the child's name and school.

We use IVR provider records to measure engagement with the calls, that is, whether the call recipient actually starts the exercises. 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, which requires tapping a phone key to confirm. We define a parent as having engaged in one of the twice weekly exercise sets if the IVR made at least one successful call in that set. Since there are 9 exercise sets per wave, engagement can take values between 0 and 9. Call records are available immediately, and they are complete and accurate.

The implementer measures children's ORF scores during the midterm and endterm examination periods. In 2-3 hour periods set aside for the fluency test, a teacher examines each child by counting the number of words on a list that 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 teacher then submits the scores to the school's grade record system. ORF scores range between 0 and 85 correct words per minute (cwpm) based on the length of the provided word list.[14]

There are a number of issues with ORF measurement, which were partly revealed only after wave 1 of the experiment had already started. Figure 1 shows the high variation in ORF scores between midterm and endterm. Among the non-missing scores, an unusually high proportion are multiples of five, and in some classrooms, there are implausibly many very high scores. In addition, a high percentage of scores are missing or submitted late to the recording system: ORF scores were available for only 73.5%-88.9% of children depending on the exam.[15] Teacher reports on why a given score is missing are often ambiguous. Overall, the data quality for ORF scores is fairly low.

Targeted Outcome. In order to use adaptive sampling, it is necessary to define an outcome measure that decides which is the "best" arm, which in turn determines which treatment arms will be sampled more. In many settings, this is not straightforward, given that multiple indicators related to the desired outcome(s) are typically available. Here, the implementer wants to increase parents' engagement in their children's education in general, because parental engagement is known to have positive effects on children's performance in school; at the same time, the calls explicitly encourage a set of reading exercises with the aim

[14] The implementer chooses a standardized, grade-appropriate word list, trains teachers and provides equipment. The measure can in principle range from zero to over 200, but for first graders it is typically not above 120.

[15] The total share of scores that are multiples of 5 is 36%, and the observed score distributions show unusual heaping even when accounting for censoring at 0 and 85. Teachers sometimes delay submission or entirely fail to submit exam scores for their class. We describe the patterns of missingness and suspected rounding in more detail in Appendix B in Tables A.1 and A.2. Part of the reason that the problems of missing and rounded scores persist is that at elementary school level, these scores do not affect the student's progression into the next grade, nor do they affect the teacher's evaluation.

to improve reading. Call engagement measures whether parents listen to the reading exercises, but we do not observe the interactions they have with their children. As discussed above, ORF scores are an imperfect measure of the child's reading ability.

In principle, both call engagement and ORF scores could be used to create a combined outcome measure. Moreover, if there is a (known) relationship between the two measures, e.g. higher call engagement implies greater reading improvements and the reverse, then an adaptive experiment could equivalently target either outcome.[16] A priori, we conjectured that call engagement is positively correlated with reading gains. First, someone actually listening to the exercises is a necessary condition for the child's exposure to these exercises. Beyond the first couple of calls, a simple model of marginal returns also suggests that parents are more likely to engage with the calls if they feel that the child learns something and they plan on actually doing the exercises. However, there could be reasons that call engagement and ORF are not aligned: any increase in reading ability is a combination of (i) the child's *exposure* to the exercises, and (ii) conditional on exposure, how effectively the delivery and content of the exercises in this arm improve reading (*efficacy* of the arm for short). Treatment variants (T1) and (T2) could potentially have different exposure, conditional on observed call engagement, and the treatment arm design choices regarding leveling (A, B, and C) may exhibit differences in efficacy.

Without any constraints, the implementer might have chosen the best arm based on a weighted average of ORF and engagement. However, we were unable to determine assignment shares in wave 2 based on the midterm ORF scores.[17] The grading day was moved during wave 1 and took place after the start of the second half of the term. Due to the submission delays described above, ORF scores "trickle in" for several weeks, and even after the end of the term, more than a quarter of the midterm data was missing (see Table A.1). The choice in practice was therefore to either exclusively target call engagement in an adaptive experiment, start the second wave late and with incomplete data for some form of adaptive assignment based on ORF scores, or conduct an experiment with uniform assignment (or abandon the test).

In this decision, it played a role that even in the best case of timely and accurate ORF measurements, any effects of IVR calls on reading ability were likely to be only incompletely realized by the end of the trial intervention. Comparable early-reading interventions measure effects after an intervention period of several months or a whole school year (Doss et al., 2019; York et al., 2019). Moreover, cumulative effects – e.g. due to habit formation – are likely to accrue for a significant period of time after intervention end, so it is

[16] This relationship would need to be established, e.g. from pilot data. Caria et al. (2020) make reference to the literature on statistical surrogates – measurable or short term outcomes that can "stand in" for harder to measure or longer-term outcomes – to argue that adaptive experiments could target short-term outcomes to achieve higher welfare in the long term; see also Athey et al. (2019) for a proposal to create "surrogate indices" from multiple variables.

[17] Initially, we planned to use adaptive sampling to target ORF scores. The change is documented in the pre-analysis plan, see (Sautmann, 2022).

unlikely that the impacts of the treatment were already fully realized by the end of the term.

Based on these considerations, it was decided to exclusively target call engagement. Ultimately, the implementer valued parental engagement sufficiently to focus on maximizing call response rates, rather than attempting to choose a treatment arm based on very noisy effect estimates of fluency gains and risking inconclusive results. Another way to view this decision is to maximize learning about which arm has the highest call engagement rates, at the expense of learning more precisely which arm has the greatest ORF gains. While this solution may not be optimal, it reflects another reality of policy choice, that policymakers sometimes have to make do with imperfect data.

Sample and Randomization. We determined the sample using the phone number on record for the parent.[18] We dropped 2 schools that had fewer than 5 students, and 2 schools with very inflated ORF scores, leaving us with 108 schools with 3,163 unique student-phone number combinations.

We first randomly assigned half of the sample to wave 1 and 2 (1,581 and 1,582 phone numbers, respectively). We did not formally assess the best sample split between first and second wave, but small-sample simulations support equal-sized waves (see supplement of Kasy and Sautmann (2021a)). Before the start of each wave, parents received an introductory call, followed by a text message confirming enrollment and explaining procedures for opt out and for switching phone number. Some parents opted out explicitly and some phone numbers were invalid, leaving a sample of 1,494 in wave 1 and 1,384 in wave 2.

The randomization was stratified at the school level.[19] In wave 1, the assignment shares for the 6 treatment arms were equally 1/7; in wave 2, we used the assignment shares given by exploration sampling, keeping 1/7 of the sample as a control group in each wave. Due to indivisibilities, the total shares are close but not equal to the targeted shares, as shown in Table 2 in section 5.

Estimating ORF effects. In many applications, outcomes other than the targeted outcome are of interest to the experimenter. Here, we estimate the effects of the treatments on ORF with reading fluency exam scores obtained after the experiment was completed to learn whether the treatment arm with the highest engagement sees increases in children's reading performance. We also briefly discuss the possibility that there are differences in how engagement with the calls translates into reading gains, which might imply that the call format with the highest call engagement may not be the format with the highest reading gains.

[18] The implementer has parental consent to use this phone number for school related communications. Based on enrollment data from the start of term 3, we randomly selected one student ID for measurement in the few cases where several student IDs were associated with the same parental phone number (likely siblings). Phone numbers and schools are de-identified by the implementer before sharing with the researchers.

[19] We also stratified assignment on whether the opt-in call or confirmation text message were answered. For example, in wave 1, a large proportion of the sample (796 student IDs) neither opted in nor explicitly opted out. However, the extensive-margin results (Appendix C.4) showed that most numbers answered the phone at least once during the experiment, and so we ignore this in the estimation.

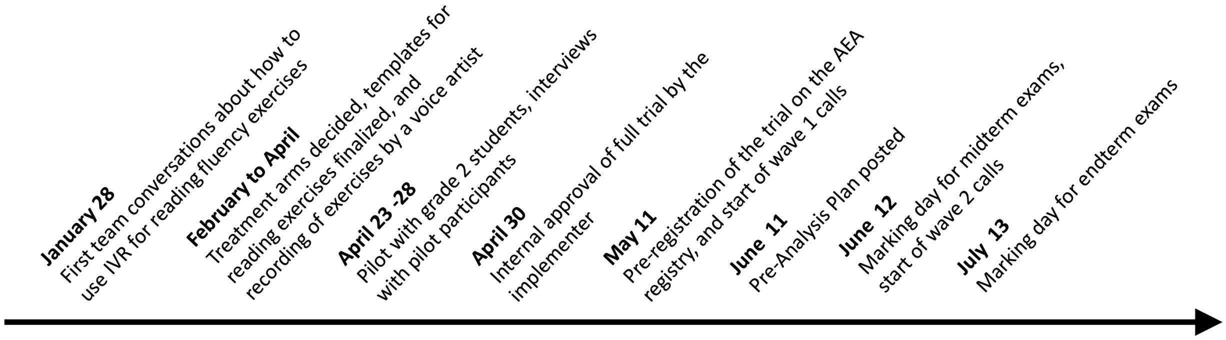

January 28
First team conversations about how to
use IVR for reading fluency exercises
February to April
Treatment arms decided, templates for
reading exercises finalized, and
recording of exercises by a voice artist
April 23-28
Pilot with grade 2 students, interviews
with pilot participants
April 30
Internal approval of full trial by the
implementer
May 11
Pre-registration of the trial on the AEA
registry, and start of wave 1 calls
June 11
Pre-Analysis Plan posted
June 12
Marking day for midterm exams,
start of wave 2 calls
July 13
Marking day for endterm exams

Figure 2: Timeline for the study.

Implementation and Pre-Specification. Figure 2 shows that the IVR experiment was carried out on a very short timeline. Development and implementation, including designing the reading exercises and recording and programming the calls for all treatment arms, were completed in three months up to April 30. The research team developed the statistical model for parental engagement and carried out the treatment assignment during wave 1 (term start May 11 to midterm exam on June 12), and the model for estimating reading fluency during wave 2 (midterm to endterm exam July 13) and after. This has downsides; for example, not enough pilot data was available to improve our priors e.g. about school random effects, and while the first wave of the experiment was ongoing, new information was still learned, such as the delay to obtaining ORF scores. The timeline also shows that the experiment was pre-registered prior to the first wave, but by the time the pre-analysis plan was filed on June 11 (before start of wave 2), the plans for the experiment had changed significantly. In general, short time windows and incomplete information in the experimental design phase may make adaptive sampling more attractive, but will also make pre-specification more challenging.

Remark: Pre-Analysis Plans and Trial Registration. A question for the research community will be whether adaptive policy choice experiments should be subject to the same norms of registration and pre-specification as "standard" experiments for causal effect estimation.[20] A full analysis of the incentives at play requires a larger body of evidence on the method, but a priori, the need for *pre-specification* seems less pressing: depending on context there is often no specific incentive to demonstrate the effectiveness of one treatment arm over another; the metric of expected policy regret has no established cut-offs akin to p-values for conventional significance levels; and, most importantly, after the first adaptive wave it is not possible to change the targeted outcome or the estimation approach, creating commitment before the data is fully

[20] Results showing significant effects often have higher value to both researchers and policy organizations, which contributes to issues such as data mining, the file drawer problem, publication bias (e.g., Andrews and Kasy, 2019) and so on, familiar from the literature on research transparency (Christensen and Miguel, 2018).

known. The opposite is true for trial *registration*: policy choice experiments are likely to be used to learn about the effectiveness of many different policy options for the same outcome. Adaptive trials may inform preliminary work where less successful interventions are never implemented or tested at scale. The file drawer problem seems particularly salient in the context. In fact, a natural extension of adaptive sampling across waves is to incorporate existing evidence into the priors that inform the research design of new experiments (see e.g. Pouzo and Finan, 2022). This form of iterative learning requires a complete record of all prior evidence gathered on the treatments under consideration.

4 Models and Estimation

This section describes how we estimate treatment effects on parental engagement and ORF measures, and how the engagement estimates are used for adaptive treatment assignment and final arm choice. We also comment on the modeling choices and implications for policy choice experiments more generally.

4.1 The Models for Call Engagement and Oral Reading Fluency

Call Engagement. Let Z_i^{sk} be the number of successful calls to a parent of child i in school s allocated to treatment arm $k \in \{1, \ldots, 6\}$. We assume that potential engagement is stationary across the two terms and for simplicity suppress the index for wave t. No calls were made to the control group, so we restrict the sample to enrolled phone numbers in the 6 treatment arms. We assume that Z_i^{sk} is a draw from a Binomial distribution with at most 9 successes and average probability of engagement $\theta^{sk} \in [0, 1]$. This is motivated by the distribution of the observed numbers of successful engagements in each treatment arm, shown in Appendix C.1. We model the average engagement probability with a hierarchical logistic regression model with school random effects. Thus, we have

$$\begin{aligned} Z_i^{sk} \mid \theta^{sk} &\sim Binomial(9, \theta^{sk}) \ , \\ \theta^{sk} &= \text{logit}^{-1}(\beta^E x^k + \kappa^E \eta_s^E) \ . \end{aligned} \tag{2}$$

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. We do not include baseline ORF information in this model − the only individual-level information we have − because of the problems with missing and noisy data outlined earlier.

We do not have much prior information on expected engagement, so we use a non-informative improper prior on $\{\beta_k^E\}_{k=1}^6$ and a Half-Normal prior distribution for κ^E (the standard Normal on $[0, +\infty)$), and assume a Standard Normal distribution for the school random effects.[21]

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

The hyperparameters $\{\beta_k^E\}_{k=1}^6$ and κ^E describe the average engagement probability in each treatment arm and the arm-independent variance of the engagement probability across schools. Each θ^{sk} is a realization of the average success probability specific to the school and treatment arm.

While our main estimates focus on the average number of calls per phone number, in appendix C.4 we also report estimates of the extensive margin of take-up. These use binary logit models, with the only change to the model above that the outcome has a Bernoulli distribution with probability of success θ^{sk}.

Remark: Modelling Treatment Effects. The model is agnostic about potential interaction effects and uses dummies for all treatment arms. A common approach to estimating the effects of cross-randomized interventions is to impose additional structure, e.g. by assuming additive effects of the intervention variants T1/T2 and A/B/C. However, note that this imposes constraints across treatment arms that may interfere with efficient learning if the underlying assumptions are incorrect. Conversely, if it is known that the treatment effects have a specific structure, the optimal assignment shares change, as observations from one treatment arm provide information about other arms, and the efficiency properties of algorithms such as exploration sampling are not known in this setting.

Oral Reading Fluency. Our estimation of oral reading fluency uses ORF scores from three periods: the endterm exam of term 2 (E2), and the midterm and endterm exams of term 3 (M3, E3). This means we capture all students pre-treatment, wave-1 students in two periods post treatment, and wave-2 students in one period post treatment (provided their ORF score is not missing). We use a Bayesian approach for consistency and because Bayesian inference is valid even with adaptive sampling.

Let Y_{it}^{sk} denote the ORF score of a student i at time t in school s, assigned to treatment arm k. Define γ_{it}^{sk} as the average ORF score of student i in school s for period $t \in \{E2, M3, E3\}$ and arm $k \in \{0, \ldots, 6\}$, where $k = 0$ now includes the control group. We assume that Y_{it}^{sk} has a normal distribution, and model the

[21] We use this random effects parameterization 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).

average ORF score with a hierarchical linear regression:

$$\begin{aligned} Y_{it}^{sk} \mid \gamma_{it}^{sk} &\sim N(\gamma_{it}^{sk}, \sigma^2) \ , \\ \gamma_{it}^{sk} &= \beta_0 + \beta^F x_t^k + \kappa^F \eta_s^F + \phi\alpha_i + \rho\iota_t \ . \end{aligned} \tag{3}$$

As before, β^F is a 1×6 vector of average treatment effects. The vector x_t^k, $k \in \{1, \dots, 6\}$ is a unit vector that indicates whether the student experienced treatment k in period t or earlier, as in a simple difference-in-difference specification wit time-invariant treatment effects.

The product $\kappa^F \eta_s^F$ is the realization of a school-level random effect, $\phi\alpha_i$ is the realization of a student-level random effect and $\rho\iota_t$ is the realization of a period-level random effect. We use a non-informative improper prior on $\{\beta_k^F\}_{k=0}^6$ and a Half-Normal prior distribution for each one of the random effect variance terms $\{\sigma, \kappa^F, \phi, \rho\}$, and assume a Standard Normal distribution for each of the random effects $\{\eta_s^F, \alpha_i, \iota_t\}$. We have:

$$\begin{aligned} p(\beta_0) &\propto 1 \ , \\ p(\beta_k^F) &\propto 1 \quad \forall k = 1, \dots, 6 \ , \\ \{\sigma, \kappa^F, \phi, \rho\} &\sim \text{Half-Normal}(0,1) \ , \\ \{\eta_s^F, \alpha_i, \iota_t\} &\sim N(0,1) \ . \end{aligned}$$

Remark: Note that, unlike for call engagement, we expect that ORF scores increase over time independently of the intervention, as students' reading ability improves over the course of the school term. The control group helps distinguish the pure time trend, captured by $\rho\iota_t$, from any common effects of the IVR calls on ORF. In pure policy choice experiments with stationary outcomes, a control group is not needed. But sampling a control group and including a period random effect in the model can be useful if the outcome targeted for adaptive sampling is expected to vary over time, even if the treatment effects have the same distribution across waves.

Model Fitness. We conduct standard checks on the distribution of predicted outcomes for the call engagement and the oral reading fluency model to validate whether our models are correctly replicating the characteristics of the observed outcome variable. We also check the sensitivity of our results to different prior distribution specifications. For the call engagement model, we select four different prior distributions for β_k^E (β_k^F for ORF): (i) a normal distribution centered on 0 and variance equal to 100, (ii) a T-Student distribution with 1 degree of freedom, mean 0 and variance equal to 100, (iii) a normal distribution centered on 0 and variance equal to 1, and (iv) a T-Student distribution with 1 degree of freedom, mean 0 and variance equal to 1. Next, we follow the same approach with κ^E (κ^F for ORF) and test the following prior distributions: (i) a half normal distribution with mean 0 and variance equal to 100, (ii) an inverse χ^2 distribution with 1 degree of freedom, and (iii) a half T-Student distribution with 1 degree of freedom, mean 0 and variance equal to 1. In all these cases, the results are not affected by the selection of the prior distribution. Given the large sample size, the likelihood is dominating the prior.

4.2 Treatment Assignment and Exploration Sampling

In wave 2, we want to use the Exploration Sampling algorithm proposed in Kasy and Sautmann (2021a) to assign experimental units to treatment arms. Doing so requires calculating the probability optimal p_t^k after each wave. In the policy choice model in Kasy and Sautmann, the outcome is binary, there are no covariates, and the parameter of interest is simply the arm mean θ^k with a Beta prior. The posteriors used to derive p_t^k therefore have a closed form. Here, we estimate a generalized linear model that allows for a school-specific average call success rate; appropriate if we expect outcomes to vary significantly between clusters (such as schools). However, this implies that the expected outcome in arm k, $\bar{\theta}^k = E_T[\theta^{sk}|k]$, depends on the random effects (note that θ^{sk} is the re-scaled expectation of call engagement Z_i^{sk}). Moreover, we sample the posterior distribution of all parameters using MCMC which requires many numerical draws.

In order to simplify the calculation of p_t^k, we use that $\bar{\theta}_t^k > \bar{\theta}_t^{k'}$ if and only if $\beta_k > \beta_{k'}$. In our model, this is the case since θ^{sk} is strictly increasing in β^E for any realization of the school effect η_s^E or the dispersion parameter κ^E. This implies that

$$\Pr_t(k = \operatorname*{argmax}_{k'} \bar{\theta}_t^k) = \Pr_t(k = \operatorname*{argmax}_{k'} \beta_{k'}), \tag{4}$$

and therefore we can simulate the probability that arm k is optimal using just the posterior of the parameters $\{\beta_k^E\}_{k=1}^6$, rather than the (joint) distribution of all the parameters entering θ_i^k.[22] This shortcut can simplify deriving the exploration sampling assignment shares for many models with covariates or random effects.

Posterior Probability of Successful Engagement and Posterior Expected Regret. At the end of the experiment, we want to implement the arm with the highest average outcome, or equivalently, lowest policy regret. Here, we translate this to choosing the treatment arm with lowest posterior estimated regret in terms of the engagement probability, $E_T[\Delta^k] = E_T[\theta^{s(1)} - \theta^{sk}|k]$ (where the expectations are formed over the posteriors for β and κ and the normally distributed school random effects, and $\theta^{s(1)}$ denotes the

[22] Note that the same approach would also be valid if we had targeted ORF and were to simulate probability optimal base on the $\{\beta_k^F\}_{k=1}^6$.

school-specific success probability under the optimal treatment arm).

The expected probability of a successful engagement $\bar{\theta}^k = E_T[\theta^{sk}|k]$ and the expected regret $E_T[\Delta^k]$ cannot be derived from the distribution of the $\{\beta_k^E\}$ alone because of the non-linear inverse logit transformation $\text{logit}^{-1}(x) = \frac{e^x}{1+e^x}$.[23] We therefore draw from the posterior distributions of κ^E and β^E and the standard normal distribution of η_s^E to calculate the success probability in each arm and school. Then we average over these θ^{sk} draws to obtain $\bar{\theta}^k$ as well as $E_T[\Delta^k]$.

Remark: Predicting probability of success and policy regret with school-level effects. By drawing the school random effects from the normal distribution, we implicitly take an "out-of-sample" approach that ignores the distribution of realized random effects in the student sample. This is informed by the fact that we did not find important differences by school size, such as a correlation between average ORF scores and size. We therefore treat new generations of students as random draws from the distribution of school random effects. An alternative would be to treat the school random effects as persistent and combine the posteriors of the school random effects with assumptions about (future) class sizes to obtain the expected (future) engagement probability and regret. The use of expected regret based on predicted treatment outcomes as the decision criterion requires making explicit what assumptions are used to make predictions.

Remark: Heterogeneity. Relatedly, our approach to calculating p_t^k rests on the monotonicity of the θ^{sk} in β^k. The approach does not apply when "preference reversals" occur. As a simple example, suppose arm k' has a strong effect in some schools and none in others, whereas k'' has a moderate effect in all schools. In this case, it depends on the treatment effect distribution which arm has the highest average treatment effect; here, for example, the size of the different schools. If such heterogeneity is expected, the researcher needs to estimate the distribution of θ_i^k more flexibly, for instance by allowing interactions between covariates and treatment, in which case deriving both the probability optimal p_t^k and expected regret $E_T[\Delta^{k^*}]$ requires assumptions about the covariate distribution in the population. Note also that preference reversals imply that treatment k' is optimal for some schools, whereas for others it is k'', in other words, the unconstrained optimal policy is specific to each school. Targeted policy choice is discussed briefly in Kasy and Sautmann (2021a), and Caria et al. (2020) describe a targeted adaptive experiment using their proposed tempered Thompson algorithm. Targeting has the advantage that we do not need to "trade-off" strata for which different policies are optimal, but it is not always easy to implement in real-world contexts.

[23] Note for example that the estimate of the average success probability, $\bar{\theta}^k = E_T[\text{logit}^{-1}(\beta^E x^k + \kappa^E \eta_s^E)|\beta_k^E]$ is different from both $\text{logit}^{-1}(\hat{\beta}_k^E)$, the inverse logit of the point estimate of the treatment effect, and from $E_T[\text{logit}^{-1}(\beta_k^E)]$, the expected success rate at the median school with $\eta_s^E = 0$.

4.3 Frequentist Inference

As discussed, treatment effect estimates from adaptively collected data are subject to sampling bias, and focusing on the effect in the best arm leads to "winner's curse". Corrections for these sources of bias are rapidly evolving fields of research.

To our knowledge, there is no method yet available to correct for adaptive sampling bias in models with random effects, but there exist weighting approaches for a range of settings that make estimators asymptotically normal (Hadad et al., 2021; Zhang et al., 2021, 2020). In particular, the square root inverse propensity weighting proposed by Zhang et al. (2021) – which in our setting corresponds to weights $\sqrt{\frac{1}{q_t^k}}$ for observations in arm k — applies to m-estimators including Binomial GLM. Using these adaptive weights results in an estimator that is asymptotically normal. In section 5.3, we examine how these weights affects point estimates and confidence intervals compared to an unweighted Binomial GLM estimate.

In addition, Andrews et al. (2021) have developed corrections for the "winner's curse" that arise when estimating the treatment effect in the best arm. We construct confidence intervals with "unconditional coverage," which allow valid inference on the effect of IVR calls on engagement when the best call format is implemented (but regardless which of the six formats that is).[24] These corrections require normally distributed estimates. Following a suggestion by Hadad et al. (2021), we use the adaptively weighted Binomial GLM estimates as inputs into these corrections and show how this changes the point estimates and confidence intervals (section 5.3). These approaches are not directly comparable to the Bayesian estimates with random effects, but they allow us to gain some intuition about how the treatment effect estimates change. Two recent software packages make it easy to apply the "winner's curse" corrections (Shreekumar, 2020; Bowen, 2022).

5 Results of the IVR Experiment

5.1 Call Engagement

Table 1 presents estimates of the treatment effects from Bayesian Binomial GLM models as specified in Equation (2). We show both the estimate with only wave-1 data and with data from both waves. The table reports the means and, in brackets, the 95% highest-probability density (HPD) intervals of the posterior distributions.[25] A higher coefficient is associated with a greater average probability of successful engagement.[26]

[24] One may debate whether conditional or unconditional coverage is appropriate. In an experiment that compares different types of interventions – say, conditional cash transfers and IVR calls – we may be interested in the effect of IVR calls only if they yield better outcomes than the cash transfer. We see this as a case of conditional inference, because the identity of the best arm matters.

[25] The 95%-HPD region H is defined by the highest k such that $\int_l^u f(\theta)d\theta = 95\%$ and $f(\theta) \geq k$ for all $\theta \in H$, where f denotes the posterior pdf of θ. For unimodal distributions, H is an interval.

[26] Recall that, for a point estimate for the treatment effect β_k^E and the median school with random effect 0, we would have that the probability of success in arm k equals $\theta^k = \frac{\exp{(\beta_k^E)}}{1+\exp{(\beta_k^E)}}$.

Table 1: Call engagement estimates after wave 1 and 2.

	Bayesian Binomial GLM	
	Wave 1	Full sample
	(1)	(2)
T1A	-2.84^{*}	-2.63^{*}
	$[-3.09; -2.60]$	$[-2.81; -2.46]$
T1B	-2.64^{*}	-2.49^{*}
	$[-2.87; -2.42]$	$[-2.63; -2.36]$
T1C	-2.75^{*}	-2.78^{*}
	$[-3.00; -2.52]$	$[-2.93; -2.63]$
T2A	-2.94^{*}	-2.89^{*}
	$[-3.19; -2.70]$	$[-3.11; -2.68]$
T2B	-2.83^{*}	-2.67^{*}
	$[-3.08; -2.60]$	$[-2.85; -2.50]$
T2C	-3.46^{*}	-3.32^{*}
	$[-3.74; -3.20]$	$[-3.57; -3.07]$
Num. students	1283	2462
Period	1	1 and 2

Notes: * Value of zero lies outside of the 95% credible interval. We simulate 4 independent Markov chains of 4,000 posterior draws each and discard the first 2,000 as warm up. The remaining 8,000 draws are used to generate the posterior distributions of the coefficients. The Split-$\hat{R}$ of every posterior distribution is below 1.01 and there are no divergent transitions.

Table 2: Treatment allocation in waves 1 and 2.

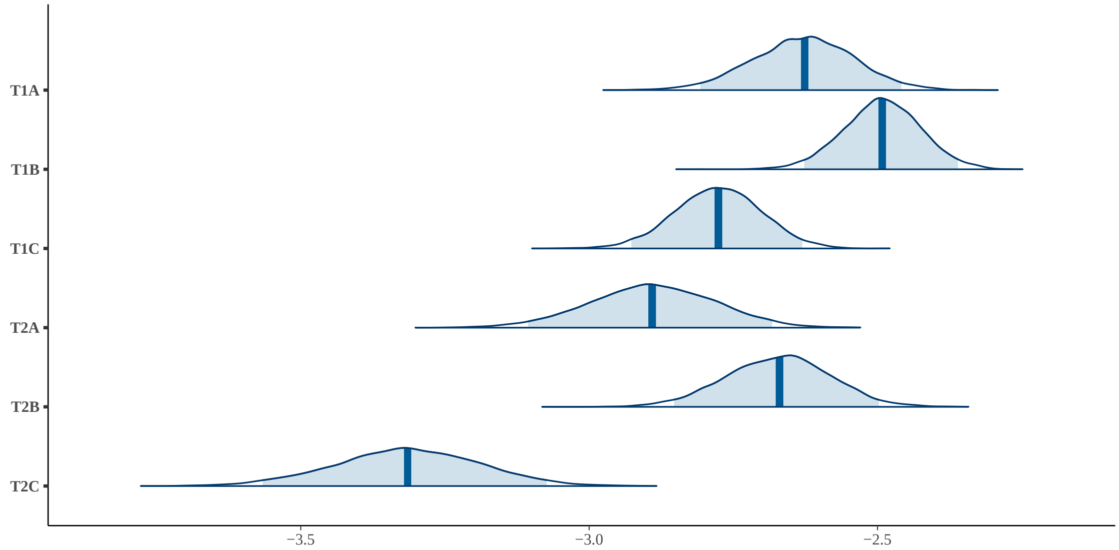
	Wave 1			Wave 2		
Treatment	Target %	Actual %	Num. students	Target %	Actual %	Num. students
T1A	14.28%	14.12%	211	7.44%	8.45%	117
T1B	14.28%	14.73%	220	39.26%	40.46%	560
T1C	14.28%	13.86%	207	28.45%	26.81%	371
T2A	14.28%	14.32%	214	0.89%	1.01%	14
T2B	14.28%	13.72%	205	9.68%	8.53%	118
T2C	14.28%	15.13%	226	0.00%	0.00%	0
Control	14.28%	14.12%	211	14.29%	14.74%	204

Notes: treatment arm sample allocation on waves 1 and 2. Target % shows the target theoretical shares of each treatment arm. Observed % shows the actual treatment allocation after randomization with stratification. Num. students is the number of students in each treatment arm.

The estimates from wave 1 in Table 1 were used to determine the exploration sampling shares for wave 2. Table 2 shows the theoretical sample shares in each treatment group, as well as the assigned sample shares after stratifying by school, both for wave 1 and wave 2. Exploration sampling reduced the sampling share assigned to treatments T2A and T2C to zero or almost zero. Moreover, T1A and T2B received only slightly over 8% of the sample. The bulk of the allocation went to T1B and T1C (aside from the control). These are both calls where the IVR instructs the parent to lead reading exercises, but in B the same intermediate exercise sequencing is used for all, whereas in C the parent can choose the exercises.

Column (1) in Table 1 shows that some differences in treatment effects already emerged in wave 1, which led to the differences in treatment assignment in wave 2. The full sample estimate in column (2) both shows slightly different point estimates and significantly tighter HPD intervals, especially for the higher-performing treatments. Figure 3 displays the treatment effect posterior distributions after wave 2, corresponding to the estimate in column (2) of Table 1. The shape of the distributions shows that the higher treatment effects are estimated with significantly greater precision. This allows a finer distinction between T1A, T1B, and T2B. After wave 2, T1B is the treatment arm with the highest level of engagement, with a point estimate of $\hat{\beta}^E_{T1B} = -2.49$, whereas T2C has the lowest engagement with $\hat{\beta}^E_{T2C} = -3.32$.

Table 3 provides additional information. Columns (1) and (2) show the raw numbers of attempted engagements and share of successful engagements (dividing the number of successful calls by the number of call attempts). Columns (3) to (5) are based on the posterior of the treatment effect vector β^E. The mean and standard deviation in each arm replicate the estimation results in column (2) of Table 1 and show once more that higher means are associated with lower dispersion of the estimate. Column (5) shows the probability optimal p^k_2 for each arm k. The posterior probability that T1B is the optimal choice is over 93%; three



Notes: the figure shows the posterior distribution of parent engagement coefficients after wave 2. Greater values are associated with a higher probability of a successful engagement. The vertical bar marks the median of each posterior distribution. The shaded areas indicate the 95% credible intervals. A total of 8,000 posterior draws sampled from 4 independent Markov chains were used.

Figure 3: Posterior distributions of parent engagement coefficients.

arms (T1C, T2A, and T2C) have essentially zero posterior probability that they deliver the highest level of engagement. Arm T1A, parent-led reading with leveled exercises, has the second highest engagement rate of 7.43%, but has only a 5.24% probability optimal.

The last two columns transform the posterior estimates into an average probability of successful engagement for each arm, $\bar{\theta}^k$, and report the expected policy regret based on the probability of engagement, the objective of interest (see section 4). This statistic shows that implementing T1B would lead to an expected loss in terms of the probability of a successful call of only 0.02 percentage points. For the other treatment arms, the loss ranges between 0.99pp and 4.49pp. These expected losses are equivalent to less than 1%, 12%, and 53% of the highest estimated success probability in arm T1B (of 8.40%).

In order to look more into parents' decision to answer the biweekly IVR calls, we also analyze the extensive margin of engagement. Appendix C.4 shows estimates for the probability of any successful engagement (i.e., whether the recipient started the reading exercises in any of the calls received) and the probability of answering the phone at least once. Tables A.6 and A.7 report the coefficient estimates and the corresponding treatment arm averages. The arms had nearly identical initial response rates: in five arms at least one call was answered with 84.1%-86.6% probability, and the response rate was only slightly lower in T2A (81.7%). The share of phone numbers with at least one successful engagement varies somewhat more across arms, and is particularly low in T2C, where the rate is only about half of what it is in other arms. However, T1A, T1B and T2B have nearly identical engagement probabilities. It is instructive to also compare the

Table 3: Call engagement: treatment effect estimates after wave 2.

	Raw numbers		Posteriors of β^E			Average engagement $\bar{\theta}^k$	
Arm	Call attempts (1)	Share successful (2)	Mean (3)	SD (4)	Prob. optimal p_t^k (5)	Success prob. $\bar{\theta}^k$ (6)	Post. exp. policy regret $E_T(\Delta^k)$ (7)
T1A	$2,952$	7.28%	-2.63	0.09	5.24%	7.43%	0.99%
T1B	$7,020$	8.40%	-2.49	0.07	93.19%	8.40%	0.02%
T1C	$5,193$	6.47%	-2.78	0.08	0.00%	6.49%	1.93%
T2A	$2,052$	5.95%	-2.89	0.11	0.00%	5.86%	2.56%
T2B	$2,907$	7.05%	-2.67	0.09	1.57%	7.15%	1.27%
T2C	$2,034$	3.98%	-3.32	0.13	0.00%	3.93%	4.49%

Notes: (1) A call attempt is a scheduled call to a parent, 9 per wave (not counting repeated attempts and call backs). (2) The share successful is the percentage of call attempts in which the exercises were started. (3-4) The posterior mean and standard deviation of β^E were calculated from a total of 8,000 posterior draws sampled from 4 independent Markov chains. (5) The probability optimal is calculated as in Eq. 4. (6) The average probability of success is calculated as in Eq. **??**. (7) The posterior expected policy regret is the expected loss from choosing this arm, expressed in terms of the probability of a successful call, after observing both waves of the experiment.

"probability highest" for each arm based on the extensive margin estimates, reported in column (2) of Table A.7. These probabilities are the analog of the probability optimal in column (5) of Table 3 (as these can be calculated for any outcome in any experiment, regardless whether adaptive sampling was used). These probabilities never exceed 35.3%. The point estimates and probability highest indicate that the six call formats are much less clearly differentiated based on the probability of "any engagement" than based on the overall call engagement rate. This suggests that the intensive margin matters, and differences in response rates emerge more clearly as parents learn about the calls and decide about continued engagement.

One interpretation of the results, comparing A and B arms, is that leveling exercise content in this setting is not valuable – perhaps because of the noisy and often missing ORF scores used for leveling – or at least not valued by parents, who may perceive the exercises as too easy or too difficult. Both C arms have relatively low call engagement rates. It is worth noting that the option to choose between exercises increases the length of the call, which may discourage the listener. The call success rate in T2C is particularly low, and we conjecture that this is because the listener is not only asked to choose which exercises to play, but the IVR here also addresses the child directly. This "gamification" aspect may lead the parent to worry about overly long calls in which the child skips around between exercises. Between T1 and T2, the posterior means suggest that T1 arms have slightly higher engagement rates, perhaps because the "listen now, practice later" format allows the parent more flexibility.

The sampling shares in Table 2 and the numbers of attempted and successful engagements in Table 3 also demonstrate a property of adaptive sampling that is attractive in the context of policy choice: the reassignment of treatment arm shares in later waves means that a larger percentage of participants benefit from the treatment arms with better outcomes. Here, this means more students get IVR calls with high engagement levels. At the end of this experiment, 27.10% percent of students participated in T1B compared to only 7.85% in arm T2C.

5.2 Oral Reading Fluency

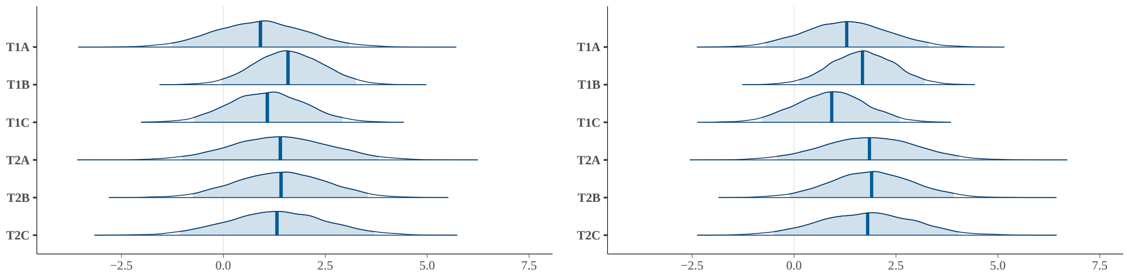
Even though the adaptive sampling algorithm was geared towards learning about call engagement, we would also like to estimate treatment effects on reading fluency. ORF may increase directly if parents regularly carry out the actual exercises delivered with their children, improving their reading. The calls may also increase parents' awareness of their child's reading ability more generally, leading them to express interest and encourage reading practice in day-to-day interactions.

Table 4 presents estimates from two different samples. Column (1) in both panels shows the estimated treatment effects on ORF scores using only the sample of students with complete score information in all three exams, whereas column (2) uses all students for whom we have at least one treated and one untreated exam score. Figure 4 shows the posterior distributions of the ORF coefficients corresponding to Panel A of Table 4, panel (a) for the balanced panel data and panel (b) for the unbalanced panel data.

In both samples, the ORF treatment effects shown in Panel A are small and estimated noisily, ranging from 0.90 to 1.90 correct words per minute. By comparison, in the control group, ORF increased on average by 1.62 cwpm and 2.92 cwpm in the first and second half of the term, respectively. Overall, going from column (1) to column (2), the treatment effects tend to be estimated larger, although with similar credible intervals; despite the much larger sample in the unbalanced panel, the precision of the estimates does not increase much, perhaps due to the student-level random effects. In the balanced panel, the credible interval for all six coefficients includes 0. However, the unbalanced panel estimate for the arm with the highest call engagement, T1B, indicates an increase in fluency by 1.68 cwpm, and the credible interval does not include zero. Note that T1B has a relatively large share of the sample because of the use of adaptive sampling, and therefore the effect on fluency is more precisely estimated in this arm than in others, even though the mean estimated ORF effects are slightly larger in some other arms. The larger sample size in the treatment arms chosen for implementation is an advantage of adaptive sampling for the estimation of non-targeted outcomes.

In order to test whether simply receiving any calls has an effect on fluency, we pool the six treatment groups in Panel B. In both samples, the HPD intervals do not include 0, and the effect is 1.31 cwpm in the balanced panel and 1.53 cwpm in the unbalanced panel.

It is worth emphasizing once more that the fluency estimates are only indicative, because of the low data quality and because the effects of any treatment would have likely been incompletely captured due



(a) Balanced panel

(b) Unbalanced panel

Notes: the figures present the posterior distribution of treatment effects after wave 2. The vertical bar marks the median of each posterior distribution. The shaded areas indicate the 95% credible intervals. A total of 8,000 posterior draws sampled from 4 independent Markov chains were used.

Figure 4: Posterior distributions of treatment effects for ORF scores.

to the short exposure and the one-off measurement of ORF immediately after treatment. That said, the estimates suggest that an IVR intervention for parental engagement in their children's reading will have a positive impact on children's reading skills. This is an encouraging finding given the relatively "light touch" of this intervention. Based on the estimates from the unbalanced panel, it is more than 95% likely that implementing the arm with the highest engagement, T1B, which asks parents to carry out a few simple reading exercises sequenced the same for all children, will lead to positive reading fluency gains. While the effects of the 4.5-week intervention tested here were moderate, it stands to reason that exposure for the full term or even the full school year will generate larger effects. The program may also lead to continued joint reading between parents and children after the calls end.

A remaining question is how the treatment effects on fluency compare between the different arms and whether call exposure and efficacy vary significantly strongly so that one of the arms with lower call engagement could be more effective for reading outcomes. Unfortunately, the answer is hampered by the quality of the data and the relatively small effect sizes. From Figure 4, there is significant overlap in the credible intervals of all arms, even for the treatment arms with a large share of observations. To get a sense of the uncertainty, Table A.4 in Appendix C shows the "probability highest" and the expected regret for each arm based on the posterior distributions *of the ORF model*. The probability that T1B leads to the highest possible reading gains among the six arms lies between 12% and 20% according to these estimates. T1B generates a posterior regret of 0.94 cwpm in the balanced panel and 1.14 in in the unbalanced panel. In the balanced panel, T1B is the arm with the lowest posterior regret. In the unbalanced panel, arm T2B has the lowest posterior regret, with 0.92 cwpm. While the probability optimal is higher than for T1B for three arms (T2A, T2B, and T2C), it is below 26.3% for all of them, and the difference in expected regret is less than 0.24 cwpm.

The low probability optimal for the arms with lowest regret reflects the noise in these estimates. Note also

Table 4: ORF scores estimates.

Panel A: Treatment effects

	Balanced Panel (1)	Unbalanced Panel (2)
(Constant)	46.90*	46.54*
	[43.98; 49.92]	[43.76; 49.31]
T1A	0.90	1.29
	[−1.26; 3.09]	[−0.70; 3.30]
T1B	1.60	1.68*
	[−0.04; 3.26]	[0.13; 3.21]
T1C	1.08	0.91
	[−0.72; 2.92]	[−0.79; 2.59]
T2A	1.40	1.85
	[−1.01; 3.81]	[−0.42; 4.04]
T2B	1.41	1.90
	[−0.75; 3.55]	[−0.12; 3.92]
T2C	1.32	1.79
	[−1.06; 3.71]	[−0.49; 4.05]

Panel B: Pooled treatment effects

	Balanced Panel (1)	Unbalanced Panel (2)
(Constant)	46.91*	46.63*
	[43.87; 50.02]	[43.77; 49.49]
Pooled treatment	1.31*	1.53*
	[0.08; 2.52]	[0.34; 2.69]
Num. obs.	5469	6701
Num. students	1823	2439

Notes: Reporting means and 95% HPD intervals (in square brackets) of the posterior distributions of treatment effects. *: zero outside 95% credible interval. We simulate 4 independent Markov chains of 4,000 posterior draws each and discard the first 2,000 as warmup. The remaining 8,000 draws are used to generate the posterior distributions of the coefficients. The Split-$\hat{R}$ of every coefficient is below 1.01 and there are no divergent transitions.

that T2A has a higher probability optimal than T2B in both the balanced and unbalanced panel, highlighting that the arm with the highest probability optimal may not always have the lowest policy regret. This can occur if some "unlikely" states of the world have very high regret realizations and occurs more often when the best arm is fairly uncertain.

Overall, based on these results there is significant uncertainty about which arm has the highest ORF gains. There is no strong evidence that choosing a policy based on maximal call engagement is systematically at a tension with also increasing oral reading fluency, but we can also not conclude that the two outcomes are definitely aligned. If the implementer would like to revise the decision to target engagement only and learn which call format maximizes ORF gains, additional testing would likely be needed.

5.3 Correcting for Sampling Bias and Winner's Curse

While most of our analysis is Bayesian, researchers may also be interested in conducting frequentist inference with the data obtained from a policy choice experiment to draw broader conclusions about the interventions tested, and this requires correcting sampling and winner's curse biases.

Table 5: Call engagement estimates applying the adaptively weighted m-estimator by Zhang et al. (2021) and the "winner's curse" correction by Andrews et al. (2021).

Panel A: Binomial model estimates, unweighted and with adaptive weighting.

	Unweighted With school RE (1)	Unweighted Without school RE (2)	Adaptively weighted Without school RE (3)
T1A	-2.63^*	-2.54^*	-2.52^*
	$[-2.80; -2.45]$	$[-2.79; -2.3]$	$[-2.78; -2.27]$
T1B	-2.49^*	-2.39^*	-2.39^*
	$[-2.62; -2.36]$	$[-2.54; -2.24]$	$[-2.55; -2.24]$
T1C	-2.77^*	-2.67^*	-2.66^*
	$[-2.92; -2.62]$	$[-2.86; -2.49]$	$[-2.85; -2.46]$
T2A	-2.88^*	-2.76^*	-2.79^*
	$[-3.09; -2.67]$	$[-3.09; -2.43]$	$[-3.20; -2.39]$
T2B	-2.67^*	-2.58^*	-2.57^*
	$[-2.84; -2.49]$	$[-2.82; -2.34]$	$[-2.81; -2.33]$
T2C	-3.31^*	-3.18^*	-3.18^*
	$[-3.55; -3.06]$	$[-3.59; -2.77]$	$[-3.59; -2.77]$
Num. students	2462	2462	2462
School RE	Yes	No	No

Panel B: "Inference on winners" correction on T1B.

	With school RE (1)	Without school RE (2)	Re-weighted (3)
T1B	-2.49^*	-2.39^*	-2.39^*
	$[-2.66; -2.32]$	$[-2.59; -2.19]$	$[-2.60; -2.18]$

Notes: *Value of zero lies outside of the 95% confidence interval. (1) Frequentist estimate, unweighted and with school random effects as in the original model specification (Table 1, Column 2). (2) Frequentist estimate without school random effects. (3) Frequentist estimate without random effects, applying adaptive weights as in Zhang et al. (2021). Panel A: full estimates for all treatment groups. Panel B: Median estimate and adjusted confidence intervals for T1B, applying corrections for inference on the best arm as in Andrews et al. (2021). Note that this correction is only theoretically valid in column (3) where the underlying estimator is asymptotically normal.

As discussed in section 4.3, a method to correct for the biases that arise from adaptive sampling when there are random effects does to our knowledge not yet exist. We therefore present results without random effects for illustrative purposes. In Table 5, we show a set of frequentist estimates that iteratively apply adaptive weighting and the winner's curse correction. In column (1), we show unweighted estimates from a Binomial model with random effects. These are the frequentist equivalent to the Bayesian estimates in column (2) of Table 1 (and they are very similar).

Column (2) shows unweighted estimates again, but this time without random effects. As is common, this shifts the estimated coefficients somewhat towards 0. In Column (3), we apply the adaptive weights proposed by Zhang et al. (2021) to obtain asymptotically normal estimates. It is instructive to compare columns (2) and (3) in Panel A: for the best arm, the estimates are almost identical, whereas for example for T2A the point estimate is shifted and the confidence interval significantly wider. This reflects that arms who initially perform poorly receive only a small share of the sample, and the weighted estimator therefore gives those few observations significantly greater weights, with the potential to change the overall treatment effect estimate. As Hadad et al. (2021) observe, this is an indirect consequence of the fact that sampling bias primarily affects the sub-optimal arms (which are "dropped" from the sample) rather than the optimal arm, where initial biases have a chance to self-correct.

In Panel B, we apply the winner's curse correction by Andrews et al. (2021) to the treatment effect estimate for the empirically best arm, T1B. Note that the method requires normally distributed estimators, so it is strictly speaking only applicable with the weighted estimates in column (3). However, for illustration purposes we carry out the same correction in all columns. The corrected confidence intervals we obtain are somewhat wider than the "naïve" estimates in Panel A. However, the point estimates for the treatment effects remain virtually the same. This reflects that at least in the IVR experiment the best arm is fairly unambiguously identified, and the distribution of the estimator is therefore not significantly truncated. This means also that a winner's curse is less likely. As Andrews et al. (2021) also point out, uncorrected frequentist estimates are asymptotically valid.

We may deduce that we need not be too worried about taking the Bayesian treatment effect estimates for the IVR experiment at face value. However, in experiments with smaller samples, both sampling biases and the winner's curse problem may be more pronounced.

6 Alternative Research Designs

In this section, we turn back to the question of how to choose the research design. Potential users of exploration sampling and adaptive experiments more generally will be interested in the learning gains from adaptivity, as well as the best design for their adaptive experiment.

A first question is whether adaptive sampling improved learning in the IVR experiment. The motivation to use adaptive methods is to increase efficiency and make the most of a limited sample and time. However, asymptotic convergence results for exploration sampling and other best-arm algorithms (Kasy and Sautmann, 2021a; Russo, 2020; Qin et al., 2017) only apply to specific outcome distributions and when the number of waves grows large. In this experiment, we learn only from one prior wave and adapt the assignment shares for half of the sample in a second wave. Possible learning gains from adaptivity are further limited by the fact that the exploration sampling algorithm can only approximate the optimal assignment. In a first exercise below we therefore use simulations to evaluate the gains from adaptive sampling in wave 2, relative to non-adaptive sampling where an equal share of the sample is allocated to each treatment arm (a "standard RCT"). We use the data actually gathered in this experiment. The goal is to quantify the performance of exploration sampling ex post and for the specific context of the IVR experiment. This contributes to an evidence base about the gains from adaptive sampling in policy choice.

A second question is how researchers should ex ante compare and make decisions about research designs based on prior information, and whether such comparisons are reliable. As discussed above, operational and logistical constraints in the IVR restricted the space of possible research designs essentially to either conducting one experimental wave (possibly in only one half of the term) or two waves. With reference to the two scenarios laid out in 2, ex ante, we might have asked whether we should simply conduct a one-wave, non-adaptive RCT with the full sample, or if there are significant gains from holding back half of the sample and adjusting the treatment assignment using exploration sampling in wave 2. Alternatively, after carrying out wave 1 and observing the results, we might have asked whether the learning gains from the second wave make the effort worthwhile. In the second and third exercise below, we therefore carry out simulations that answer these questions, in the same way an experimenter might have done to make decisions about the IVR experiment. These simulations are by necessity not based on the actual data collected, but on the Bayesian model and parameter distributions we specified. The purpose is both to compare the predicted gains from adaptive sampling obtained ex ante from the model with those obtained ex post from the data, and to illustrate how one might go about conducting such simulations.

6.1 Ex Post Counterfactual: Non-Adaptive Experiment

In a first exercise, we ask what expected regret and probability optimal in the experiment might have been if we had carried out a "standard RCT", that is, an experiment with uniform assignment shares. Since the assignment shares in wave 1 were equal, we simulate learning outcomes from a large number of bootstrapped samples for wave 2, drawn from real experimental observations in wave 1 and 2.

All our bootstrap samples for wave 2 have $N = 1384$ observations, the draws are stratified by school,

Table 6: Ex post counterfactual: performance of exploration sampling and standard RCT

	Exploration Sampling				Standard RCT			
Treatment	Success prob. mean (1)	Success prob. SD (2)	Prob. treat optimal (3)	Posterior exp. policy regret (4)	Success prob. mean (5)	Success prob. SD (6)	Prob. treat optimal (7)	Posterior exp. policy regret (8)
T1A	6.96%	4.45%	4.33%	1.56%	7.28%	5.29%	6.14%	1.38%
T1B	8.50%	5.24%	91.93%	0.03%	8.59%	6.06%	85.13%	0.06%
T1C	6.89%	4.39%	1.59%	1.64%	7.27%	5.28%	6.40%	1.39%
T2A	6.12%	3.99%	0.11%	2.41%	5.92%	4.44%	0.05%	2.74%
T2B	6.77%	4.34%	2.03%	1.76%	6.92%	5.07%	2.28%	1.73%
T2C	3.89%	2.67%	0.00%	4.64%	4.03%	3.17%	0.00%	4.63%
Selected	8.50%	5.25%	92.58%	0.02%	8.60%	6.07%	86.39%	0.05%

Notes: The table shows averages of estimates for each treatment arm obtained from 1,000 simulated samples drawn from the observed experimental data. Columns (1) and (5): mean posterior probability of a successful call. Columns (2) and (6): standard deviation of the posterior success probability. Columns (3) and (7): probability that the treatment arm is optimal. Columns (4) and (8): posterior policy regret in terms of engagement success probability.

and we append the bootstrapped sample to the observed wave 1 data to estimate a hierarchical Bayesian Binomial GLM as described in Eq. 2. We carry out 1,000 draws that simulate an RCT and 1,000 draws that simulate an exploration sampling experiment. For the simulated RCTs, we bootstrap a wave 2 of equal-sized treatment arms. For the simulated exploration sampling experiments, we use the treatment assignment shares derived from the original wave 1 posterior distributions.[27] For each sample draw, we calculate the posterior mean and standard deviation of $\bar{\theta}^k$, the probability optimal, and the posterior expected policy regret for each arm. The averages for each arm across draws are shown in in Table 6. In addition, we show the average of the posterior regret and probability optimal of the selected (lowest-regret) arm k^* in each simulated experiment.

The average posterior mean of the probability of a successful call is similar between exploration sampling and standard RCT, as seen in columns (1) and (5). As expected, the standard deviation of the posterior distribution of the mean success probability $\bar{\theta}^k$ is lower under exploration sampling for the high-performing treatments, but higher for the low-performing arms. In both research designs, the treatment arm that is most often associated with the highest probability of engagement is T1B. However, in the exploration sampling experiment, T1B is chosen 97.4% of the time, whereas this is the case 94.9% of the time in the simulated RCTs. This reflects the greater uncertainty and consequently higher variance in the final decision that results

[27] This exercise is not perfect, because we re-sample from the six arms at different proportions for the two designs. Since we use data from both waves, the bootstrapped wave-2 sample is always smaller than the original sample we draw from. However, the probability of repeat draws is affected by both the size of the original arm and the target size, and this ratio varies across the two designs. An alternative approach is to use a randomly drawn sub-sample of the original data that is proportional to the targeted wave size. This equalizes the chance of repeat sampling across arms, but it implies that the two bootstraps draw from different underlying populations. Ultimately this second drawback seemed more problematic than the first.

from a non-adaptive experiment.

Exploration sampling increases the probability optimal of the best arm on average from 86.39% to 92.58% and reduces the average posterior regret from 0.05% to 0.02%. The reduction is small in absolute terms for two reasons; first, the student sample is large enough so that even an RCT would lead the researcher to relatively firm conclusions here, and second, in this particular problem instance it turns out that the arm averages are clustered closely together, meaning that even a suboptimal choice is likely to be benign. However, in relative terms the improvement is large, and in a policy choice problem where the best arm is actually implemented, even small per-unit gains in payoffs may accumulate into large welfare differences. Overall, the ex-post simulations suggest that we can achieve a meaningful decrease in uncertainty and improved decisions from just one wave with adaptive sampling involving half of the experimental sample.

Remark: Decision Metrics. These simulations highlight an advantage of the proposed Bayesian approach: the metrics of expected policy regret and probability optimal provide the decision maker with easy to understand, intuitive measures of the uncertainty attached to the policy choices they are making. This facilitates the comparison of treatment arms as well as experimental research designs.

6.2 Ex Ante Comparison: Model-Based Simulation of Exploration Sampling vs. RCT

In the second exercise, we imagine the experimenter asking before the IVR experiment, "should I carry out one (non-adaptive) wave with the whole sample, or two (adaptive) waves with half the sample each?" For these simulations, take a given parameter vector (β^E, κ^E). Based on this vector and Eq. (2), we can simulate outcomes Z_i^{sk} for the students in each wave (drawing the school effects η_s^E from the Standard Normal distribution). The first simulated sample uses equal assignment shares, the second is generated under an adaptive design, where the assignment shares for wave 2 are obtained from estimating our model above from the simulated wave 1 data. We can then compare the estimation results under these two sampling strategies to calculate the predicted gains from the adaptive vs. the non-adaptive design for the given parameter vector. This is reminiscent of conducting power calculations for an assumed effect size.

Panel A of Table 7 shows the result of such an exercise, using as the parameter vector the mean of the posterior distributions of β^E and κ^E after wave 2, as reported in Table 3. Using the wave-2 estimates from the experiment serves to show how well the ex ante simulation does in predicting these estimates, and how ex ante simulation results compare with the ex post simulation above. The predicted gains from using adaptive sampling in terms of posterior regret are very similar to our previous exercise based on the actual IVR data. The average posterior expected regret from arm T1B is 0.02% with adaptive sampling but 0.08% with the "standard RCT" on average. The average posterior probability optimal for both sampling strategies is also similar to what we obtained in Table 6.

Table 7: Ex ante comparison: performance of exploration sampling and standard RCT in simulated samples based on parameter vector $(\hat{\beta}^E, \hat{\kappa}^E)$.

Panel A: Averages of Posterior Estimates.

	Exploration Sampling		Standard RCT	
Treatment	Avg. posterior expected policy regret	Avg. posterior probability optimal	Avg. posterior expected policy regret	Avg. posterior probability optimal
T1A	1.18%	4.39%	1.08%	12.61%
T1B	0.02%	92.2%	0.08%	81.29%
T1C	2.10%	0.45%	2.01%	0.48%
T2A	2.67%	0.07%	2.66%	0.03%
T2B	1.52%	2.89%	1.38%	5.60%
T2C	4.51%	0.00%	4.61%	0.00%

Panel B: Average Realized Values.

Exploration Sampling		Standard RCT	
Average policy regret	Percentage best arm identified	Average policy regret	Percentage best arm identified
0.01%	99.00%	0.07%	93.00%

Notes: The table shows averages from 100 simulated samples drawn using the parameter vector given by the means of the estimated posteriors from wave 2 of the IVR experiment, $\hat{\beta}^E = (-2.63, -2.49, -2.78, -2.89, -2.67, -3.32)$ and $\hat{\kappa}^E = 0.5$. For each sample draw, the same first wave was used, the second wave was drawn either using the exploration sampling shares based on the estimates from the first wave, or using equal assignment shares.

Panel B of Table 7 uses the fact that we know the parameter vector that generated the simulated samples, and therefore know the policy regret from choosing a different arm from T1B. This means we can calculate the average policy regret and share of optimal decisions from making the final choice after each simulated experiment (which is based on posterior policy regret). According to panel B, in 99% of the time (93% in the RCT) the experimenter correctly chooses T1B based on this decision metric. The average posterior regret from T1B is only slightly higher than the realized average policy regret;[28] both show an 0.06% reduction in regret from adaptive over non-adaptive sampling. Panel B shows the decision metric that should be used to choose between the adaptive and the non-adaptive design (Panel A shows the expected value of the posterior estimates after the experiment). The posterior estimates show some remaining uncertainty. This is partly due to the school random effects: some of the measurement effort is spent on estimating the school averages, which adds uncertainty to the final estimates.

Prior to an experiment, the researcher of course does not know what the true parameters are, and they

[28] Note that regret in Panel B only occurs when the experimenter does *not* choose T1B.

may want to carry out the calculation in Panel B of Table 7 for multiple parameter vectors in order to get a sense of the distribution of gains from adaptivity. The most consistent approach would be to draw many values from the prior distribution of the model parameters, but this can give a misleading picture of the gains from adaptivity when uninformative priors are used (not to mention that the computational cost is high). As an example, in the IVR experiment, the flat priors combined with the logit transformation in the model mean that treatment arm averages based on random draws from the distributions of the β^E are almost always close to 0 or 1. In Appendix C.3, we therefore show results from a modified exercise in which we independently and randomly draw the θ^k from the uniform distribution on $[0, 1]$. As it turns out, this exercise is not meaningful either: in many cases, the drawn parameters are so far apart that, given our large sample of students, even equal assignment shares lead to a very high probability of picking the correct arm. A more meaningful approach might be to assume correlated prior distributions for the β^E_k, or use the prior same distribution for each θ^k but with a mean obtained from pilot data. An alternative to drawing from a prior distribution is to examine learning gains for a few well-chosen parameter vectors. Again, this is similar to the approach taken in typical power calculations for experiments, see (e.g. Duflo et al., 2007).

Remark: When is Adaptive Sampling Most Valuable? As the simulations show, the efficiency gains from adaptive sampling vary significantly across different problem instances. For best-arm identification, closely clustered treatment effect averages make the problem "hard," as it is difficult to distinguish these arms. From a welfare-maximization (regret minimization) perspective, however, two or more treatment arms with very similar success rates may often lead to a sub-optimal choice, but the loss from that choice will be small. Intuition suggests that adaptive sampling is particularly valuable when there are two or more "near-optimal" arms but also several "far from optimal" arms that can be quickly ruled out. An example could be an experiment that compares two or more different types of interventions but also tests several variants within each type. It will be fruitful to explore these questions in more detail.

6.3 Comparison after Wave 1: Model-Based Simulation of a Second Adaptive Wave

In our last exercise, we imagine the experimenter, after having carried out wave 1, asking, "should I conduct a second adaptive wave?" This is somewhat less computationally costly than the above exercise because after wave 1, the exploration sampling shares for wave 2 are known. As before, for a given parameter vector β^E and κ^E, we simulate a second wave of the experiment by generating a random sample of size $N = 1384$ following the model in Eq. (2) and using the assignment shares in Table 2. We draw 200 parameter vectors from the wave-1 posteriors and calculate average policy regret and percentage of times the best arm is identified for each.

Panel A of Table 8 shows the posterior expected regret and probability optimal for arm T1B after wave

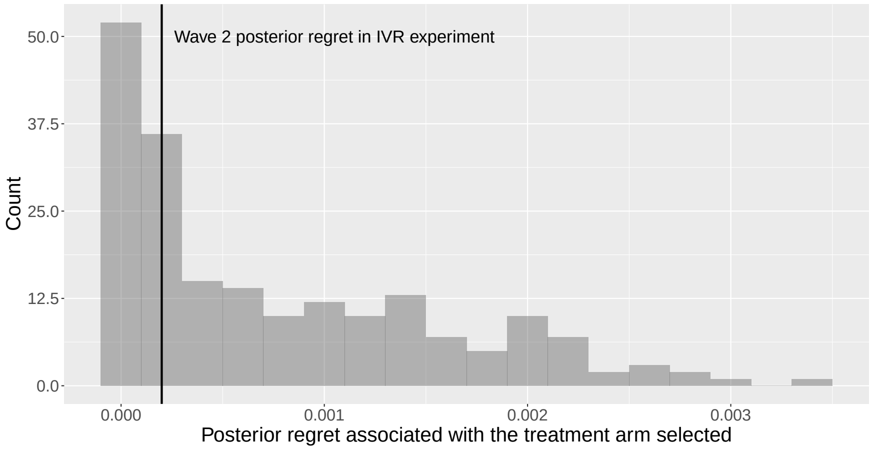


Figure 5: Distribution of the posterior expected regret from wave 1 data and 200 simulated samples for wave 2, based on β^E, κ^E and η^E drawn from their posterior distributions after wave 1.

1. The posterior expected regret at $t = 1$ would be the basis for decision making if no other wave was conducted, and T1B was the arm with the lowest expected regret at that point. Panels B and C show the results of the simulations of wave 2. Panel B shows the average and median posterior policy regret and probability optimal of the chosen arm. On average, the simulation predicts an improvement in expected regret from continued experimentation of 0.04%, and an increase in the probability optimal for the chosen arm from 74.14% to 77.86%. Using the median of the distribution, the improvement would be 0.07% and to a probability optimal of 83.01%. Note that the average posterior expected regret has a heavily skewed distribution. The actual value of 0.02% observed after the second wave in the IVR experiment is at the 38th percentile of that distribution, as seen in Figure 5.

Panel C shows the average realized policy regret and the percentage of times the best arm is identified, both after wave 1 (where the experimenter would have chosen arm T1B) and after wave 2. Comparing the numbers for wave 1 with Panel A shows that the distribution of the simulated draws replicates the theoretical posteriors, as expected. The numbers for wave 2 show realized gains that more than halve the predicted policy regret of wave 1 and increase the share of optimal decisions by 10%. In the actual IVR experiment, had we conducted these calculations between waves, we would have likely concluded that the low monetary cost of sending the IVR calls to the second half of the sample would have more than justified the gains in certainty about the optimal choice. The actual IVR experiment performed even better than these simulations predict. A better prior for our parameters, for example based on pilot data, is likely to generate more reliable answers to research design questions.

Table 8: Comparison after wave 1: ending the experiment vs. conducting a second wave.

Panel A: Posterior Estimates after Wave 1.		
	Exploration Sampling	
Treatment	Posterior expected policy regret	Posterior probability optimal
T1B	0.12%	74.14%

Panel B: Posterior Estimates after Wave 2.		
	Exploration Sampling	
Wave	Avg. posterior expected policy regret [median]	Avg. posterior probability optimal [median]
1 and 2	0.08% [0.05%]	77.86% [83.01%]

Panel C: Average Realized Values.		
	Exploration Sampling	
Wave	Average policy regret	Percentage best arm identified
1	0.13%	71.00%
1 and 2	0.06%	81.00%

Notes: The table summarizes the results of 200 simulated samples based on β^E, κ^E and η^E drawn from their posterior distribution after wave 1.

7 Conclusion

This paper shows a concrete application of the exploration sampling algorithm to demonstrate the successful use of adaptive sampling in real-world policy choice problems. The experiment we conducted provides an opportunity to answer many implementation questions surrounding this new method. For instance, as part of the IVR experiment, we give two examples of Bayesian modeling for the outcomes of interest – here call engagement and oral reading fluency – and show how to use such models to compute the assignment shares in each wave and the posterior expected regret that is used to choose one arm for implementation. We discuss some of the constraints on the research design that are unique to adaptive experiments as well as the approaches to choosing between alternative designs based on simulations.

Our sample application tests six different designs for a new parent outreach method, interactive voice response calls, to encourage home reading with children in Kenya, which is known to improve early literacy. Even though the time and budget for the experiment were limited, the adaptive design is able to identify the call format with the highest level of engagement with 93% probability, leading to minimal expected losses from mistakenly selecting the wrong call format. Despite the short exposure period of just 5 weeks (9 calls in total) and despite the moderate uptake, the call format with the highest engagement level, which asks parents to carry out exercises after the call with the child and uses the same "intermediate" exercise sequence for all children, leads to a moderate but detectable improvement in ORF test scores of 1.68 correct words per minute ([0.13-3.21], or 0.065 standard deviations of the baseline reading fluency level). These findings make IVR calls a promising method of educational outreach. Identifying such methods has become an urgent policy priority, given the delays to schooling experienced by millions of children in the wake of the Covid-19 pandemic.

This EdTech application provides a compelling example for using adaptive sampling in policy choice experiments, showing that there are expected gains in the targeted outcome with even moderate adaptivity and a relatively large sample. We would expect even larger gains when more waves are possible, and in problem instances where (for example) a few inferior arms can be ruled out quickly, focusing sampling effort on a smaller subset of promising candidates.

As long as the added (per wave) cost is low, adaptive sampling has the potential to improve learning in many areas of policy, in particular when outcome data is regularly received as part of ongoing administrative data collection. The range of contexts in which this is the case continues to expand as public administrations shift towards digital record keeping and online interactions with beneficiaries and citizens. In other situations, the cost of adaptivity may be high, for example due to the added data collection effort, but the gains from increased efficiency are potentially also high; for example when the available sample is small or the welfare gains from implementing an effective policy faster are potentially large. From an ethics perspective, adaptive methods for policy choice can reduce the burden of experimentation with human subjects twofold; first, because the share of experimental subjects who receive the highest-performing policies increases as learning progresses, and second, because the same sample size generates greater learning gains with an adaptive over a non-adaptive design, increasing the potential for better policy outcomes afterwards.

As part of describing the design of this experiment, the paper tackles many implementation questions that we anticipate others will encounter as well. As more economists and policy makers begin to use adaptive methods, we hope they benefit from this example and the solutions we propose. The paper also reveals some potential challenges and highlights that an important – and in practice often difficult – step in the research design is choosing the right outcome measure. This may in future applications involve more formal methods of eliciting preferences from the policymaker in order to be able to correctly construct the posterior outcome distributions and select the optimal arm. Many of the issues raised point to fruitful areas for future research and will hopefully spur ongoing innovation to improve the method further.

References

Andrews, I. and M. Kasy (2019). Identification of and correction for publication bias. *American Economic Review 109*(8), 2766–94.

Andrews, I., T. Kitagawa, and A. McCloskey (2021). Inference on winners. *Working paper*.

Angrist, N., P. Bergman, and M. Matsheng (2020a). School's out: Experimental evidence on limiting learning loss using "lowtech" in a pandemic. *NBER Working Paper 28205.*

Angrist, N., P. Bergman, and M. Matsheng (2020b). School's out: Experimental evidence on limiting learning loss using "low-tech" in a pandemic. Technical report, National Bureau of Economic Research.

Athey, S., S. Baird, J. Jamison, C. McIntosh, and B. Özler (2021). A sequential and adaptive experiment to increase the uptake of long-acting reversible contraceptives in Cameroon. *AEA RCT Registry May 14*. https://doi.org/10.1257/rct.3514.

Athey, S., R. Chetty, G. W. Imbens, and H. Kang (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research.

Athey, S. and G. W. Imbens (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments*, Volume 1, pp. 73–140. Elsevier.

Audibert, J.-Y., S. Bubeck, and R. Munos (2010). Best arm identification in multi-armed bandits. In *COLT*, pp. 41–53. Citeseer.

Bahety, G., S. Bauhoff, D. Patel, and J. Potter (2021). Texts don't nudge: An adaptive trial to prevent the spread of COVID-19 in India. *Journal of Development Economics 153*, 102747.

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. V., S. Chassang, S. Montero, and E. Snowberg (2020). A theory of experimenters: Robustness, randomization, and balance. *American Economic Review 110*(4), 1206–30.

Banerjee, A. V., S. Cole, E. Duflo, and L. Linden (2007). Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics 122*(3), 1235–1264.

Bergman, P. (2021). Parent-Child Information Frictions and Human Capital Investment: Evidence from a Field Experiment. *Journal of Political Economy 129*(1), 286–322.

Bergman, P. and E. W. Chan (2021). Leveraging parents through low-cost technology: The impact of high-frequency information on student achievement. *Journal of Human Resources 56*(1), 125–158.

Berlinski, S., M. Busso, T. Dinkelman, and C. Martínez (2021). Reducing parent-school information gaps and improving education outcomes: Evidence from high-frequency text messages. Technical report, National Bureau of Economic Research.

Bettinger, E., N. Cunha, G. Lichand, and R. Madeira (2021, May). Are the Effects of Informational Interventions Driven by Salience? *Working Paper*.

Bowen, D. (2022). Multiple inference. https://dsbowen-conditional-inference.readthedocs.io/en/latest/?badge=latest.

Bubeck, S. and N. Cesa-Bianchi (2012). Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends® in Machine Learning 5*(1), 1–122.

Bubeck, S., R. Munos, and G. Stoltz (2009). Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pp. 23–37. Springer.

Caria, S., M. Kasy, S. Quinn, S. Shami, and A. Teytelboym (2020). An adaptive targeted field experiment: Job search assistance for refugees in jordan.

Christensen, G. and E. Miguel (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature 56*(3), 920–80.

de Barros, A. and A. J. Ganimian (2021). Which Students Benefit from Personalized Learning? Experimental Evidence from a Math Software in Public Schools in India. *Working Paper*.

Doss, C., E. M. Fahle, S. Loeb, and B. N. York (2019). More than just a nudge: supporting kindergarten parents with differentiated and personalized text messages. *Journal of Human Resources 54*(3), 567–603.

Duflo, E., R. Glennerster, and M. Kremer (2007). Using randomization in development economics research: A toolkit. *Handbook of development economics 4*, 3895–3962.

Garivier, A. and E. Kaufmann (2016). Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pp. 998–1027. PMLR.

Hadad, V., D. A. Hirshberg, R. Zhan, S. Wager, and S. Athey (2021). Confidence intervals for policy evaluation in adaptive experiments.

Hadad, V., L. R. Rosenzweig, S. Athey, and D. Karlan (2021). Practitioner's guide: Designing adaptive experiments.

ICTworks (2016, August). The blind spot of sms projects: Constituent illiteracy.

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis 24*(3), 324–338.

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

Kasy, M. and A. Sautmann (2021b). Correction regarding "adaptive treatment assignment in experiments for policy choice". *Working paper*.

Knauer, H. A., P. Jakiela, O. Ozier, F. Aboud, and L. C. Fernald (2020). Enhancing young children's language acquisition through parent–child book-sharing: A randomized trial in rural Kenya. *Early Childhood Research Quarterly 50*, 179–190.

Kraft, M. A. and M. Monti-Nussbaum (2017, November). Can Schools Enable Parents to Prevent Summer Learning Loss? A Text-Messaging Field Experiment to Promote Literacy Skills. *The ANNALS of the American Academy of Political and Social Science 674*(1), 85–112.

Lai, T. L. and H. Robbins (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics 6*(1), 4–22.

Lattimore, T. and C. Szepesvári (2020). *Bandit algorithms.* Cambridge University Press.

Madaio, M. A., V. Kamath, E. Yarzebinski, S. Zasacky, F. Tanoh, J. Hannon-Cropp, J. Cassell, K. Jasinska, and A. Ogan (2019). "you give a little of yourself": Family support for children's use of an ivr literacy system. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*, COMPASS '19, New York, NY, USA, pp. 86–98. Association for Computing Machinery.

Mayer, S. E., A. Kalil, P. Oreopoulos, and S. Gallegos (2019, October). Using Behavioral Insights to Increase Parental Engagement: The Parents and Children Together Intervention. *Journal of Human Resources 54*(4), 900–925.

Melfi, V. F. and C. Page (2000). Estimation after adaptive allocation. *Journal of Statistical Planning and Inference 87*(2), 353–363.

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.

Offer-Westort, M., A. Coppock, and D. P. Green (2021). Adaptive experimental design: Prospects and applications in political science. *American Journal of Political Science 65*(4), 826–844.

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.

Pouzo, D. and F. Finan (2022). Reinforcing RCTs with multiple priors while learning about external validity. *NBER Working Paper 29756.*

Qin, C., D. Klabjan, and D. Russo (2017). Improving the expected improvement algorithm. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5387–5397.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society 58*(5), 527–535.

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.

Russo, D. (2020). Simple Bayesian algorithms for best-arm identification. *Operations Research* (6), 1625–1647.

Sautmann, A. (2021a). Online supplement: Bridge Kenya IVR literacy intervention materials. https://bit.ly/3LosOgM.

Sautmann, A. (2021b). Text messaging for parental engagement in student learning. *AEA RCT Registry May 6*. https://doi.org/10.1257/rct.6701.

Sautmann, A. (2022). Interactive phone calls to improve reading fluency. *AEA RCT Registry April 9*. https://doi.org/10.1257/rct.7663.

Shang, X., R. Heide, P. Menard, E. Kaufmann, and M. Valko (2020). Fixed-confidence guarantees for Bayesian best-arm identification. In *International Conference on Artificial Intelligence and Statistics*, pp. 1823–1832. PMLR.

Shreekumar, A. (2020). winference. https://github.com/adviksh/winference.

Tabord-Meehan, M. (2018). Stratification trees for adaptive randomization in randomized controlled trials. *arXiv preprint arXiv:1806.05127*.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika 25*(3/4), 285–294.

Xu, M., T. Qin, and T.-Y. Liu (2013). Estimation bias in multi-armed bandit algorithms for search advertising. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Volume 26. Curran Associates, Inc.

York, B. N., S. Loeb, and C. Doss (2019, July). One step at a time: The effects of an early literacy text-messaging program for parents of preschoolers. *Journal of Human Resources 54*(3), 537–566.

Zhang, K., L. Janson, and S. Murphy (2020). Inference for batched bandits. *Advances in Neural Information Processing Systems 33*, 9818–9829.

Zhang, K., L. Janson, and S. Murphy (2021). Statistical inference with m-estimators on adaptively collected data. *Advances in Neural Information Processing Systems 34*.

A Intervention Design

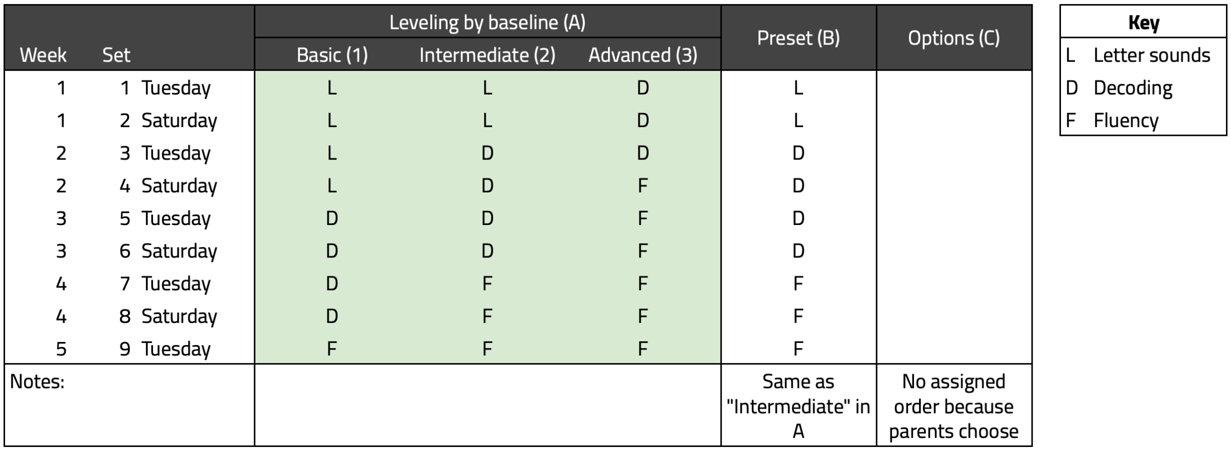
The three content intervention variants A, B, and C are as shown in figure A.1:

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 assigns 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 baseline scores, we assign them their class median. For classes with missing scores, we assign the intermediate level (which in this sample also happens to be the sample median).



		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 A.1: Exercise leveling variations.

B Oral Reading Fluency Data Quality

In this section we provide more details on the data quality issues with ORF scores.

Table A.1: Non-missing ORF scores in each exam, by treatment arm, and by wave.

	Wave 1 and 2								Wave 1	Wave 2
Period	C	T1A	T1B	T1C	T2A	T2B	T2C	Total	Total	Total
T2 ET	89.6%	88.2%	88.2%	87.8%	91.8%	90.0%	88.9%	88.9%	88.9%	88.8%
T3 MT	74.0%	74.8%	73.8%	71.4%	74.0%	73.3%	74.8%	73.5%	74.3%	72.6%
T3 ET	79.3%	81.5%	81.9%	78.0%	83.5%	79.9%	77.4%	80.3%	80.5%	80.1%
Total	81.0%	81.5%	81.3%	79.1%	83.1%	81.1%	80.4%	80.9%	81.2%	80.5%
N. students	415	330	781	581	231	329	226	2893	1509	1384

Notes: the table presents the percentage of valid ORF measurements for students allocated to each treatment arm in Wave 1 and 2.

Table A.1 displays the percentage of non-missing ORF measures across treatment arms and periods. There is no evidence of a systematic relationship between ORF attrition and the treatments arms. However, after the last data delivery received by the researchers in fall 2021, the endterm exam of term 2 has the highest average percentage of ORF measures (88.9%) compared to the midterm of term 3 (73.5%) and the endterm of term 3 (80.3%).

There are many possible reasons for these patterns. One reason for the endterm difference could be that teachers even at the last data delivery had not submitted all their scores for term 3. The number of scores collected in the midterm may be lower because the examination period for ORF was shorter (2 hours) than in the endterms (3 hours). Children may also be more likely to miss the midterm than the endterm.

Table A.2: Average ORF scores from separate data deliveries for endterms of term 2 and term 3.

	Treatment arm							Number of
	C	T1_A	T1_B	T1_C	T2_A	T2_B	T2_C	students
E2	39.7	38.8	41.7	40.1	39.7	41.2	42.1	2285
E2 updated	62.3	53	60.6	58.6	53.6	55.6	58.2	286
E3	48.7	49.9	52	49.1	51.8	51.2	50.2	1897
E3 updated	42.2	46.1	47.3	44.6	35.9	43.8	48.4	425

Notes: first set of scores obtained for each endtem exam shortly after grading day. Original E2 scores were used for leveling for term 3. The updated scores are ORF scores for children whose grades were uploaded to the system later and obtained in a second data delivery for all exams weeks after intervention end. Midterm of term 3 not shown because there were less than 20 students with an updated score in the second data delivery.

Table A.2 shows average scores from separate data deliveries we received for the endterm exams of term 2 and term 3. Each data delivery included all scores that were submitted up to that point. The first delivery was received shortly after each exam took place. Crucially, for term 2, this was also the time when exercise leveling based on reading ability for the next term was determined, in order to start IVR calls in time for the next term. The second data delivery (for all exams) was received in Fall 2021.

The data show large differences between the scores submitted soon after the exam vs. later (during the next term). This is especially true for the data from term 2. This gap in scores could be an explanation for why leveling reading exercises is not as successful: children with missing scores tended to have better reading skills, and they might have received too easy exercises on average. Interestingly, while the average in the second data delivery is higher for endterm 2, for endterm 3 it is lower. These differences could be due to systematic patterns in the time of score submission, such as remote locations having poorer internet connectivity and also lower reading levels: note that the second delivery for term 2 was much smaller than for term 3. But the difference could also stem simply from variation that arises because scores for a whole school or classroom are sent at once and there is a lot of inter-school variance.

In any case, the two tables show that even after many weeks, a substantial share of ORF scores for each exam was still missing. When examining scores, we additionally found an unusually large percentage of scores that are multiples of 5 ("rounded" scores). One reason for this could be measurement error, stemming for example from the teacher having only imprecise means to measure time.

C Additional Results

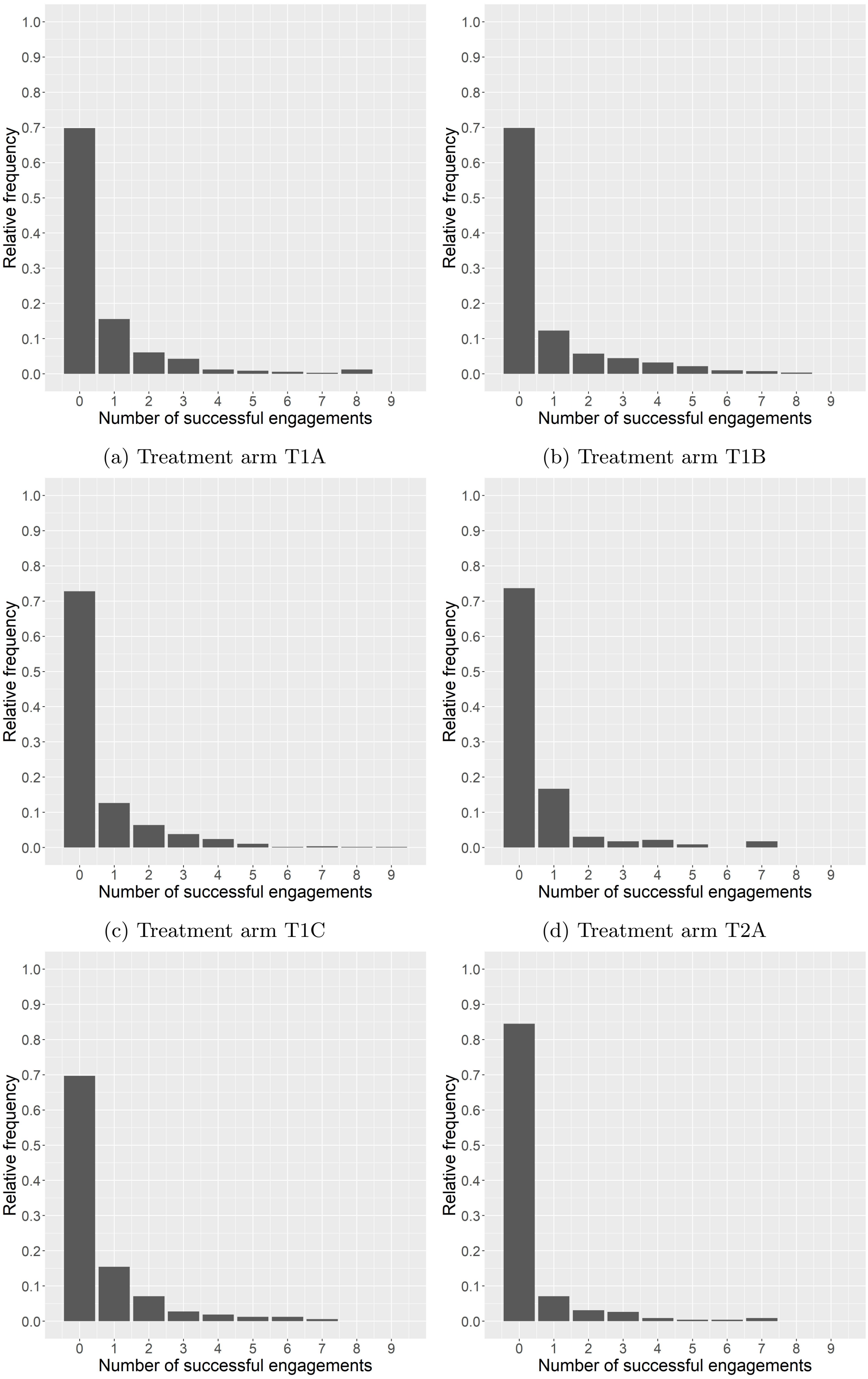
C.1 Observed call engagement

Table A.3 show sthe average number of calls (out of nine calls) with successful engagement, by treatment arm and wave, and Figure A.2 shows the histogram of observed call engagement. The first bar shows the number of phone numbers with zero engagement. This share is nearly the same in every call format except in treatment arm T2C, suggesting that the same share of parents listen to the exercises at least once. Differences in sustained engagement arise from the second call onward.

Table A.3: Mean number of successful engagements by treatment arm and wave.

	Wave 1	Wave 2*	Wave 1 and 2
T1A	0.602	0.752	0.655
T1B	0.745	0.761	0.756
T1C	0.633	0.554	0.582
T2A	0.542	0.429	0.535
T2B	0.595	0.703	0.635
T2C	0.358	-	0.358

Notes: * No observations were allocated to treatment T2C in Wave 2.



(a) Treatment arm T1A

(b) Treatment arm T1B

(c) Treatment arm T1C

(d) Treatment arm T2A

(e) Treatment arm T2B

(f) Treatment arm T2C

Figure A.2: Observed call engagement, by treatment arm

C.2 Probability optimal for ORF

Table A.4: Posterior regret and probability of highest ORF score gains.

	Balanced		Unbalanced	
Treatment	Posterior regret	Prob. highest	Posterior regret	Prob. highest
T1A	1.636	8.55%	1.528	9.11%
T1B	0.935	20.30%	1.143	11.54%
T1C	1.458	8.33%	1.911	2.35%
T2A	1.134	22.23%	0.973	26.34%
T2B	1.123	20.60%	0.918	25.80%
T2C	1.219	20.00%	1.029	24.86%

Notes: Posterior regret expressed in terms of correct words per minute. The table contains information from 8,000 posterior draws sampled from 4 independent Markov chains.

C.3 Ex Ante Comparison of Exploration Sampling and RCT

Table A.5: Ex ante comparison: performance of exploration sampling and standard RCT in simulated samples based on many parameter draws from the prior.

Panel A: Averages of Posterior Estimates.

Exploration Sampling		Standard RCT	
Avg. posterior expected policy regret	Avg. posterior probability optimal	Avg. posterior expected policy regret	Avg. posterior probability optimal
0%	98.97%	0.01%	98.91%

Panel B: Average Realized Values.

Exploration Sampling		Standard RCT	
Average policy regret	Percentage best arm identified	Average policy regret	Percentage best arm identified
0%	98.99%	0%	98.99%

Notes: The table shows averages from 100 simulated samples drawn using the parameter vector $\{\beta^E, \kappa^E, \eta^E\}$, drawn from their prior distributions. For each sample draw, the same first wave was used, the second wave was drawn either using the exploration sampling shares based on the estimates from the first wave, or using equal assignment shares.

Table A.5 shows simulation results when drawing hypothetical treatment arm averages θ^k from a uniform distribution, simulating two experimental samples (one with exploration sampling, one with non-adaptive sampling) for each draw, and estimating the model parameters from these samples. Note that both equal and adaptive sampling shares essentially lead to zero regret on average. This is because random independent draws for the average success rate in the different treatment arms often lead to one arm that is clearly a "winner". In reality, it is likely that the success rates in the different arms are highly correlated and will be clustered more closely than typical random draws from the unit interval.

C.4 Extensive margin for call engagement

Table A.6: Extensive margin for call engagement: model coefficients.

Treatment	Any successful engagement (1)	At least one second in call (2)
T1A	-0.85^{*}	1.71^{*}
	$[-1.09; -0.61]$	$[1.42; 2.03]$
T1B	-0.85^{*}	1.84^{*}
	$[-1.01; -0.70]$	$[1.63; 2.06]$
T1C	-1.00^{*}	1.91^{*}
	$[-1.19; -0.81]$	$[1.67; 2.18]$
T2A	-1.05^{*}	1.54^{*}
	$[-1.34; -0.76]$	$[1.20; 1.90]$
T2B	-0.84^{*}	1.86^{*}
	$[-1.08; -0.60]$	$[1.54; 2.19]$
T2C	-1.72^{*}	1.82^{*}
	$[-2.10; -1.37]$	$[1.46; 2.22]$
Num. students	2462	2462

Notes: * Null hypothesis value outside 95% credible interval. We simulate 4 independent Markov chains of 4,000 posterior draws each and discard the first 2,000 as warmup. The remaining 8,000 draws are used to generate the posterior distributions of the coefficients. The Split-$\hat{R}$ of every coefficient is below 1.01 and there are no divergent transitions.

Table A.7: Extensive margin for call engagement: probability of engagement.

	Any successful engagement		At least one second in call	
Treatment	Mean	Prob. highest	Mean	Prob. highest
	(1)	(2)	(3)	(4)
T1A	30.12%	32.15%	84.12%	5.75%
T1B	30.04%	26.55%	85.70%	12.09%
T1C	27.15%	2.26%	86.59%	35.15%
T2A	26.20%	3.78%	81.70%	0.91%
T2B	30.28%	35.26%	85.92%	25.11%
T2C	15.41%	0.00%	85.40%	20.99%

Notes: (1) and (3): The average probability is calculated in analog with Eq. **??**. (2) and (4) The probability optimal is calculated as in Eq. 4.