Evaluation of a Teacher Training in Uganda: Specifications for Endline

Moustafa El-Kashlan Stefan Faridani Vesall Nourani

Sara Restrepo-Tamayo

January 2025

Abstract

We evaluate the impact of a general skills teacher training program in rural Uganda. Nourani et al. (2023) found large impacts of the training on student test scores in primary schools. In this new randomized controlled trial, we offer the program to a fresh sample of 640 teachers in 39 secondary schools. We use a new teacher-randomized design to study several follow-up questions: Does the treatment effect replicate in secondary schools and when teachers cannot easily self-select into training? Are there teacher-to-teacher knowledge spillovers? We have three main data sources: student assessments given by our team, teacher surveys taken by our team, and high-stakes government administered exam scores. Randomization occurred in November 2021. Midline data was collected in the Falls of 2022 and 2023. Endline data was collected in the Fall of 2024 but has not been analyzed. Learning outcome data for 2023 and 2024 has not yet been processed. This plan presents results from midline data and computes statistical power for when those analyses are run again on endline data.

* Corresponding author: Vesall Nourani, vesall_nourani@g.harvard.edu. We would like to thank Kimanya Ngeyo Foundation for Science and Education for providing us the opportunity to study their teacher training program in Uganda. We also acknowledge and thank the Echidna Foundation for making this research possible.

1 Introduction

While 49% of Ugandan children enter secondary school, fewer than 10% enter the final year (Rooke, 2014). In 2007 Uganda eliminated most secondary school fees as part of its Universal Secondary Education (USE) subsidy program. Yet completion rates stagnated three years later (World Bank, 2024). Low quality teaching driven by inadequate teacher training may be partly to blame. Although 90% of Ugandan secondary teachers have the required formal qualifications (Vasiliev and Demas, 2018), teachers in schools receiving USE subsidies scored 0.5 standard deviations lower on a numeracy test and 0.28 standard deviations on a literacy test than teachers in non-implementing schools (Najjumba and Marshall, 2013). Recent reforms to the Ugandan national secondary school curriculum and teacher training standards reflect policymaker awareness of the teaching quality challenge (Businge, 2019).

Kimanya Ngeyo Foundation for Science and Education (Kimanya) was founded in 2007 and has offered a teacher training in primary schools since 2014. The program ultimately trains teachers to deeply consider *why* they use the teaching practices that they use, to analyze whether their actions have their intended effects, and to consider changes in their actions that can enhance those effects. In short, it trains them to be lifelong learners and to refine their professional activities through a process of continuous learning. The training itself is multifaceted, but some of the core experiences of the training include (i) to utilize the process of scientific investigation in daily life, (ii) to reflect on how a learning experience that emphasizes investigation benefits from participatory and learner-centered pedagogy, and

(iii) to reflect on the purpose of education and the important role teachers play in shaping the experiences of their students. Kimanya training occurs in three rounds of about 11 days each over the course of a school year. In addition, Kimanya tutors visit participating teachers monthly. A randomized evaluation of Kimanya training in primary schools found school-level intention-to-train raised student scores on the high-stakes primary leaving exam by 0.5 standard deviations which raised the pass rate from from 51% to 75% after two years (Nourani et al., 2023).

The present experiment is a separate randomized controlled trial that differs from the first in at least two key respects.[1] First, the present trial randomizes at both the teacher and school level while (Nourani et al., 2023) randomized only at the school level. The teacher-level design allows us to study teacher-to-teacher spillover effects and also to check whether the program effect remains when teachers cannot easily self-select in training. Second, the present trial takes place in secondary schools, and not primary schools, during a period of curriculum transition in Ugandan secondary schools. The new competency-based secondary curriculum — offered for the first time to half of the baseline cohort in our study — is a frame shift from education as static knowledge collection to learning as a dynamic tool for the individual, community, and societal development. Competency-based education shifts educational goals from what a learner is meant to know to what a learner is meant to do. Thus, the secondary-school setting allows us to study whether a training in which teachers experience something akin to a competency-based approach to learning makes them more effective as teachers under the new

[1] AEA registriation is here: https://www.socialscienceregistry.org/trials/12347

curriculum. The secondary school context also allows us to check whether the treatment effect can be replicated in a related but distinct context with older students and a distinct professional class of teachers.

The experiment is ongoing and Figure 1 presents the full study timeline. This document is a plan that describes main research questions that the trial will answer by the end. The plan specifies empirical methods to answer each question. The three datasets needed to answer our questions will be collected at endline in fall 2024 after all treatment teachers have been given the opportunity to train. However, we have already collected midline versions of each of these samples in 2022 and 2023. This document runs our specifications on the midline data in order to demonstrate that our empirical methods are fully specified and sufficiently precise. In addition, this document also describes a separate, exploratory set of analyses that will not be pre-specified.

We detect no midline effects on student exam scores after one year. This is somewhat expected (and similar to results from primary schools), possibly because it takes time for teachers to internalize changes to their teaching practices and for students to reap the benefits of these changes. Additionally, student exposure to trained teachers was still low and teacher turnover was high (about 15% of teachers in our study left their jobs prior to 2022 training). However, we do detect average changes in teacher attitudes at the second midline—particularly among teachers who were randomly assigned to train in groups of friends. We will re-run analyses on endline data, which has not been examined at the time that this report was written.

Whatever its findings, this trial will contribute to the literature on improving teaching quality in developing countries. Supply-side interventions often need to come alongside improved teaching to be effective (Glewwe et al., 2009; Kremer et al., 2013; de Ree et al., 2017). Meta-analyses on the impact of teacher training and professional development have yielded promising but highly heterogeneous results (McEwan, 2015; Popova et al., 2021). Recent randomized evaluations find that teachers in developing countries can be quick to effectively adopt innovative new teaching practices (Nourani et al., 2023; Alan and Mumcu, 2024). The present experiment will contribute to understanding of when and how teacher training can improve learning outcomes in developing countries.

The rest of this plan is organized as follows. Section 2 specifies the main confirmatory research questions. Section 3 describes our study population, the randomized design, and teacher compliance. Section 4 describes our data and discusses balance and attrition. Section 5 specifies the empirical methods that we will use to answer each of the "confirmatory" research questions, while 6 runs these methods on midline data to demonstrate that they are fully specified and calculates power for when our methods are re-run on endline data. Section 7 describes the exploratory part of the study, which seeks to update and revise a theory of how training affects teacher behavior. The remaining sections list the members of the research team, describe our intended deliverables, present the full study calendar, and list our IRB approvals.

2 Research Questions

Three primary research questions are enumerated below. Question Q1 asks whether Kimanya's teacher training program affected student exam scores. Question Q2 asks whether exam score effects can spill over across teachers. Research question Q3 asks whether the training affected teacher attitudes towards the purpose of education and gender. In Section 5.6 we will link each research question to a small number of primary hypothesis tests.

Q1 Student Learning

(a) Can we replicate the impact on student learning outcomes found by (Nourani et al., 2023) in secondary schools? Does the program effect remain when teachers do not self-select into training?

Q2 Teacher Spillovers

(a) Does training with a friend improve program effectiveness on student exam scores?

Q3 Teacher attitudes and perceptions of their students and the purpose of education.

(a) Do teachers' perceptions of the purpose of secondary education change? How do spillovers in such perceptions depend on connections to trained teachers?

(b) Do teachers' perceptions of their students' abilities reflect a deeper appreciation of each student's potential? Is this reflected in more equitable gender attitudes?

3 Research Design

3.1 Study Population

Our study consists of 39 secondary schools located in the Jinja and Buikwe districts of southeastern Uganda. We chose the schools in the following fashion. In April-June 2021 we obtained teacher rosters from the 52 secondary schools in both districts and conducted a baseline telephone survey of the teachers on the roster. The survey is described in Section 4.1.

We used these survey responses to determine eligibility criteria for teachers. We deemed a teacher to be eligible for training if they (i) taught at least one of the first three years of secondary school and (ii) reported spending at least four hours per week at the school on whose roster they appeared.[2] This left us with 879 eligible teachers. Then we removed the 15 schools with fewest eligible teachers to arrive at 39 schools containing 807 eligible teachers. From these we selected 640 teachers to be part of our randomization.

The set of teachers included in the randomization was chosen to meet two criteria. First, Kimanya wanted to invite approximately 100 teachers per year to

[2]Since the survey was taken during the COVID-19 pandemic, four hours worked per week was above the mean in our survey. A national lockdown was implemented on June 7th, midway through our baseline survey.

train over three years. Second, our research team wanted each school to contribute approximately the same percentage of its eligible teachers to the training, which turned out to be about 60%. We prioritized including teachers who said that they worked more hours on our survey.

Table 3 compares teachers not in the randomization to those whom we chose for the randomization. Compared to the other 524 teachers, our chosen 640 teachers were on average 1.4 years older, 8 percentage points more likely to be female, 5 percentage points more likely to be married, 8 percentage points more likely to teach only at the school where they were surveyed, and worked one more hour per week in their surveyed school.

3.2 Assignment to Treatment

We randomized at the school and teacher level. The aim of our design was to not only identify the program effect on individual teachers, but also to study teacher-to-teacher spillovers and effects on school culture. We therefore randomize not only who receives training, but also whether teachers are trained in groups of friends or not. To achieve this, the experimental design consisted of four steps: constructing teacher clusters, constructing school triples, school-level randomization, and within-school randomization. Figure 2 visualizes the experimental design and the following paragraphs explain the four steps.

The first step in our design was to partition the teachers within each school into four equally-sized treatment clusters using two different methods. For each school we first sorted teachers into what we called *clique clusters*. Clique clusters

were chosen in order to maximize the number of teacher-to-teacher social ties within each cluster. Then for each school we sorted those same teachers into four different clusters that we called *anticlique clusters*. Anticlique clusters were chosen in order to minimize the number of social ties within each cluster. The result is that each teacher belonged to one Clique cluster and one Anticlique cluster.

Figure 3 visualizes an example of the teacher-level randomization in a small school that sends 2 teachers per year—the minimum and modal number of sent teachers in our sample. The picture visualizes how the teacher social network maps into the Clique and Anticlique clustering schemes. The baseline data used to construct teacher social networks is described in Section 4.1.

In the second step, we sorted the schools into thirteen matched triples of three schools each. Triples were chosen in order to make the number of eligible teachers per school roughly the same within each triple. This was to guarantee that for every realization of the randomization, roughly the same number of teachers would be invited to training.

The third step was to randomize at the school level. Within each triple we randomly selected one school as a Pure Control, one as a Clique school, and one as an Anticlique school. In the 13 Pure Control schools, no one would be invited to train. In the 13 Clique schools, teachers would be invited to train in groups of friends. In the 13 Anticlique schools, teachers would be invited to train in groups with few social ties. Thus our school-level design has three experimental arms: two different treatments and one pure control. We refer to both Clique and Anticlique schools as "treatment schools."

The fourth and final step is to randomize within schools. Each treatment school's teachers were already partitioned into four treatment clusters depending on the school's assignment. In Clique schools we use the four Clique clusters and in Anticlique schools we use the four Anticlique clusters. From each treatment school, one cluster is randomly invited to train in 2022, one cluster is invited to train 2023, one cluster is invited to train in 2024, and the last cluster is kept as a control. See Figure 2 for a flowchart.

It is crucial to realize that Clique vs Anticlique assignment has no effect on the conditional probability that any individual teacher is invited to training. Conditioning on Clique vs Anticlique assignment affects only the *joint* distribution of teacher invitation probabilities but not the *marginal* distributions. So, if a teacher-level SUTVA holds where each teachers' outcomes depend only on their own training invitation, then any difference between Clique and Anticlique schools could be interpreted as a violation of SUTVA. This special design allows us to detect teacher-to-teacher spillover effects on student exam scores even when we cannot link individual students to specific teachers.[3]

Two schools changed ownership between our baseline survey and the second training in April 2023. Consequently both schools experienced very high faculty turnover. Before training began, we re-randomized the 2023 teacher invitations for both schools using the post-ownership-change teacher rosters.

[3]We have attempted to link students to teachers via classrooms, but the links so far are conservative and the resulting networks are quite dense. We therefore do not discuss this kind of analysis in this plan. Our team is working on improving this match and we hope to be able to do this kind of analysis soon.

3.3 Compliance at Midline

While teacher *invitations* were randomized, actual *attendance* at Kimanya's training is voluntary and therefore cannot be randomized. Moreover, our context includes relatively high teacher employment turnover. To account for imperfect attendance at the trainings, all of the estimands studied later on in this plan are intent-to-treat. In this section we summarize teacher attendances as reported by Kimanya.

While most invited teachers attended significant portions of the training during the first two years of the campaign, some did not. By far the most common reason for non-attendance was that the invited teacher had stopped working at the school that we had initially surveyed them at. Table 1 displays all of the compliance information for 2022 and 2023. By the end of 2023, 176 teachers had been invited to training and 131 (75%) of these had attended at least three days of training. Moreover, 79 (45%) of the invited teachers had attended for at least twenty days. Figure 4 shows the full CDF of days of training attended for teachers who had been invited by the end of 2023. All 26 treatment schools had sent teachers to train for at least three days by the end of 2023. Among invited teachers, those who attended for at least 3 days were slightly more likely to be female, but did not differ in age from teachers who did not attend training for at least 3 days. Table 2 quantifies these comparisons.

The most common reason for noncompliance was that the teacher had stopped working at the school where we had initially surveyed them at. Out of 33 invited teachers whom headteachers told us would not attend in 2022 or 2023, for 72% the reason given was that the invited teacher had left the school. This kind of non-

compliance will decrease our intent-to-treat effects in a way that is likely unrelated to the impacts of Kimanya training.

In addition, several teachers came to training even though they were not invited. By the end of 2023, 17 uninvited teachers had attended for at least three days and 5 of these had attended at least twenty days.

Teacher attendance patterns suggests that teachers talk to one another about the training and raises the possibility of teacher-to-teacher spillover effects. All 17 uninvited teachers who attended anyway for at least three days worked at treated schools. Nine were from Clique schools and eight from Anticlique.

To avoid wasting capacity, Kimanya wished to minimize the number of empty seats at its trainings due to noncompliance. When a teacher is unavailable or declines an invitation, we invited a replacement teacher from a prepared list. Prior to randomization, for each treatment cluster we prepared an ordered list of "replacement" teachers in the school that were not otherwise part of the randomization. As teachers from a treated cluster declined invitations, we invited replacement teachers in order down the list associated with the cluster of the declining teachers. A total of 65 teachers had been invited from replacement lists by February 2024. Replacement teachers are not included in Table 1 or any of the analyses in this plan.

4 Data, Attrition, and Balance

This plan discusses five data sources: standardized exams administered by the Uganda National Education Board (UNEB), a teacher survey administered by our team at baseline, an exam given by our team to students, and a midline survey given by our team to teachers. Table 4 enumerates all of the datasets discussed in this plan, including future samples. Sections 4.1-4.4 later on discuss each dataset in detail.

We face significant data attrition in our context. Teachers sometimes change schools or exit the study sample, and one school did not allow its students to sit for our exam. This means that all of our midline and endline datasets are samples, not populations. If treatment causally affects who appears in our data, then this can confound our estimates of treatment effects. We will need to assume that this is not the case.

Assumption 1 states that treatment does not affect which students, teachers, and schools attrit out of our sample. Importantly, Assumption 1 would be violated both by imbalance in the *number* and the *kind* of observations across treatment arms. Without Assumption 1, the estimands studied later on do not have causal interpretations.

Assumption 1. *The sample of units observed in each of our datasets would have been the same under any counterfactual treatment assignment.*

Assumption 1 is not verifiable but we can try to falsify it with balance checks. A balance check tests for differential rates of attrition by treatment arm within each

sample. The primary statistic of interest for each balance check is the p-value of a Fisherian exact test, sometimes called "Randomization Inference" (Young, 2018). Here Assumption 1 is the sharp null hypothesis that Fisher's test is testing. The test statistic is always the overall F-score of a regression of treatment status indicators on demographic characteristics. The critical values for the Fisherian exact test are calculated by re-randomizing, computing our test statistic under the counterfactual randomizations, and calculating the percentiles of the distribution of permuted statistics. This test would reject its null hypothesis if, for example, men were over-represented in followup surveys of treated teachers but not control teachers.

The following sections describe how each baseline or midline dataset was collected, provide summary statistics, and check for balance. Assumption 1 is never rejected in any of our samples.

4.1 Baseline Teacher Survey

During April-June 2021 our team surveyed 1164 teachers from 54 secondary schools in Jinja and Buikwe. Our survey team was contracted with Innovations for Poverty Action. 834 teachers were surveyed in person and the rest were surveyed over the telephone due to COVID-19 lockdowns. Survey questions included teacher demographics, grades taught, and how many hours they worked at the school that they were surveyed at, and whether they taught at any other schools.

Table 3 presents summary statistics for the demographic characteristics of teachers surveyed at baseline. We used this data in November 2021 to select our study population of 640 teachers as described in Section 3.1. The summary statistics in

Table 3 are partitioned into teachers whom we chose to be part of our randomization and teachers whom we did not choose.

We check for balance at baseline by comparing school characteristics across treatment arms. Table 5 compares characteristics of treatment schools to pure control schools. Table 6 uses the same data to compare Anticlique schools to Clique schools. The Fisherian exact test (randomization inference) fails to reject the null hypothesis of balance at baseline for both comparisons.[4]

The baseline survey was used to get the social networks used to construct the teacher clusters discussed in Section 3.2 and illustrated in Figure 3. The survey elicited teacher social networks in the following fashion. For each surveyed teacher (called the "ego") we randomly chose fifteen other teachers (called "alters") at the same school. For each alter, we asked the ego whether they had spent 30 minutes talking to the alter during the last academic term. If the ego answered in the affirmative, then we asked the following three followup questions:

- Have you and [ALTER] spoken about how to better **manage** your classroom?
- Have you or [ALTER] **visited** any of each others classroom to help improve teaching practices?
- Have you **planned** classroom activities together with [ALTER]?

We considered two teachers to be "linked" if they both answered in the affirmative about the other to at least one of the three followup questions. Teacher

[4]The test statistic for the exact Test is the overall F-score of a regression of school baseline characteristics on school treatment status.

clusters in the Clique schools described in Section 3.2 were chosen to maximize links and clusters in Anticlique schools were designed to minimize links.

4.2 Student Assessment

In October and November of 2022 we gave an exam to 11,495 students. One school (assigned to Anticliques) did not allow us to give the exam, so our exam data comes from 38 schools. The exam was administered by a team contracted with Innovations for Poverty Action. The exam-takers were every student, from S1 to S4,[5] present at the school on the day that the IPA team arrived to give the assessment. The students weren't forced to participate in the assessment but the teacher was present while the students were taking the assessment, increasing the participation of the students on the assessment. Table 7 provides summary statistics for the demographics of students who took our assessment.

Our exam consisted of 24 multi-part questions and students had one hour to complete it. A subset of questions were designed bespoke for this project by El-Kashlan and Nourani, while the remaining were taken from the 2011 TIMSS standardized examination, which at the time of study was the most recently available version of the exam available to researchers. The objective of the exam was to measure students' ability to use scientific language and to apply critical thinking skills to scientific problems. The exam questions were similar in spirit to those that the teachers are asked during Kimanya's training sessions. Figure 5 provides an example of a question from the "Use of Language" section where students are asked

[5]The first year of secondary school is called S1, the second S2, and so on.

to decide when to use the word "science" and when to use the word "technology".[6] Figure 6 shows an example of a question from the "Critical Thinking" section where students are asked to construct a simple experiment to test a hypothesis.

We computed exam scores by taking 35 question-parts that had single correct answers and adding up the number of correct answers per exam. Then we subtracted the mean and divided by the standard deviation over all exams.

Since one Anticlique school did not take our assessment, we are concerned that student exams might not be balanced between treatment and control. Table 7 compares the demographic characteristics of students who took our assessment across treatment and control schools. We check for balance in the following characteristics: student age, student gender, indicators of whether the student was in year 2, 3, or 4 of secondary school, and the mean UCE score for the student's school prior to baseline.

To check for balance overall, we run a Fisherian exact test using the overall F-score of a regression of the treatment indicator on the six characteristics in Table 7 as our test statistic. The test fails to reject Assumption 1.

We gave a similar exam in Fall 2023 and Fall 2024 but has not been digitized yet and so we cannot analyze it here.

[6]Since the S4 students were preparing for their UCE exam, we solely administered Section 1 - Use of Language to them.

4.3 UNEB Exam Scores

Students in the fourth year of secondary school (called S4) take the Uganda Certificate of Education (UCE) exam every year. The exam is administered by the Uganda National Board of Education (UNEB). The exam is mandatory and high-stakes and is comparable to the "O-level" exam in the United Kingdom.

The UCE exam contains six mandatory subjects: Mathematics (150 minutes), Biology (120 minutes), English language (120 minutes), Chemistry (120 minutes), Physics, (120 minutes), and either Geography (150 minutes), or History (120 minutes) or Religious education (90-120 minutes).[7] Scores are aggregated across all subjects by UNEB. We process aggregate scores by subtracting the mean and dividing by the standard deviation.

We attempt to match student's UCE exam scores to their score on the Primary Leaving Exam (PLE). The PLE is a high-stakes exam taken in the last year of primary school that is used for admissions into secondary school. We use standardized PLE scores as a student-level covariate in our analyses because they were determined prior to treatment and are strongly correlated with UCE exam scores. Similar to the UCE, our PLE scores have been aggregated across all subjects by UNEB.

UCE and PLE scores were provided and matched by UNEB. We obtained data by requesting that UNEB provide scores for exams taken by students attending schools in our study area. Unfortunately, UNEB's data is linked not to students' schools but only to the testing center that serviced that school. To solve this problem, we asked schools to provide the index numbers and testing centers of exams taken by their

[7] http://nada.uis.unesco.org/nada/fr/index.php/catalogue/47

students. We then use students' names to improve the match.

We have matched data from 2018-2020 and 2022. The exams prior to baseline are processed differently than exams after baseline, so we will discuss them separately.[8]

4.3.1 UCE Prior to Randomization

Our data contain 12731 UCE exams taken at 49 testing centers during 2018-2020. These testing centers were selected to correspond to the schools in our randomization. Since these exams were taken before randomization we treat them as predetermined covariates. We were able to match 11483 out of 12731 pre-baseline exams to 37 out of our 39 schools. One missing school was assigned to Cliques and the other to Anticliques.

Since not every exam was successfully matched to a school, we may be concerned that Assumption 1 was violated somehow. We do not find evidence that the number of observations differed by school treatment status. 3623 scores were from Clique schools, 4080 from Anticlique schools, and 3780 from Pure Control Schools. Fisher's exact test does not reject the null hypothesis that the number of exams per school was unaffected by school treatment status.

We use these pre-baseline exam scores as covariates in our analyses. We do this by taking the mean of standardized UCE exam scores within each school-year and then meaning the school-years within each school. For the two schools without

[8]While we do have 2023 data from UNEB in hand, we need to obtain more information from schools to complete the match. We will attempt to access 2024 data from UNEB in Spring 2025. However, due to significant exam reforms during 2024, we anticipate some delays or difficulties in acquiring the last year of UNEB data.

any baseline scores, we impute their mean standardized scores as zero. Imputation does not affect the validity or interpretation of our analyses because these means are used as predetermined covariates in the doubly robust estimators described in Section 5.5. Balance remains even after imputation. School mean standardized scores are only .04 standard deviations higher in treated schools than control schools. Fisher's exact test finds that this difference is statistically insignificant.

4.3.2 UCE Scores in 2022

Our data contain the universe of exams in Buikwe and Jinja districts taken in 2022, which is 14,595 exams.[9] Since these exams were taken post-baseline we treat 2022 scores as an outcome variable, though we do not expect a significant treatment effect in 2022 given our hypothesis that the treatment takes time to manifest in student learning outcomes. We were able to match 4456 exams to students in 38 out of 39 of the schools in our randomization. The missing school did not provide index numbers for their students in 2022.

The number of matched exams does not seem to differ by treatment arm. Our matched data contain 1462 exams from Clique schools, 1436 from Anticlique, and 1558 from Pure Control. These do not differ significantly from the 1485 exams per treatment arm that we would expect under perfect balance.

Table 8 shows balance for our sample of 2022 UCE test-takers. Treatment and control schools are very similar in this sample across all pre-baseline characteristics. The Fisherian exact test finds no evidence against Assumption 1.

[9]We have data from a significantly larger area in 2022 than in previous years. This makes the sample size of matched data is somewhat larger (4456 in 2022 vs 3782 in 2020).

We have UCE exam scores from 2023 but the data has not been fully matched to schools in our randomization and is therefore not analyzed in this plan.

4.4 Teacher Survey

In October and November of 2023 our team surveyed 1153 teachers at all 39 schools. We attempted to reach all 640 teachers in the randomization but only reached 455. The other 698 survey responses were from teachers not in the randomization and will not be discussed further in this document.

The key outcome variables discussed in this document are the responses to one particular survey module that asks teachers the extent to which they agree with each of eight statements about the purpose of education and how gender is related to it. The purpose of these questions is to determine whether training changes teachers' attitudes towards boys' and girls' education. The statements are listed below. Table 9 provides summary statistics.

1. Boys have more ability than girls to learn science subjects.
2. Boys need more education because they are better suited than girls to find high-paying jobs.
3. Parents should maintain stricter control over their daughters than their sons.
4. Girls should attain higher education so that they find better husbands.
5. Boys should attain higher education so that they find better wives.

6. Boys should be just as capable as girls at preparing a meal for visitors when guests visit.

7. Girls have more ability than boys to learn science subjects.

8. Girls need more education because they are better suited than boys to find high-paying jobs.

Since 185 teachers from our randomization are missing, we must test whether this attrition violated Assumption 1. Our first piece of evidence in favor of Assumption 1 is that 65% of survey respondents come from treated schools which is very close to the 66% of the study population of 640 teachers were in treatment schools. Second, 53% of survey respondents had been invited to training by 2023 which is very close to the 50% of the study population that was assigned to receive an invitation. This suggests that treatment had no influence on the number of survey responses.

We next wish to check that treatment did not change who responded. Table 10 compares characteristics of survey respondents across pure control and treated schools while Table 11 compares characteristics of respondents who had been invited by 2023 to uninvited respondents. We do not find statistically significant differences for any one characteristic. To check for balance overall, we run the Fisherian exact test using the overall F-score of a regression of the treatment indicator on the teacher characteristics as our test statistic. Fisher's test finds no evidence against the null hypothesis of Assumption 1.

We collected another round of teacher survey data in October and November of

2024 but this data has not been analyzed yet.

5 Emprical Methods

5.1 Primary Outcome Variables

We study effects on three primary outcome variables:

- Research questions Q1 and Q2 consider student scores on two exams:
  - The fraction of correct responses on our student exam described in Section 4.2.
  - All-subject aggregate score on the Uganda Certificate of Education standardized exam described in Section 4.3.
- Research question Q3 considers indicators of whether teachers agreed with eight statements on our teacher survey described in Section 4.4.

5.2 Notation

Consider a sample of m teachers and n students who appear in our data. Let T_{jts} be the treatment assignment of teacher j at school s in year t and let $\mathbf{T}$ be the full $3m \times 1$ vector of all teacher treatment assignments for all three years of the study. Let $Y_{its}(\mathbf{T})$ be the potential exam score for student i in year t of the study attending school s under teacher treatment assignment vector $\mathbf{T}$. Let $D_s(\mathbf{T})$ be the

assignment of school s (Clique, Anticlique, or Pure Control). For brevity we will sometimes suppress the argument $\mathbf{T}$.

In our framework, potential outcomes are fixed and treatment assignment is random. This means that expectations are taken over all treatment assignments and not over any sampling procedure. This requires Assumption 1 to hold, i.e. that treatment assignment does not change who is observed.

The following sections describe our estimands and hypotheses. These are subscripted according to the following convention. The first part of the subscript corresponds to the research question that the estimand or hypothesis is meant to address, and the second part of the subscript is a unique identifying number. So $\Theta_{Q2.2}$ is a treatment effect that addresses question Q2. Clicking on the first part of the subscript will take the reader to the specification of the research question. Clicking on the second part of the subscript will take the reader to the definition of the estimand.

5.3 ITT Effects on Student Exam Scores

The following treatment effects attempt to address the main research questions Q1 and Q2 by studying treatment effects on student exam scores Y_{its}. We will study effects on both our own assessment that we give to the students described in Section 4.2 and the UCE exam described in Section 4.3.

All of these are intent-to-treat effects, meaning that they are the causal effects of receiving an invitation to train, not the effect of attending training. Recall that in our framework treatment assignment is random, but under Assumption 1 the sam-

ple is fixed and non-random. So Assumption 1 guarantees that all of the treatment effects in this section are identified even when some units from the randomization do not appear in the sample. Take note however that these are effects only on the (fixed) sample of units who appear in our data.

The estimand $\Theta_{Q1.1}$ captures the school-level causal effect of receiving training on student test scores. It most closely resembles the effect estimated by Nourani et al. (2023). It compares assessment scores across treatment vs control schools within a single year t.[10]

$$\Theta_{Q1.1} \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Y_{its} \mid D_s \neq \text{CTL}\right] - \mathbb{E}\left[Y_{its} \mid D_s = \text{CTL}\right] \tag{1}$$

The next estimand studies teacher-to-teacher knowledge spillovers. Estimand $\Theta_{Q2.2}$ leverages the special design of the experiment to compare schools where teachers who trained with a friend (Clique schools) to schools where those who did not (Anticlique schools) within a given year t. If we find that $\Theta_{Q2.2} \neq 0$ then we will have found evidence of teacher-to-teacher spillover effects on students without needing to link students to teachers—a difficult data task.

$$\Theta_{Q2.2} \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Y_{its} \mid D_s = \text{CLIQUE}\right] - \mathbb{E}\left[Y_{its} \mid D_s = \text{ANTI}\right] \tag{2}$$

The third estimand $\Theta_{Q1.3}$ captures the school-level causal effect of receiving training on student test scores for female students only. It compares girls' scores

[10]The expectation is over randomized teacher invitations within each school, not over a sampling process.

across treatment vs control schools within a single year t.

$$\Theta_{Q1.3} \equiv \frac{\sum_{i=1}^{n} \mathbf{1}\left\{\text{FEMALE}_i\right\}\left(\mathbb{E}\left[Y_{its} \mid D_s \neq \text{CTL}\right] - \mathbb{E}\left[Y_{its} \mid D_s = \text{CTL}\right]\right)}{\sum_{i=1}^{n} \mathbf{1}\left\{\text{FEMALE}_i\right\}} \tag{3}$$

5.4 ITT Effects on Teacher Survey Responses

The next three estimands study the effect of invitation to training on teacher attitudes towards the purpose of secondary education for boys vs girls—research question Q3. Let Z_{itsu} be a dummy variable indicating that teacher i at school s responded with agreement to the gender statement u on the teacher survey in year t. The survey questions are described in Section 4.4.

The estimand $\Theta_{Q3.4.u}$ compares the rate of agreement among invited teachers to teachers in pure control schools. By excluding uninvited teachers in treatment schools we avoid diluting the treatment effect with spillovers.

$$\Theta_{Q3.4.u} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Z_{itsu} \mid T_{jts} = 1\right] - \mathbb{E}\left[Z_{itsu} \mid D_s = \text{CTL}\right] \tag{4}$$

The estimand $\Theta_{Q3.5.u}$ compares average agreement rates for question u for invited teachers across Clique and Anticlique schools. This captures the effect of training with friends vs training with co-workers who are not friends.

$$\Theta_{Q3.5.u} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Z_{itsu} \mid D_s = \text{CLIQUE} \cap T_{jts} = 1\right] - \mathbb{E}\left[Z_{itsu} \mid D_s = \text{ANTI} \cap T_{jts} = 1\right] \tag{5}$$

The estimand $\Theta_{Q3.6.u}$ compares average agreement rates for question u by unin-

vited teachers across treatment and pure control schools. This comparison captures the effect of being exposed to trained colleagues on teachers who were not themselves invited to train.

$$\Theta_{Q3.6.u} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Z_{itsu} \mid D_s \neq \text{CTL} \cap T_{jts} = 0\right] - \mathbb{E}\left[Z_{itsu} \mid D_s = \text{CTL}\right] \tag{6}$$

5.5 Estimation

We estimate Θ in Sections 5.3-5.4 with the regression-adjusted estimator from Wooldridge (2010) on pp. 930-931. The relevant STATA command is: "teffects ipwra." We describe here the estimation procedure for $\Theta_{Q2.2}$, but the procedure is nearly identical for all of the estimands. There are two steps. In the first step we run the OLS regression in Equation 7. No weights are necessary because the true propensity scores are all constant.

$$Y_{its} = 1\left\{D_s = \text{CLIQUE}\right\} X_{is}\beta_1 + 1\left\{D_s \neq \text{CLIQUE}\right\} X_{is}\beta_0 + \epsilon_{its} \tag{7}$$

To increase precision, we include the $k \times 1$ vector of baseline covariates X_{is}. These always include a constant, pre-baseline UCE exam scores meaned within each school, student (or teacher) age, student (or teacher) gender, and (when estimating effects on students) student cohort indicators. When we estimate effects on UCE scores we also include student PLE scores as an additional covariate. The regression in Equation 7 predicts outcomes using the covariates separately among students in Clique schools and those in Anticlique schools. The regression yields

the two $k \times 1$ vectors coefficients $\hat{\beta}_1, \hat{\beta}_0$.

In the third step, we compute the final estimate by taking the difference in the predicted values of Y_{its} under the two treatment conditions.

$$\hat{\Theta}_{Q2.2} = \frac{1}{n} \sum_{i=1}^{n} X_{is} \left(\hat{\beta}_1 - \hat{\beta}_0 \right)$$

In order for $\hat{\Theta}_{Q2.2}$ to be consistent for $\Theta_{Q2.2}$, we must verify the overlap and unconfoundedness conditions. The experimental design guarantees that every student and teacher has a chance to be in any group that the estimands compare, so overlap is satisfied. Since all treatments were randomly assigned, if we could observe data from every teacher and student at every school, unconfoundedness would hold whenever we condition on covariates pre-determined at baseline. Since we only observe a sample of teachers and students, we need Assumption 1 to hold in order to have unconfoundedness.

We use the regression-adjusted estimator because it has the double robustness property. Since we know the true treatment probabilities of each student, school, and teacher, our propensity score model is true by design. This means that we can adjust for any covariates using any model so long as the covariates are determined prior to randomization and Assumption 1 holds. We choose to specify a linear model for the outcomes in Equation 7. Even if the linear model is false, we can still consistently estimate all of our estimands because of double robustness.

Adjusting for covariates is vital in our case because they greatly increase precision by explaining much of the variation in our outcome variables. For example,

student PLE scores from primary school explain 44% of the UCE exams scores taken in secondary school. The doubly robust estimator allows us to exploit these powerful covariates without imposing any additional assumptions on how outcomes are determined.

5.6 Primary Hypotheses

While we will test many hypotheses, we specify a small number of "primary" tests for each of our three research questions from Section 2. These tests are enumerated in Table 12. All primary tests will be tested exclusively using outcome data from 2024—which has not been analyzed at the time that this report is written. The table makes explicit the sharp null hypothesis being tested, provides an interpretation in plain English, and notes the test statistic to be used. We do not adjust for multiple hypothesis testing across the distinct research questions because they are of separate interest.

The first three primary tests in Table 12 check for effects of training on student exam scores. The primary null hypotheses for Q1 are (i) that school treatment status will have no effect on any student's score on the 2024 UCE exam and (ii) that school treatment status had no effect on any female student's score on the 2024 UCE exam. The primary null hypothesis for Q2 is that school assignment to Clique vs Anticlique will have no effect on any student's score on the 2024 UCE exam.

The last three primary tests in Table 12 consider effects on teacher attitudes toward the purpose of education for girls and boys which is research question Q3.

First we test the sharp null hypothesis that receiving an invitation to training did not affect any teacher's responses to any of the eight survey questions in our the module from Section 4.4. Next we test the sharp null hypothesis that no teacher's responses to the survey module concerning the purpose of education for boys and girls were affected by whether they were invited to train alongside friends or not. Finally we test the sharp null hypothesis that no uninvited teacher's responses to the module were affected by whether any other teachers at their school had been invited to train.

5.7 Hypothesis Testing Methods

Next we describe how each of these sharp null hypotheses will be tested. We will not use t-tests for inference because the number of treatment clusters is small. While we report robust standard errors clustered at the school level for all estimates $\hat{\Theta}$ discussed above, these standard errors are unreliable because the number of clusters (38 or 39 schools) is small.

Our solution is to exploit the randomized design to use a Fisherian exact test known to economists as "randomization inference" (Young, 2018). This procedure only tests the "sharp" null hypotheses under which all outcomes used to construct a particular test statistic are known. Table 12 explicitly lists each sharp null that is being tested. The validity of Fisher's test requires only Assumption 1, i.e. that treatment never causes (non)attrition from the sample. Fisher's test is exact and requires no other assumptions or asymptotic approximations and will control the type-I error rate regardless of the number or size of the treatment clusters.

Table 12 lists the statistics that we will use to test each research question. For the three research questions concerning student outcomes Q1 and Q2 we specify the test statistics to be the point estimates of the treatment effects, $\hat{\Theta}_{Q1.1}$, $\hat{\Theta}_{Q2.2}$ and $\hat{\Theta}_{Q1.3}$ respectively.

The research question Q3 considers several outcome variables each. To produce a single test statistic per hypothesis, we use the following three F-scores. Define the statistic $F_{Q3.8}$ in the following way. First run the OLS regression in Equation 8 below. Exclude all teachers in treatment schools who had not yet been invited to train in year t. Let T_{jts} indicate that teacher j at school s had been invited to train by year t. Let Z_{jstu} indicate that the teacher agreed with the statement in question u. Include that teacher's baseline age and gender, as well as school s mean UCE score prior to randomization as covariates. Define $F_{Q3.8}$ as the F-score from the usual F-test of the null hypothesis that every γ_u coefficient is equal to zero where standard errors were clustered at the school level. This F-score will be large when teacher gender question responses predict whether the teacher had been invited to training or was in a pure control school.

$$F_{Q3.8}: \quad T_{jts} = \beta_0 + \beta_1 \text{AGE}_{js} + \beta_2 \text{MALE}_{js} + \beta_3 \text{UCE}_s + \sum_u \gamma_u Z_{jstu} + \epsilon_{jts} \tag{8}$$

To test the null hypothesis that clique and anticlique schools shifted attitudes differently, define $F_{Q3.9}$ as the F-score from the usual F-test of the null hypothesis that every γ_u coefficient is equal to zero in Equation 9 below. Teachers in pure control schools are excluded. This F-score will be large when gender question answers

predict whether the teacher was invited to train in a Clique or an anticlique.

$$F_{Q3.9}: \quad \mathbf{1}\{D_s = \text{CLIQUE}\} = \beta_0 + \beta_1 \text{AGE}_{js} + \beta_2 \text{MALE}_{js} + \beta_3 \text{UCE}_s + \sum_u \gamma_u Z_{jstu} + \epsilon_{jts} \tag{9}$$

Finally, to test whether teacher attitudes changed when their colleagues were invited to training, define $F_{Q3.10}$ as the F-score from the usual F-test of the null hypothesis that every γ_u coefficient is equal to zero in Equation 10 below. Teachers who were invited to training are excluded. This F-score will be large when gender question answers among teachers who were never invited to train nevertheless predict whether their school was assigned to pure control.

$$F_{Q3.10}: \quad \mathbf{1}\{D_s \neq \text{CTL}\} = \beta_0 + \beta_1 \text{AGE}_{js} + \beta_2 \text{MALE}_{js} + \beta_3 \text{UCE}_s + \sum_u \gamma_u Z_{jstu} + \epsilon_{jts} \tag{10}$$

6 Midline Results and Endline Power

We will test the primary hypotheses in Table 12 at endline once the 2024 teacher survey and student exam scores are available. To demonstrate that the testing and estimation methods are fully specified, we present here results using midline data collected at the end of 2022 and 2023. We detect no midline effect on student exam scores, likely because student exposure to trained teachers at midline was still low. We do detect midline effects on teacher gender attitudes. This effect is significantly larger in Clique schools, implying teacher-to-teacher spillovers. Midline results are not used to draw final conclusions about the intervention, but they do provide a

full set of specifications for endline.

6.1 Effects on Student Exam Scores

Table 13 shows estimates of $\Theta_{Q1.1}$ which compares student exam scores in treated to untreated schools. Table 14 shows estimates of $\Theta_{Q2.2}$ which compares student exam scores in Clique to Anticlique schools. Table 15 shows $\Theta_{Q1.2}$ which is the treatment effects for girls. The columns of each table show the point estimate, cluster-robust standard error, and the p-value of the Fisherian exact test that uses the point estimate as the test statistic. The top row of each table uses data from our 2022 student assessment described in Section 4.2 and the bottom row uses the 2022 UCE exam scores described in Section 4.3. While the point estimates are nearly always positive, Fisher's test does not reject any of the three null hypothesis on either of the two exams.

The null midline effect on exam scores is unsurprising for two reasons. First, Nourani et al. (2023) also did not find program effects on exam scores after only one year. Second, when students took our assessment in 2022, the first cohort of teachers were still going through Kimanya training.

We expect larger effects on exam scores at endline. By the end of 2024, three times as many teachers will have been trained as 2022, the treated teachers will have been trained for longer, and therefore more students will have had more years of exposure to trained teachers. We calculate power to detect a larger effect in the following way. We simulate endline samples by resampling the midline data with replacement within each school and adding a hypothetical treatment effect to the

resampled scores. Power is calculated as the fraction of bootstrapped samples in which our tests reject their sharp null hypotheses at the 5% level. Resampling in this way overstates within-cluster correlations and is likely to understate true power.

The right panel of Table 13 shows the results of the power calculations. Here the null hypothesis is that school assignment to treatment or pure control had no effect on any exam score. We calculate at least 92% power to detect an endline effect .38 standard deviations larger than midline on our assessment and an effect .24 standard deviations larger than midline on the UCE exam.

The right pane of Table 14 shows power to reject the sharp null hypothesis that Clique and Anticlique treatment have the same effect on exam scores for all students. We find 90% power to detect a Clique effect of .38 standard deviations on scores for our student assessment and 95% power to detect an effect of .19 standard deviations on the UCE scores.

The right pane of Table 15 shows power to reject the sharp null hypothesis that treatment affected girls' exam scores. We find at 93% power to detect a Clique effect of .25 standard deviations on scores for our student assessment and 95% power to detect an effect of .39 standard deviations on the UCE scores. We conclude that fir girls' exams, our assessment is much more likely to detect an effect than for UCE exams.

These detectable effect sizes are significantly smaller than the effects found by Nourani et al. (2023) of 0.5 standard deviations on high-stakes exam scores. Moreover these power calculations are likely to understate true statistical power

because resampling within schools overstates within-cluster dependence for small schools. Finally, these effect sizes, while still large, are detectable with very high probability.

6.2 Effects on Teacher Attitudes

We do find significant effects at midline on teacher attitudes towards the purpose of education for boys vs girls. Table 16 shows estimates of $\Theta_{Q3.4}$, which compares invited teachers to pure control teachers. While Fisher's test fails to reject the null hypothesis for any individual survey question, the joint null hypothesis of no effect on any question is rejected with $p = .01$. We are interested to see whether the endline survey yields more insight but not much can be said based on the midline data.

Table 17 shows much stronger effects. Here we estimate $\Theta_{Q3.5}$, which compares invited teachers in Clique schools to invited teachers in Anticlique schools. The joint null hypothesis of no effect on any of the eight survey questions is rejected with Fisher's $p = .02$. In particular, we find that trained teachers from Clique schools are significantly less likely to agree with each of the following three statements:

- Boys need more education because they are better suited than girls to find high-paying jobs. ($p = .00$)
- Girls should attain higher education so that they find better husbands. ($p = .00$)

- Girls need more education because they are better suited than boys to find high-paying jobs. ($p = .04$)

The fact that midline effects on Clique vs Anticlique invited teachers are larger than effects on invited vs pure control teachers leads us to hypothesize that training with a friend is more effective than not training with a friend. We look forward to confirming this result with the endline survey.

We do not find evidence of spillovers onto untrained teachers at midline. Table 18 shows estimates of $\Theta_{Q3.6}$ which compares uninvited teachers in treatment vs control schools. Neither the joint test nor the individual tests reject their sharp null hypotheses of no spillover effect of having trained colleagues on untrained teachers. Since the joint p-value of $.12$ is relatively low, we are interested to find out whether spillover effects appear at endline when more teachers are trained.

We wish to calculate power to reject the three joint null hypotheses in Table 12 at endline. In contrast to students, we did not collect a new sample of teachers in 2024 but rather re-surveyed the same set of teachers as before. To calculate power, we ask two questions. First, how much would teacher responses have to change in order for us to fail to detect effects on already seen at midline? Second, how much would the spillover effects on untrained teachers in 2024 have to be for us to detect it at midline? In other words, we run power calculations that condition upon midline survey responses.

Power calculations suggest that the finding in Table 16 might be difficult to replicate. In a simulation we randomly chose 15% of teacher-questions and changed their survey responses at random. Then for two survey questions, we changed 35%

of "Agrees" among teachers trained in or before 2024 to "Disagrees". In 76% of simulations, Fisher's test using $F_{Q3.8}$ still rejected the null. This suggests that if even a modest amount of noise is added to teacher responses, then the program effect in year 3 would need to be very strong for us to replicate the p-value at the bottom of Table 16.

In contrast, conditional power calculations suggest that we are very likely to replicate the rejection of the joint null hypothesis in Table 17 even if survey responses are noisy. In a simulation we randomly chose 15% of teacher-questions and changed their survey responses at random. But here we did not add any program effects at all for teachers newly trained in 2024. Fisher's test still rejected the joint null hypothesis in 98% of simulations. This suggests that a contrast between Clique and Anticlique schools is likely to hold up at endline—even if the endline survey is noisy and third year program effects are small.

While Table 18 does not show midline effects on uninvited teachers, we wish to know what magnitude of effect is likely to be detectable at endline. A simulation exercise showed that if every teacher-question answered by a teacher still uninvited in 2024 has a 40% chance to change from agree to disagree, then Fisher's test using $F_{Q3.10}$ only rejects the sharp null hypothesis 66% of the time. This is because only one quarter of teachers in treatment schools will remain uninvited in 2024. Although the true spillover effect is likely highest when most teachers are trained, the sample used to detect that spillover is at its smallest when most teachers are trained. Power to detect an endline effect on untrained teachers will depend on whether the spillover has become large by that time.

7 Measuring Purpose and Alignment

This section describes the "exploratory" portion of the study. While the previous sections pre-specify confirmatory analyses that test whether KN training had impact at all, this section signposts an exploratory process that aims to discover *why* and *how* KN training works. First we sketch an initial theory that guides our own inquiry and is also likely to evolve.

7.1 Theory of Change

Our conversations with teachers and KN tutors suggest that trained teachers transition from a traditional "banking" pedagogical approach to a new "capabilities" approach over several school years. Here we describe an initial theory of this transition that we will update and revise as the study progresses and endline data is analyzed.

We posit that most teachers who enter training initially espouse what we call a *banking* pedagogical approach. A banking approach emphasizes information transfer from teacher to student and rote memorization of facts as the main learning objective. We hypothesize that KN training gradually transitions teachers to a *capabilities* approach that helps teachers understand how a diverse set of learning objectives can advance students' abilities to act intentionally across the diverse contexts where knowledge is ultimately used in action. Examples where knowledge and action converge include: using precise and language in dialogue with others, articulating questions that advance inquiry associated with a relevant domain

(e.g., agricultural activity, social well-being, environmental issues, etc.), advancing experiments and activities that generate new insights into said questions, disciplining observations by distinguishing between objective and subjective observation statements, using concepts to advance analogical thinking, and placing activities into contexts that consider their meaning to society (beyond the goal of individual learning).

Transition from a banking to a capabilities approach requires both a shift in the purpose with which teachers approach education as well as a change in the activities which they implement to realize that purpose. KN training creates an environment that facilitates these shifts. As the teacher experiences new approaches to learning during training and increasingly invests in the process of *learning to learn*, they resolve to change their own practices accordingly, which may initially be a re-formulation of their own sense of the purpose and pedagogical objectives.

Teachers may try to mimic the same educational methods used by their tutors, in conjunction with the *Preparation for Social Action* texts, during training. This may drive an uneven transition process where teachers' goals advance faster than their own actions — or that teachers mechanically change their actions before considering the goals of their actions — and the degree of support in their environment. We posit that teachers who are *learning to learn* are working towards bringing their actions and purposes into alignment. Misalignment is a state in which the actions of the teacher are inconsistent with the teacher's objectives, while alignment is a state in which the actions and objectives are matched. We hypothesize that performance, potentially measured in student learning outcomes,

is positively correlated with alignment.

We posit three major types of alignment. We define *internal alignment* as a teachers' ability to instantiate the purpose with which they approach education through pedagogical practice. For instance, an aligned "banking" teacher may spend a vast majority of their time lecturing or dictating notes for their students to write down in their notebooks. This may serve the purpose of improving students' ability to memorize facts, but not to place these facts in a broader context that serves society, for example.

Meanwhile, *meta-alignment* describes the institutional forces that may impose a set of activities, associated with a specific purpose, on the teacher. For example, this is manifested in our context under two curriculum regimes: under one regime teachers are asked by the Ministry of Education to teach within a "competency-based" educational framework with objectives similar to a capabilities framework, whereas the second regime reflects a more traditional, "banking", approach. Finally, the third major type of alignment is what we term *external alignment*. This capture the degree to which the purpose and practice perceived by others, for example the teachers' students, is aligned with that of the teacher. Internal alignment is a necessary condition for external alignment.

Teachers who are transitioning from a "banking" to a "capabilities" approach may initially enter a state of misalignment as they gain experience with the capabilities approach. This misalignment could reduce a teacher's effectiveness in enhancing student learning as they experiment with new methods. Eventually, however, a teacher who is critically reflective of their practice — and has an increasingly clear

understanding of the practices and objectives of a capabilities approach — will be more likely to overcome misalignment and move towards a capabilities approach through a process of learning (by doing and from others). While we do not elaborate much here, we acknowledge that the rate of transition towards a state of capabilities-alignment is influenced by the meta-alignment of the institution they are acting within in addition to other factors.

Our exploratory analysis will help us understand each type of alignment. The following three subsections discuss each type of alignment one at a time. Each subsection (i) lists our initial hypotheses, (ii) describes the relevant data that we are collecting, and (iii) discusses broadly how our exploratory analyses will help us learn about that type of alignment.

7.2 Internal Alignment

7.2.1 Exploratory Hypotheses

1. Teachers in the treatment group will demonstrate higher internal alignment than those in the control group, particularly in aligning capability-building purposes with corresponding activities.

2. The increased alignment between purpose and practice may take some time to emerge, with teachers developing the capacity

3. Treated teachers will also become better able to be aligned inter-temporally. That is, to be able to come up with a plan for pedagogical activities which aligns with their purpose, and then be able to follow through with that plan

in the face of challenges.

7.2.2 Data Collection

To capture the notion of internal alignment, we ask teachers to describe their perceived purpose of education broadly and reflect on a particular lesson they have recently taught.

- We ask teachers what they perceive to be the role of education and their role as a teacher specifically, from a menu of options. For example, they may answer “to impart or give new information to my students”, “to help my students pass”

- We then ask teachers about a *specific* lesson which they recently gave. We ask them to list the specific activities which they included as part of the lesson (e.g. lecturing using blackboard, breaking class into small groups for activities) as well as the purpose of that particular lesson, giving us a measure of internal alignment at a particular point in time.

- We also ask the teachers how, if at all, the lessons outcomes deviated from the objectives in the plan. We follow up by asking what they would change in order to move towards alignment. This gives a measure of intertemporal alignment.

7.2.3 Analytical Approach

Our aim is to explore how teachers' broad educational philosophies translate into specific classroom practices and how these relationships evolve across treatment conditions. We hope to capture the following notions through this analysis:

1. Purpose-practice consistency: quantifies how well teachers currently align their stated purposes with their chosen in activities.

2. Intertemporal Alignment: The degree to which teachers are able to follow-through on lesson plans made before class

For each lesson, we can construct a matrix where rows represent the different purposes (e.g. "fostering critical thinking") and columns represent the various activities (e.g., "lecturing using the blackboard," "group activities"). The entries in the matrix represent the proportion of total points allocated to each purpose-activity pair.

More simply, we will attempt to classify each of the stated purposes and practices as "capacity-building" or "information-delivery", as well as measuring the extent to which teachers align the purpose of the lesson with the activities contained therein.

7.3 Meta-Alignment

7.3.1 Exploratory Hypotheses

1. Teachers who undergo the training may feel more empowered to discuss pedagogical practices with others, conditional on an open school/institutional environment

2. Teachers may then become more influential within their social networks on other teachers pedagogical practices as well as on school decision-making more broadly

3. The curriculum regimes create differences in the meta-alignment for different teachers, which will affect rates of change towards internal alignment. E.g., a trained teacher's rate of change may be faster under the "competency-based curriculum" than the traditional curriculum.

7.3.2 Data Collection

As previously mentioned, we collect detailed social network data which allows us to reconstruct the entire teacher social network in each of our schools. We also collect various measures of teacher influence and types of interactions across teachers. We supplement this with data on school decision-making processes to get a measure of the meta-alignment which captures the extent to which the teacher's environment is conducive to alternative pedagogies. Within our teacher survey we collect multiple measures which form components of this concept:

- We elicit frequency with which the teacher discusses various pedagogical practices with 6 randomly selected teachers from their school as well as the extent to which they agree that the teacher has influence over the teaching.
  - Additionally, we asked the teacher to list up to 3 further teachers who have influenced the teachers teaching practices in the current and last term.
- We also ask teachers about school-level decision-making, including policies that affect teaching practices and the manner in which these policies are decided (e.g. through discussion, through decree by superiors).
- We ask whether teachers believe that their department shares the same objectives with respect to education as them.
- Finally, we ask teachers about the various motivating and demotivating factors within their job as a teacher, including ”cultural” factors such as relationships with coworkers and headteachers.

7.3.3 Analytical Approach

We seek to explore the following components of meta-alignment:

1. The extent to which the purpose and practice of each teacher is aligned with those of their colleagues and key school stakeholders
2. Conditional on (1), whether and how the actions undertaken by teachers whose internal alignment is more or less “meta-aligned” differs.

We can construct a school-level measures of educational purpose, practice, and school decision-making which we can then compare to teachers' individual analogues to these school-wide measures and thus observe both changes in meta-alignment across time as well as behaviors in response to different levels of alignment. Additionally, we hope to study whether trained teachers are able to engender any changes in school decision-making processes, as well as the degree to which non-treated teachers in treated schools change their teaching practice based on the number and types of interactions with trained teachers. We also hope to understand whether particular members/components of the school stakeholders are more relevant than others on teachers' perception of their meta-alignment with the broader school.

7.4 External Alignment

7.4.1 Exploratory Hypotheses

1. Treated teachers will be better able to articulate the purpose with which they approach their lessons to their students, resulting in a higher degree of alignment between teacher and student

2. These changes may be most clearly identified after more time has passed, allowing teachers to learn from and reflect on their actions, and in environments where meta-alignment is higher.

7.4.2 Data Collection

We ask students to reflect on the a lesson taught by their teacher, similar to the questions asked to teachers. This serves the dual purpose of validating teacher responses on the classroom activities, but also provides a measure of the external alignment between teacher and student on the perceived purpose with which the teacher approaches each lesson, and that which is understood by the student. Similarly to the teacher survey, we ask students about the following:

- Their perceived purpose of education (e.g. ”To help students learn to cooperate with others.” or ”Prepare students for success in the economy when they graduate.”)
- We also prompt students to recall a particular lesson by their teacher, the same lesson we ask about in the teacher survey. We then ask them:
    1. The activities which the teacher engaged in during that lesson
    2. The purpose that they perceived the teacher approached that lesson with

7.4.3 Analytical Approach

We can measure the degree to which students are aligned with their teacher in terms of the purpose of each particular lesson, their perceived actions from that same lesson, and their general understanding of the purpose of education which each provide a few measures of external alignment. We can then combine these individual measures of external alignment to generate an external alignment scale which measures the degree to which teachers are aligned with their students on

the broad purpose of education as well as the specific purpose of a particular lesson and the actions therein.

8 Research Team

The lead principal investigator is Vesall Nourani. Nourani is responsible for coordinating the study and building relations with Kimanya, the Echidna Foundation, and other partners. Moustafa El-Kashlan and Sara Restrepo-Tamayo lead data collection and survey instrument design. Stefan Faridani developed the experimental design and the quantitative methods in Section 5. All four will contribute as co-authors to writing academic papers and other deliverables. We are greatful for the field and data assistance given by Faith Babirye and Jiya Nair's research coordination.

9 Deliverables

We expect the data and analysis from this experiment to result in two articles published in academic journals. The authors for both articles will be El-Kashlan, Faridani, Nourani, and Tamayo. The first article will study spillover effects of the training across teachers and students. The second article will study how training affects teachers' thoughts and behavior. In particular the second article will focus on the alignment between teacher's stated purpose for each pedagogical technique that they use and students' perceived purpose of the pedagogical techniques that

their teachers use. Both articles will consider how changes in the Ugandan national secondary school curriculum interacts with training to affect teachers practices and what they report the purpose of those practices to be.

10 Calendar

The timeline in Figure 1 shows all completed and upcoming tasks. The most important dates are the following. Baseline surveys began in May 2021 and endline surveys were completed in Fall 2024. Randomization occurred on November 29, 2021 and teacher training consisted of three waves beginning in April 2022, December 2022, and December 2023.

11 IRB

Local IRB approval was granted by The Mildmay Uganda Research and Ethics Committee on March 24, 2021 under approval number 0510-2017. Our current US IRB approval is from University of Chicago and was most recently approved on April 24, 2023 under number 23-0497. Prior to this our IRB approval was given by the London School of Economics in April 2020 with number 000622c.

References

Alan, Sule and Ipek Mumcu, "Nurturing Childhood Curiosity to Enhance Learning: Evidence from a Randomized Pedagogical Intervention," *American Economic Review*, April 2024, *114* (4), 1173–1210.

Businge, Conan, "What is the New National Teachers' Policy," *The New Vision*, 2019.

de Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers, "Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia*," *The Quarterly Journal of Economics*, 11 2017, *133* (2), 993–1039.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin, "Many Children Left Behind? Textbooks and Test Scores in Kenya," *American Economic Journal: Applied Economics*, January 2009, *1* (1), 112–35.

Kremer, Michael, Conner Brannen, and Rachel Glennerster, "The Challenge of Education and Learning in the Developing World," *Science*, 2013, *340* (6130), 297–300.

McEwan, Patrick J., "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments," *Review of Educational Research*, 2015, *85* (3), 353–394.

Najjumba, Innocent Mulindwa and Jeffery H. Marshall, "Improving Learning in Uganda Vol. II: Problematic Curriculum Areas and Teacher Effectiveness: Insights from National Assessments," Technical Report, World Bank 2013.

Nourani, Vesall, Nava Ashraf, and Abhijit Banerjee, "Learning to Teach by Learning to Learn," *Working Paper*, 2023.

Popova, Anna, David K Evans, Mary E Breeding, and Violeta Arancibia, "Teacher Professional Development around the World: The Gap between Evidence and Practice," *The World Bank Research Observer*, 06 2021, *37* (1), 107–136.

Rooke, Barnaby, "Teacher Issues in Uganda: A shared vision for an effective teachers policy," Technical Report, Ugandan Ministry of Education and Sport, UNESCO 2014.

Vasiliev, Kirill and Angela Demas, "Uganda Secondary Education Expansion Project (P166570)," Technical Report, The World Bank 2018.

Wooldridge, Jefrey M., *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, 2010.

World Bank, "Lower secondary completion rate. Series SE.SEC.CMPT.LO.ZS," 2024.

Young, Alwyn, "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results*," *The Quarterly Journal of Economics*, 11 2018, *134* (2), 557–598.

12 TIMSS Citation

Several questions on our secondary school assessment described in Seciton 4.2 used modified versions of questions from the TIMSS 2011 exam. While these TIMSS questions are in the public domain, it was requrested that we include the following special citation:

SOURCE: TIMSS 2011 Assessment. Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, Chestnut Hill, MA and International Association for the Evaluation of Educational Achievement (IEA), IEA Secretariat, Amsterdam, the Netherlands

13 Figures

Figure 1: Study Timeline

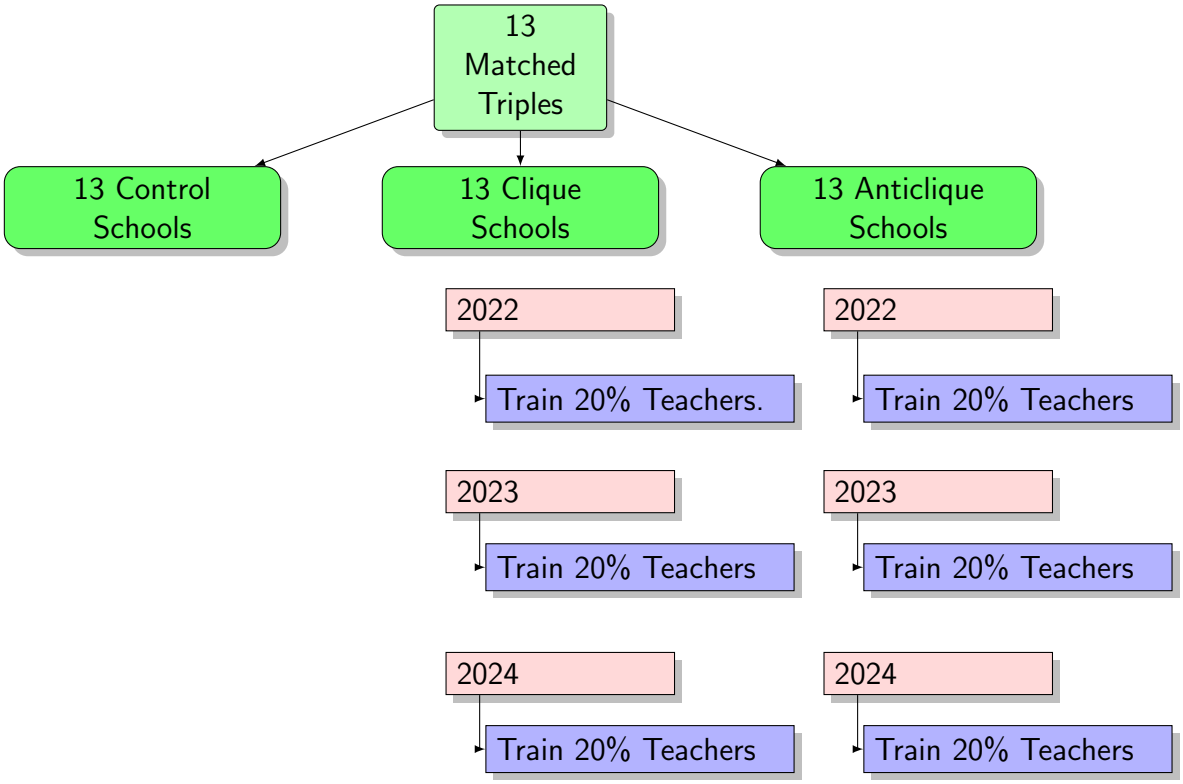

Figure 2: School-Level Design Flowchart.

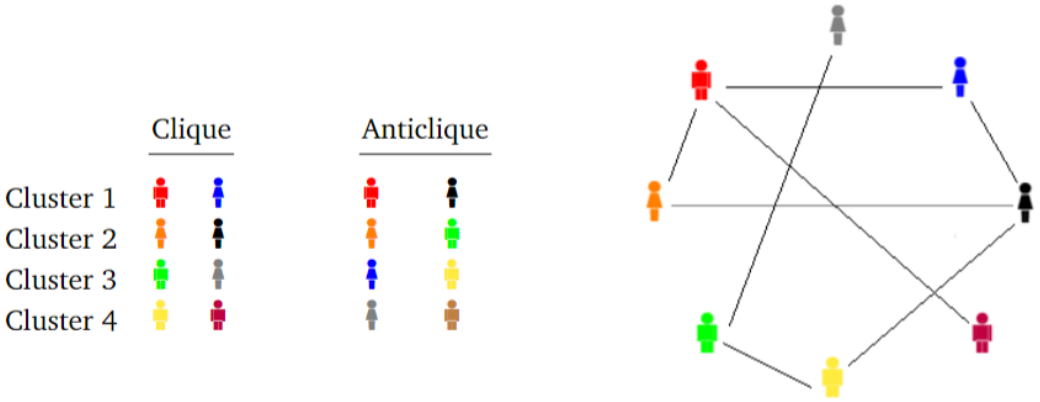

Figure 3: Example clustering of teachers In a school with cluster size two (the minimum and modal arrangement). The left panel shows the clustering scheme and the right panel shows the social network. Teachers are sorted into four clusters of two teacher each. If the school is assigned to Clique treatment, clusters almost always contain a link. If the school is assigned to Anticlique treatment, clusters almost never contain a link. One cluster will be assigned at random to each of four assignments: 2022 training, 2023 training, 2024 training, and control.

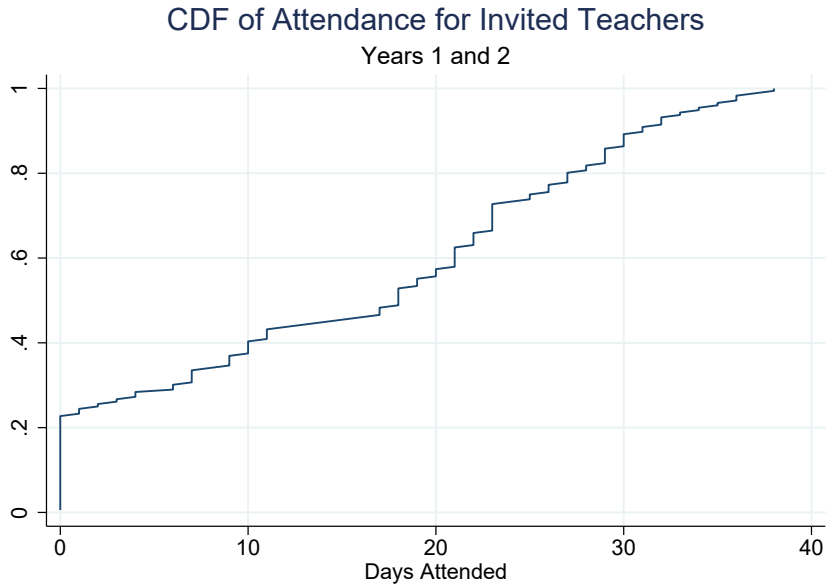

Figure 4: CDF of Teacher Attendance by end of 2023

2. Choose one of the following option choices to fill each of the sentences **a** to **d** that follow:

OPTION CHOICES: ***science, technology, both, neither.***

a. __science__ helps us discover how diseases operate.

b. Medicine is a __science(full-credit)/both(half-credit)__ that helps us prevent disease.

c. Vaccines are a _technology____ that help us prevent diseases.

d. The __science___ of nutrition can help us maintain good health.

Figure 5: Example of questions from the Use of Language section of the 2022 student assessment with correct answers highlighted in yellow.

1. Susie has a potted plant. She sets up an experiment that shows that water travels through a plant into the air.

Which experiment would show this?[☑ the one correct answer]

☐ **A.** Put water in a container under the pot; water will disappear from the container.

☐ **B.** Cover one of the stems of the plant with a plastic bag and water the plant; drops of water will be seen in the bag.

☐ **C.** Place a cut stem from the plant in a plastic bag; water will be seen in the bag.

☐ **D.** Place a cut stem from the plant in a glass of colored water; the plant's leaves will change color.

Figure 6: Example of questions from the Critical Thinking section of the 2022 student assessment. The correct answer is B.

14 Tables

	Compliance by end 2022			Compliance by end 2023		
	Not Invited '22	Invited '22	Total	Not Invited '23	Invited '23	Total
$<$ 3 days	549	24	573	447	45	492
$\geq$ 3 days	6	61	67	17	131	148
$<$ 20 days	549	42	591	459	97	556
$\geq$ 20 days	6	43	49	5	79	84
Total	555	84	640	464	176	640

Table 1: Compliance Tabulations. Columns divide teachers by intention to treat in 2022 (left panel) and 2023 (right panel). Rows show number of teachers who actually attended more than 3 days (or 20 days) of Kimanya training. Out of 33 invited teachers whom headteachers told us would not attend at all in 2022 or 2023, for 72% the reason given was that the invited teacher had left the school.

Demographics of Invitees by Compliance

	(Attended $<$ 3 days)	(Attended $\geq$ 3 days)	(Diff)
Teacher Age	40.233	38.714	-1.518
	(10.851)	(10.388)	(2.094)
Teacher Male	0.814	0.683	-0.131*
	(0.394)	(0.467)	(0.066)
Observations	45	131	176

Table 2: Differences between non-attendees (left) and attendees (right).

Baseline Selection of Teachers

	(Not)	(In Randomization)	(Difference)
Teacher Age	37.332	38.689	1.357**
	(9.976)	(10.045)	(0.590)
Teacher Male	0.761	0.678	-0.083***
	(0.427)	(0.468)	(0.026)
Teacher Married	0.714	0.769	0.055**
	(0.452)	(0.422)	(0.026)
Teacher holds Undergraduate	0.771	0.789	0.018
	(0.421)	(0.408)	(0.024)
Hours worked per week here	3.403	4.394	0.991***
	(2.906)	(1.496)	(0.132)
Teach at other schools?	0.189	0.106	-0.083**
	(0.393)	(0.309)	(0.039)
Teachers	524	640	1,164
Schools	53	39	54

Table 3: Summary statistics for teachers at baseline. Compares teachers not selected by our team to be part of the randomization vs teachers whom we did choose to be part of the randomization. Standard deviations are in parentheses in the first columns. Standard errors are in parentheses in the third column. Hours worked per week were low due to COVID-19. Uganda entered a partial lockdown midway through this survey.

Datasets

Dataset	Section	Unit	Year	Collected	Clean	N
UCE Exam	4.3	Student	'18-'20	✓	✓	12474
UCE Exam	4.3	Student	'22	✓	✓	4496
UCE Exam		Student	'23	✓		
UCE Exam		Student	'24			
Assessment	4.2	Student	'22	✓	✓	11495
Assessment		Student	'23	✓		
Assessment		Student	'24	✓		
Teacher Survey	4.1	Teacher	'21	✓		640
Teacher Survey	4.4	Teacher	'23	✓		455
Teacher Survey		Teacher	'24	✓		

Table 4: Summarizes datasets discussed in this plan—present and future. The second column displays the section of this document that describes the sample. The final column reports the number of observations that are non-missing and linked to a member of our randomization. Note: the '23 and '24 student assessments have been collected but not yet digitized.

School Balance at Baseline: Treated vs Pure Control

	(Control)	(Clique and Anticlique)	(Difference)
Enrolment	534.231	612.038	77.808
	(345.558)	(430.436)	(137.525)
Mean UCE '18-'20	0.064	0.098	0.034
	(0.636)	(0.415)	(0.169)
Private School	0.462	0.731	0.269
	(0.519)	(0.533)	(0.180)
No. Male Teachers	20.923	23.077	2.154
	(14.857)	(10.840)	(4.174)
No. Femaile Teachers	8.846	8.500	-0.346
	(4.776)	(5.736)	(1.849)
No. Teachers Eligible	20.692	20.692	0.000
	(14.285)	(12.985)	(4.559)
No. invited per year if treated	4.077	4.115	0.038
	(2.660)	(2.422)	(0.850)
Observations (schools)	13	26	39
RI p (Fisher's)			0.62

Table 5: School-Level Balance at baseline (November 2021). Standard deviations are reported in parentheses under group means and standard errors are reported under the differences. Fisher's exact p-value was calculated using the overall F-score of a regression of an indicator of whether a school was treated at all on the baseline characteristics in the rows.

School Balance at Baseline: Anticlique vs Clique

	(Anticlique)	(Clique)	(Difference)
Enrolment	639.385	584.692	-54.692
	(459.384)	(416.336)	(171.950)
Mean UCE '18-'20	0.125	0.070	-0.054
	(0.369)	(0.470)	(0.166)
Private School	0.846	0.615	-0.231
	(0.555)	(0.506)	(0.208)
No. Male Teachers	23.154	23.000	-0.154
	(9.848)	(12.159)	(4.340)
No. Female Teachers	8.385	8.615	0.231
	(5.810)	(5.895)	(2.296)
No. Eligible Teachers	19.231	22.154	2.923
	(11.039)	(14.994)	(5.164)
No. invited per year if treated	3.846	4.385	0.538
	(2.075)	(2.785)	(0.963)
Observations (schools)	13	13	26
RI p (Fisher's)			0.52

Table 6: Balance of school-level characteristics at baseline. Standard deviations are reported in parentheses under group means and standard errors are reported under the differences. Fisher's exact p-value was calculated using the overall F-score of a regression of an indicator of whether a school was treated at all on the baseline characteristics in the rows.

Balance of 2022 Student Assessments

	(Control)	(Clique or Anticlique)	(Difference)
Student Age	16.285	16.299	0.014
	(1.740)	(1.501)	(0.229)
Student Male	0.450	0.522	0.073
	(0.498)	(0.500)	(0.135)
S2	0.303	0.306	0.003
	(0.460)	(0.461)	(0.019)
S3	0.259	0.268	0.009
	(0.438)	(0.443)	(0.020)
S4	0.023	0.034	0.012
	(0.149)	(0.182)	(0.007)
School Mean UCE '18-'20	-0.192	0.117	0.309
	(0.753)	(0.379)	(0.303)
Observations (students)	3,774	7,721	11,495
Schools	13	25	38
RI p (Fisher's)			0.39

Table 7: Balance of baseline characteristics for students who took our 2022 student assessment. Standard deviations are reported in parentheses under group means and standard errors clustered at the school level are reported under the differences. Fisher's exact p-value was calculated using the overall F-score of a regression of an indicator of whether a school was treated at all on the baseline characteristics in the rows.

Balance for our sample of 2022 UCE Test-Takers

	(Control)	(Clique or Anticlique)	(Difference)
Student PLE Score	0.058	-0.031	-0.089
	(1.015)	(0.991)	(0.249)
Student Male	0.566	0.469	-0.097
	(0.496)	(0.499)	(0.105)
Student Age	18.995	19.003	0.008
	(1.432)	(1.376)	(0.188)
Schoom Mean UCE '18-'20	-0.065	0.176	0.241
	(0.690)	(0.382)	(0.245)
Observations	1,558	2,898	4,456
Schools	13	25	38
RI p (Fisher's)			.40

Table 8: Balance of demographic characteristics for students in our 2022 UCE exam score data. Standard deviations are reported in parentheses under group means and standard errors clustered at the school level are reported under the differences. Fisher's exact p-value was calculated using the overall F-score of a regression of an indicator of whether a school was treated at all on the baseline school characteristics in the rows. Baseline UCE scores from '18-'20 contain imputations for two schools.

Summary statistics for 2023 teacher survey

	mean	sd	min	max
School Mean UCE Score 2018-2020	-0.08	0.59	-1.43	0.87
Teacher Age	39.52	9.70	0.00	70.00
Teacher Male	0.67	0.46	0.00	1.00
Fraction agreeing with each statement:				
Boys have more ability than girls to learn science subjects	0.29	0.45	0.00	1.00
Boys need more education because they are better suited than girls to find high-paying jobs	0.18	0.39	0.00	1.00
Parents should maintain stricter control over their daughters than their sons	0.49	0.50	0.00	1.00
Girls should attain higher education so that they find better husbands	0.36	0.48	0.00	1.00
Boys should attain higher education so that they find better wives	0.27	0.44	0.00	1.00
Boys should be just as capable as girls at preparing a meal for visitors when guests visit	0.85	0.36	0.00	1.00
Girls have more ability than boys to learn science subjects	0.21	0.41	0.00	1.00
Girls need more education because they are better suited than boys to find high-paying jobs	0.26	0.44	0.00	1.00
Observations	455			

Table 9: Summary statistics for 2023 teacher survey. Mean UCE score was imputed as zero for two schools. Teacher age and gender were imputed at their means for nine teachers. No survey response was imputed.

Balance of 2023 Teacher Survey by School Treatment Status

	(Pure Control)	(Clique or Anticlique)	(Difference)
Teacher Age	40.447	39.019	-1.427
	(8.409)	(10.303)	(1.640)
Teacher Male	0.623	0.703	0.080
	(0.486)	(0.450)	(0.063)
School Mean UCE '18-'20	-0.320	0.054	0.373
	(0.787)	(0.396)	(0.342)
Observations (teachers)	159	296	455
Schools	13	26	39
RI p (Fisher's)			0.30

Table 10: Balance of characteristics among teachers who took our 2023 survey. Standard deviations are reported in parentheses under group means and standard errors clustered at the school level are reported under the differences. Fisher's exact p-value was calculated using the overall F-score of a regression of an indicator of whether a school was treated at all on the baseline characteristics in the rows.

Balance of 2023 Teacher Survey by Invitation Status

	(Not Invited)	(Invited)	(Difference)
Teacher Age	39.664	39.236	-0.428
	(9.285)	(10.476)	(1.317)
Teacher Male	0.649	0.725	0.076
	(0.475)	(0.440)	(0.054)
School Mean UCE '18-'20	-0.148	0.061	0.209
	(0.658)	(0.400)	(0.199)
Observations (teachers)	300	155	455
Schools	39	26	39
RI p (Fisher's)			0.55

Table 11: Balance of teachers who took the 2023 survey. Standard deviations are reported in parentheses under group means and standard errors clustered at the school level are reported under the differences. Fisher's exact p-value was calculated using the overall F-score of a regression of an indicator of whether a teacher had been invited in years 1 or 2 (2022 or 2023) on the baseline characteristics in the rows.

Primary Hypothesis Tests

Question	Test Stat	Interpretation	Sharp Null Hypothesis
Q1	$\hat{\Theta}_{Q1.1}$	School assignment to treatment (Clique or Anticlique) did not affect any student exam scores on the '24 UCE exam.	$Y_{i3s}(\mathbf{T}) = Y_{i3s}(\mathbf{T}')$ if $\mathbf{1}\{D_s(\mathbf{T}) = \text{CTL}\} = \mathbf{1}\{D_s(\mathbf{T}') = \text{CTL}\}$ and $\mathbf{T}, \mathbf{T}'$ have same teacher cluster assignments
Q2	$\hat{\Theta}_{Q2.2}$	Clique vs Anticlique school assignment did not affect any student exam scores on the '24 UCE exam.	$Y_{i3s}(\mathbf{T}) = Y_{i3s}(\mathbf{T}')$ if $D_s(\mathbf{T}) \neq$ CTL and $D_s(\mathbf{T}') \neq$ CTL and $\mathbf{T}, \mathbf{T}'$ have same teacher cluster assignments
Q1	$\hat{\Theta}_{Q1.3}$	School assignment to treatment did not affect any girls' exam scores on the '24 UCE exam.	$Y_{i3s}(\mathbf{T}) = Y_{i3s}(\mathbf{T}')$ if $\mathbf{1}\{D_s(\mathbf{T}) = \text{CTL}\} = \mathbf{1}\{D_s(\mathbf{T}') = \text{CTL}\}$ and $\mathbf{T}, \mathbf{T}'$ have same teacher cluster assignments for girls
Q3	$F_{Q3.8}$	An invitation to training did not affect any teacher survey responses to gender-related questions in '24.	$Z_{i3su}(\mathbf{T}) = Z_{i3su}(\mathbf{T}')$ for all $\mathbf{T}, \mathbf{T}', u$
Q3	$F_{Q3.10}$	Clique vs Anticlique assignment did not affect any survey responses to gender-related questions among invited teachers in '24.	$Z_{i3su}(\mathbf{T}) = Z_{i3su}(\mathbf{T}')$ if $D_s(\mathbf{T}) \neq$ CTL and $D_s(\mathbf{T}') \neq$ CTL and $\mathbf{T}, \mathbf{T}'$ have same teacher cluster assignments
Q3	$F_{T.Q3.10}$	Teacher training assignments did not affect teacher survey responses to any gender-related questions for uninvited teachers in '24.	$Z_{i3su}(\mathbf{T}) = Z_{i3su}(\mathbf{T}')$ if $T_{j3} = T'_{j3} = 0$

Table 12: Primary Hypothesis Tests for 2024. Every hypothesis will be tested using Fisher's Exact Test using the test statistic in the middle column. Additional followup exploratory tests will be run, but these tests are specified in advance so they are non-exploratory.

ITT effect of treatment school vs control school on 2022 exams

		Midline Estimate				Power Calc	
Estimand	Exam	N	Estimate	(se)	RI p	Effect	Pwr
$\Theta_{Q1.1}$	Ours	11495	.096	(.119)	.74	.32	92%
$\Theta_{Q1.1}$	UCE	4456	.014	(.067)	.89	.24	93%

Table 13: Midline results and endline power calculations for Q1: main effects on student exam scores. Compares exams scores of students in treatment (clique or anticlique) schools to students in pure control schools. The outcome is the student exam score expressed in standard deviations from the sample mean. Standard errors are clustered at the school level and reported in parentheses. The left panel shows the midline estimate. The ITT was estimated by regression-adjustment. The top row uses our 2022 assessment as the outcome and the bottom row uses the 2022 official UCE exam scores. When the outcome variable is our assessment, covariates included school mean baseline UCE scores and student age, gender and cohort. When the outcome variable is the UCE exam, covariates include school mean baseline UCE scores, student PLE score, and student age and gender. The right panel shows a conservative power calculation based on resampling the 2022 data with replacement within schools and adding an additional effect to treated schools

ITT effect of Clique school vs Anticlique school on 2022 exams

		Midline Estimate				Power Calc	
Estimand	Exam	N	Estimate	(se)	RI p	Effect	Pwr
$\Theta_{Q2.2}$	Ours	7721	.190	(.158)	.49	.38	90%
$\Theta_{Q2.2}$	UCE	2898	.109	(.082)	.53	.19	95%

Table 14: Midline results and endline power calculations for Q2: effect of Clique training on student exam scores. Compares scores of students in Clique schools to students in Anticlique schools. The outcome is the student exam score expressed in standard deviations from the sample mean. Standard errors are clustered at the school level and reported in parentheses. The left panel shows the midline estimate. The ITT was estimated by regression-adjustment. The top row uses our 2022 assessment as the outcome and the bottom row uses the 2022 official UCE exam scores. When the outcome variable is our assessment, covariates included school mean baseline UCE scores and student age, gender and cohort. When the outcome variable is the UCE exam, covariates include school mean baseline UCE scores, student PLE score, and student age and gender. The right panel shows a conservative power calculation based on resampling the 2022 data with replacement within schools and adding an additional effect to treated schools

ITT effect of treatment school vs control school on 2022 girls' exams

		Midline Estimate				Power Calc	
Estimand	Exam	N	Estimate	(se)	RI p	Effect	Pwr
$\Theta_{Q1.3}$	Ours	5765	-.016	(.067)	.84	.25	93%
$\Theta_{Q1.3}$	UCE	2214	.031	(.100)	.94	.39	95%

Table 15: Midline results and endline power calculations for Q1: main effects on girls' exam scores. Compares exams scores of girls in treatment (clique or anti-clique) schools to girls in pure control schools. The outcome is the exam score expressed in standard deviations from the sample mean. Standard errors are clustered at the school level and reported in parentheses. The left panel shows the midline estimate. The ITT was estimated by regression-adjustment. The top row uses our 2022 assessment as the outcome and the bottom row uses the 2022 official UCE exam scores. When the outcome variable is our assessment, covariates included school mean baseline UCE scores and student age and cohort. When the outcome variable is the UCE exam, covariates include school mean baseline UCE scores, student PLE score, and student age. The right panel shows a conservative power calculation based on resampling the 2022 data with replacement within schools and adding an additional effect to treated schools.

ITT effect of Invitation on Gender Attitudes

Estimand	Survey question	Estimate	(s.e.)	RI p
$\Theta_{Q3.4.u}$	Boys have more ability than girls to learn science subjects	.019	(.080)	.80
$\Theta_{Q3.4.u}$	Boys need more education because they are better suited than girls to find high-paying jobs	-.042	(.076)	.57
$\Theta_{Q3.4.u}$	Parents should maintain stricter control over their daughters than their sons	-.032	(.04)	.52
$\Theta_{Q3.4.u}$	Girls should attain higher education so that they find better husbands	.022	(.053)	.82
$\Theta_{Q3.4.u}$	Boys should attain higher education so that they find better wives	-.035	(.052)	.51
$\Theta_{Q3.4.u}$	Boys should be just as capable as girls at preparing a meal for visitors when guests visit	.021	(.047)	.68
$\Theta_{Q3.4.u}$	Girls have more ability than boys to learn science subjects	.131	(.073)	.10
$\Theta_{Q3.4.u}$	Girls need more education because they are better suited than boys to find high-paying jobs	-.101	(.097)	.20
$F_{Q3.8}$	Overall test of H_0: no effect on any teacher or question			.01**
N	Teachers (excluding non-invitees in treated schools)			314

Table 16: Midline results for Q3: main effects on teacher gender attitudes. Compares invited teachers vs teachers in pure control schools. Outcomes are the fraction of teachers agreeing with different gender-related statements. Covariates were teacher age, teacher gender, and school mean baseline UCE score. Standard errors clustered at the school level in parentheses. Data: midline 2023 teacher survey.

ITT effect of Clique vs Anticlique School on Gender Attitudes of Invited Teachers

Estimand	Survey question	Estimate	(s.e.)	RI p
$\Theta_{Q3.5.u}$	Boys have more ability than girls to learn science subjects	-.186	(.071)	.11
$\Theta_{Q3.5.u}$	Boys need more education because they are better suited than girls to find high-paying jobs	-.255	(.068)	.00***
$\Theta_{Q3.5.u}$	Parents should maintain stricter control over their daughters than their sons	-.121	(.071)	.10*
$\Theta_{Q3.5.u}$	Girls should attain higher education so that they find better husbands	-.278	(.076)	.00***
$\Theta_{Q3.5.u}$	Boys should attain higher education so that they find better wives	-.084	(.065)	.28
$\Theta_{Q3.5.u}$	Boys should be just as capable as girls at preparing a meal for visitors when guests come to visit	.026	(.059)	.67
$\Theta_{Q3.5.u}$	Girls have more ability than boys to learn science subjects	-.127	(.056)	.19
$\Theta_{Q3.5.u}$	Girls need more education because they are better suited than boys to find high-paying jobs	-.209	(.071)	.04**
$F_{Q3.9}$	Overall test of H_0: no Clique effect on any teacher or question			.02**
N	Teachers (excluding pure control schools)			155

Table 17: Midline results for Q3: Clique effects on teacher gender attitudes. Compares invited teachers in Clique vs Anticlique schools. Outcomes are the fraction of teachers agreeing with different gender-related statements. Standard errors clustered at the school level in parentheses. Covariates included school mean baseline UCE score and teacher age and gender. Data: 2023 teacher survey.

ITT effect of Treated vs Pure Control School on Gender Attitudes of Uninvited Teachers

Estimand	Survey question	Estimate	(s.e.)	RI p
$\Theta_{Q3.6.u}$	Boys have more ability than girls to learn science subjects	.065	(.075)	.41
$\Theta_{Q3.6.u}$	Boys need more education because they are better suited than girls to find high-paying jobs	-.074	(.055)	.28
$\Theta_{Q3.6.u}$	Parents should maintain stricter control over their daughters than their sons	-.009	(.041)	.85
$\Theta_{Q3.6.u}$	Girls should attain higher education so that they find better husbands	.115	(.081)	.16
$\Theta_{Q3.6.u}$	Boys should attain higher education so that they find better wives	-.052	(.061)	.30
$\Theta_{Q3.6.u}$	Boys should be just as capable as girls at preparing a meal for visitors when guests visit	.061	(.044)	.20
$\Theta_{Q3.6.u}$	Girls have more ability than boys to learn science subjects	.092	(.054)	.20
$\Theta_{Q3.6.u}$	Girls need more education because they are better suited than boys to find high-paying jobs	-.1	(.057)	.21
$F_{Q3.10}$	Overall test of H_0: no spillover effect on any uninvited teacher for any question			.12
N	Teachers (excluding invitees)			300

Table 18: Midline results for Q3: spillovers on teacher gender attitudes. Compares uninvited teachers in treatment vs control schools. Outcomes are the fraction of teachers agreeing with gender-related statements in second column. Standard errors clustered at the school level in parentheses. Data: 2023 teacher survey.