Evaluating a Whatsapp based Self Development Program to Enhance English Language Skills and Teacher Agency in Pakistan[1]

Pre-analysis Plan

Asad Liaqat, Meta and Beaj Education
Madeline Duhon, CEGA and University of California, Berkeley

07 July 2025

Summary: This document describes planned analysis to evaluate the impact of Beaj Education's Self-Development course on English proficiency, psychological well-being, and career aspirations among teachers in low-cost schools across Pakistan.

Appendix: Baseline survey, endline survey.

[1] We thank Fatima Miraj for excellent management of the field research. We gratefully acknowledge funding from the Agency Fund. This study received IRB approval from Pepperdine University (Protocol #: 25-04-2640) and Research and Development Solutions (Ref No. RADs/IRB-Beaj/16-05-2025/007).

Introduction

Summary

This document describes planned analysis to evaluate the impact of Beaj Education's Self-Development course on English proficiency and psychological well-being among teachers in low-cost schools across Pakistan.

Motivation

Teachers in underfunded Pakistani schools often face challenges related to bilingual instruction, and have few opportunities for professional development. Only 6% of teachers are proficient in English (British Council, 2013), even though many subjects are taught in English. Beaj Education currently supports over 5,000 teachers in low-cost schools throughout Pakistan, providing a variety of programs to support teachers and students in this context. Among them is the Self-Development course which is the focus of this study.

The intervention

Beaj Education's Self-Development course ("the course") was designed in early 2025 with the goal of improving English proficiency and psychological well-being for teachers in low-cost schools across Pakistan. The content for the course was created by experts in the field of English language instruction (with 20+ years of experience), and Leadership/Psychological Coaching (with 15 years of experience).

The course included 12 weeks of English language proficiency and psychological wellness content, split into three four-week "levels", and delivered primarily through self-paced videos. In general, each daily lesson had 6 "activities" of different types (listen and speak, watch and speak, multiple choice questions, etc.). As a secondary way of engaging with the course, all participants were assigned to a moderator-led group with up to 50 participants. All interactions within these groups took place over a group Whatsapp chat. Moderators would use these chat threads to nudge participants to engage, and some participants would share what they had completed, but these chat groups are not considered a primary form of engagement with the course.

One version of the course also integrates a bilingual (English and Urdu) and LLM-powered voice chatbot on WhatsApp, designed to serve as both an English language practice partner, personalized tutor, and personal self-development coach or mentor.

The study

Nearly 3,000 teachers participated in the present study, each randomly assigned to a control group or one of two treatment groups. A baseline survey was conducted in late January 2025,

the course ran between January and May 2025, and an endline survey was fielded in May 2025. The baseline and endline surveys were both conducted by phone.

In January 2025, Beaj contacted over 4,710 teachers[2] to assess interest and availability in participating in the course. Of the 3,490 individuals who completed the baseline survey, 2,931 participants gave verbal consent over the phone[3] to join the course. These 2,931 teachers who completed the baseline survey and signaled their intention to join and complete the course are considered as the sample for the present study. After providing verbal consent, these participants were randomly assigned to one of three groups (on a rolling basis, in six batches):

- Treatment group 1 (T1): Self-Development course without AI voice chatbot
- Treatment group 2 (T2): Self-Development course with AI voice chatbot
- Control group (C): No access to the course

This study seeks primarily to quantify the impact of Beaj's Self-Development course on English proficiency and psychological well-being among teachers in low-cost private schools in Pakistan, and to test the additional impact of an AI-powered voice chatbot. We hypothesize that teachers who participate in the course will demonstrate improved English proficiency and psychological well-being. We expect effects to be stronger within treatment group 2 (where an AI-powered voice chatbot complements the basic content) relative to treatment group 1 (basic content). Findings from this study will inform future programming and guide Beaj's future scale up and engagement strategy with key stakeholders.

Note**:** To date, incoming data has only been analyzed for data quality purposes (tracking outreach and survey completion, checking for expected distribution of variables, assessing balance across treatment groups, etc.). No members of the research team have done any analysis related to constructing the outcomes of interest described below or estimating treatment effects. No such analysis will be conducted until after this document is filed on the AEA registry.

Analysis

The planned analysis falls into three categories: (1) Baseline balance and attrition analysis, (2) treatment effects analysis, and (3) exploratory and descriptive analysis.

Baseline balance and attrition analysis

[2] Beaj received contact information for potential participants in two ways. First, partner schools that were interested in their teachers taking the course shared lists of teachers for Beaj to contact, and second, Beaj ran a digital ad campaign on social media platforms to solicit interest. The ad included a link to a registration form for the course trial and those interested filled out the form with their contact information.
[3] As part of the consent process, participants were asked three questions: (1) Are you interested in enrolling in Beaj's 3-month Self-Improvement course with a full scholarship? (2) Do you have any commitments in the next three months that might prevent you from completing the course? (3) If Yes to 2: Would you prefer to join the upcoming batch or the one after that, which will begin a few months later?

We will test for balance across the control and two treatment groups along several characteristics collected during the baseline survey: Age, highest level of education completed, marital status, whether they are a teacher, whether they are an administrator, employment at a public or private school, grade levels taught, subjects taught, household monthly income (collected on a 6-category scale), salary (collected on a 8-category scale), access to a private smartphone, and access to a shared smartphone.

Given the nature of a phone survey, substantial attrition is likely; based on our incoming data checks, around 30% of the sample did not complete the endline survey. To assess the extent and severity of attrition, we will first test for differential attrition (endline survey non-completion) across the control and two treatment groups. Second, we will test whether and which baseline characteristics (as listed above) predict endline survey attrition.

To address concerns about differential attrition across control and treatment groups or differential composition of non-attritors across control and treatment groups, we will estimate the main treatment effects analysis using inverse probability weights to restore comparability with the original baseline sample as a robustness check. To implement this, we will use logit regressions to estimate the probability of survey completion based on the baseline characteristics listed above, then use the inverse of fitted probabilities as weights in the treatment effects analysis.

Treatment effects analysis

The main treatment effects analysis will consist of two comparisons of interest. The goal of the first is to estimate the causal effects of the self-development course (with or without the additional AI chatbot) on English proficiency and psychological well-being. The goal of the second is to test whether the effect of the course is stronger with access to AI chatbot than without.

First, we will estimate intent-to-treat effects of the self-development course on English proficiency and psychological well-being by comparing outcomes across participants assigned to either treatment group (T1 and T2) versus those assigned to the control group (C). The main specification for this analysis will be:

$$Y_i = \beta_0 + \beta_1 T_i + \varepsilon_i \qquad (1)$$

In regression (1), T_i is defined as assignment to either treatment group (T1 or T2), where β_1 captures the causal effect of access to course, with or without the AI chatbot.

$$Y_i = \gamma_0 + \gamma_1 T1_i + \gamma_2 T2_i + \varepsilon_i \qquad (2)$$

In regression (2), $T1_i$ and $T2_i$ are defined as assignment to T1 or T2, respectively, where γ_1 and γ_2 give the treatment effect of each version of the course independently. For this analysis, we will test H_o: $\gamma_1 = \gamma_2$ for differential treatment effects with addition of AI chatbot.

Robustness
We will test robustness of the main treatment effects analysis to including enumerator fixed effects, fixed effects to indicate during which of six "batches" the participant was randomized into the study, and to estimation using inverse probability weighting to account for differential attrition as discussed above.

We will also run alternate specifications that exclude a small number of schools where the administration had asked their teachers not to participate in the course. This occured after the teachers had completed the baseline survey and had consented to participate in the study, and some of them had been randomized into one of the treatment groups. The school administrators in question asked their teachers not to participate. Beaj's sense is that this happened because the school administrators felt that participating in the course might make it more likely that their teachers would leave their jobs. This will not be our main specification, but we will run it to provide an estimate of the course's effects in cases where participants did not face active resistance from their workplace.

Heterogeneity
We will also test for heterogeneity in the main treatment effects of interest along the following dimensions as collected during the baseline survey:

- Age: Up to age 30 vs. above age 30
- Marital status: currently married vs. not
- Educational attainment: Master's degree or higher vs. bachelor's degree or lower
- Grade level taught: Primary (up to grade 5) vs. secondary or higher (grade 6 or higher)
- Subject taught: English vs. other subject; Everything aside from Urdu, Islamiat, Quran vs. Urdu, Islamiat, Quran

This heterogeneity analysis will be designed to test various predictions about which types of teachers may benefit more from the course. For example, younger teachers may find the video-based course more familiar to engage with, and so may be more engaged with the course and benefit more from the course. Teachers who teach more advanced students and/or subjects that involve more advanced usage of English may similarly value the course more, and hence engage with the course to a greater degree and benefit more from the course.

Other predictions are less clear. For example, more educated teachers may be better able to make use of the course, and hence benefit more. On the other hand, less educated teachers may have had less prior exposure to some of the concepts covered or may have had fewer professional development opportunities, and hence value the course more and/or have more to gain from the course content.

Exploratory and descriptive analysis

We will also conduct descriptive analysis to better understand which types of teachers engage more with the course (ie., how baseline characteristics predict course engagement and completion) and how engagement and completion of the course correlate with English proficiency and psychological well-being as measured at endline. All analysis will be conducted within the participants assigned to either treatment group (T1 or T2). This analysis will be exploratory, and non-causal.

First, we will explore how baseline characteristics predict engagement with the course using:

$$E_i = \alpha + \delta X_i + \varepsilon_i \qquad (3)$$

In regression (3), E_i are measures of engagement as described below. The vector X_i captures various baseline characteristics. We will estimate this first with a more limited set of characteristics (age and educational attainment), then with the full set of baseline characteristics listed above.

Second, we will explore how measures of engagement correlate with measures of English proficiency and psychological well-being at endline using:

$$Y_i = \lambda + \theta E_i + \varepsilon_i \qquad (4)$$

In regression (4), E_i is a vector of engagement-related measures as described below, each included independently and together. We will estimate this regression with the full sample of participants who started the course and with a more limited sample of participants who completed at least the first week (and so demonstrated some minimal amount of commitment to the course).

We will also run a secondary specification to test for a differential relationship between engagement and outcomes across our two treatment arms:

$$Y_i = \lambda + \theta E_i + \pi T2_i + \gamma T2_i * E_i + \varepsilon_i \qquad (5)$$

In regression (5), E_i is a vector of engagement-related measures as described below, each included independently and together. We include a control for assignment to T2, the version of the course that includes access to an AI chatbot and interactions between assignment to T2 and engagement-related measures, to allow for (and test for) any differences in the relationship between engagement with the course and outcomes of interest with and without the AI chatbot. Similar to regression (4), we will estimate this regression with the full sample of participants who started the course and with a more limited sample of participants who completed at least the first week (and so demonstrated some minimal amount of commitment to the course).

Heterogeneity
We will also test for heterogeneity in engagement in the course using regression (3) along the same dimensions as discussed in the section on the main treatment effects analysis.

Outcomes

This section describes the main outcomes as collected during the endline survey that will be used in the analysis. We distinguish between two groups of outcomes – primary outcomes of interest and secondary outcomes. The five primary outcomes include:

1. Composite English index
2. General self-efficacy index
3. Teaching efficacy index
4. Agency and empowerment
5. Goal-setting index

In addition to standard p-values on the coefficients of interest (β_1 in regression (1), γ_1 and γ_2 in regression (2)), we will also report False Discovery Rate (FDR) adjusted q-values across these five primary outcomes to account for multiple hypothesis testing (Anderson, 2008).

All other outcomes described below are considered secondary outcomes. .

Unless noted otherwise, standardized indices will be constructed by summing across component items, then subtracting the mean and dividing by the standard deviation within the control group to arrive at an index measured in standard deviation units.

English language

The English language assessment consisted of a three-question listening comprehension task and a three-question prompted dialogue task. We will construct two separate English language indices from each of these tasks, and construct a composite English language index from these.

1. **Listening comprehension index**: Standardized sum of number of correct answers to three questions.
2. **Prompted dialogue index**: Standardized sum of scores on three questions, each scored by a team of research assistants using a consistent rubric.
3. **Composite English language index**: Standardized sum of the listening comprehension index and prompted dialogue index.

In secondary analysis (likely for an appendix), we will also look at each of the 6 questions included in the listening comprehension task and the prompted dialogue task separately.

We will also assess robustness of the composite English language index to two alternative methods of constructing: (a) first giving equal weighting to the listening comprehension score and each of the three prompted dialogue scores, and (b) weight component items using inverse covariance weighting.

Psychological well-being

We captured various dimensions of psychological well-being using several different scales. We will construct a series of scores or indices to capture each of these dimensions of psychological well-being as described below.

1. **General self-efficacy index** (Schwarzer & Jerusalem, 1995)
   This index will be the standardized sum of six items, each measured on a scale from 1 (not true at all) to 4 (exactly true). We will also report as a secondary outcome (likely for an appendix) the non-standardized sum of these six items. Participants were asked how true each of the following statements are to them:

   a. *If someone opposes me, I can find means and ways to get what I want.*
   b. *It is easy for me to stick to my aims and accomplish my goals.*
   c. *I am confident that I could deal efficiently with unexpected events.*
   d. *Thanks to my resourcefulness, I know how to handle unforeseen situations.*
   e. *I can remain calm when facing difficulties because I can rely on my coping abilities.*
   f. *No matter what comes my way, I'm usually able to handle it.*

2. **Teaching efficacy index** (OECD, 2019 (page 285); Schweig et al., 2025 (page 16))
   This index will be the standardized sum of the 12 items listed below, each measured on a scale from 1 (not at all) to 4 (a lot). We will also report as secondary outcomes (likely in an appendix) the following three indices: (1) Self-efficacy in classroom management, using items d,f,h,i; (2) Self-efficacy in instruction subscale, using items c,j,k,l; (3) Self-efficacy in student engagement subscale, using items a,b,e,g. Participants were asked to what extent they can do each of the following in their teaching:

   a. *Get students to believe they can do well in school work*
   b. *Help students value learning*
   c. *Craft good questions for students*
   d. *Control disruptive behaviour in the classroom*
   e. *Motivate students who show low interest in school work*
   f. *Make my expectations about student behaviour clear*
   g. *Help students think critically*
   h. *Get students to follow classroom rules*
   i. *Calm a student who is disruptive or noisy*
   j. *Use a variety of assessment strategies*
   k. *Provide an alternative explanation, for example when students are confused*

l. *Vary instructional strategies in my classroom*
m. *Support student learning through the use of digital technology (e.g. computers, tablets, smart boards)*

3. **Agency and empowerment**
   To capture agency and empowerment, participants were asked to select which of three teachers described in short vignettes they felt was most similar or least similar to them. Following the methodology employed in Cheema et al. (2023), we will use a multinomial logit model regressing teachers' choice of identifying most with a vignette on treatment conditions. We will treat the lowest agency teacher (Asma) as the base category and estimate whether treatment affects identification with the medium-agency teacher (Salma) or the high-agency teacher (Zakia). As a secondary outcome, we will run an analogous model using teachers' choice of identifying least with a vignette.

4. **Goal-setting index (MAGNET, 2023)**
   This standardized index will be a modified index version of the Goal-Setting Capacity Scale, using a sum of the four items (out of the usual eight items) included in the endline survey, each measured on a scale from 1 (strongly disagree) to 4 (strongly agree). Participants were asked the extent to which they agree or disagree with the following statements:

   a. *I set specific, clear goals for myself.*
   b. *I make plans to help me achieve my goals.*
   c. *I feel proud when I achieve my goals.*
   d. *I am able to prioritize multiple goals*

5. **Depression index** (Radloff et al., 1977; Andresen et al., 1994)
   This standardized index will be based on summing over all 10 items, each measured on a scale from 1 (Rarely or none of the time (less than 1 day)) to 4 (4 = Most of the time (5-7 days)). Starred items below will be reverse-coded for consistency. As a secondary outcome, we will also report the CESD score, which is the sum over all 10 items, recoded so that the final score ranges from 0 to 30. As another secondary outcome, we will also construct a depression indicator, indicating a CESD score consistent with depression (ie., CESD score >= 10). Participants were asked how frequently in the past week they felt each of the following statements applied:

   a. *I was bothered by things that usually don't bother me*
   b. *I had trouble keeping my mind on what I was doing*
   c. *I felt depressed*
   d. *I felt that everything I did was an effort*
   e. **I felt hopeful about the future*
   f. *I felt fearful*
   g. *My sleep was restless*
   h. **I was happy*

i. *I was lonely*
j. *I could not "get going"*

6. **Labor market aspirations indices** We will report the following standardized indices: (1) a job satisfaction index (based on summing items i-iv) and (2) a career development index (based on summing items v-viii), where items are measured on a scale from 1 (strongly disagree) to 4 (strongly agree). Participants were asked the degree to which they agree or disagree with the following:

    a. *I am satisfied with the salary of my current employment*
    b. *I am satisfied with the workload of my current employment*
    c. *I am satisfied with the recognition I receive at work*
    d. *I am satisfied with the opportunities I have to grow and improve in my work*
    e. *I have the tools and resources to develop my skills as a teacher.*
    f. *I anticipate improving as a teacher in the coming years.*
    g. *I expect to have advanced in my career in five years*
    h. *I expect to have more responsibility at work in five years*

7. **Satisfaction with classroom autonomy index** (OECD, 2019 (page 285))
    This index will be the standardized sum of the 5 items listed below, each measured on a scale from 1 (strongly disagree) to 4 (strongly agree). Participants were asked the extent to which they agree or disagree with with having control over:

    a. *Determining course content*
    b. *Selecting teaching methods*
    c. *Assessing students' learning*
    d. *Disciplining students*
    e. *Determining the amount of homework to be assigned*

8. **Locus of control score** (Haerpfer et al., 2022)
    This measure will be scored continuously, on a scale from 1 (no choice at all) to 10 (a great deal of choice). Respondents were asked to indicate the degree to which they felt they have choice and control over their lives in response to the following:
        a. *Some people feel they have completely free choice and control over their lives, while other people feel that what they do has no real effect on what happens to them. Please use this scale where 1 means "no choice at all" and 10 means "a great deal of choice" to indicate how much freedom of choice and control you feel you have over the way your life turns out.*

Other: Engagement
We will use three primary measures of engagement for the descriptive and exploratory analyses described above.

- Indicator for starting the course and finishing the first lesson
- Fraction of the course completed (T1, T2): Proportion of lessons completed

- Indicator for completing all 12 weeks of the course

Other: Household economic circumstances

As secondary outcomes of interest, we will also explore impacts on economic circumstances.

9. Indicator for increase in earnings since the start of 2025
10. Indicator for increase in household economic situation since the start of 2025

Bibliography

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American Statistical Association*, 103(484):1481–1495.

Andresen, E. M., Malmgren, J. A., Carter, W. B., and Patrick, D. L. (1994). Screening for depression in well older adults: Evaluation of a short form of the CES-D. *American Journal of Preventive Medicine*, 10(2):77–84.

Cheema A, Khan S, Liaqat A, Mohmand SK. Canvassing the Gatekeepers: A Field Experiment to Increase Women Voters' Turnout in Pakistan. *American Political Science Review*. 2023;117(1):1-21.

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin & B. Puranen (eds.). 2022. World Values Survey: Round Seven – Country-Pooled Datafile Version 6.0. Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat. doi:10.14281/18241.24

MAGNET (2023). Goal-setting Capacity Scale. https://magnet.ifpri.info/goal-setting-capacity-scale/

OECD (2019), *TALIS Starting Strong 2018 Technical Report,* TALIS, OECD Publishing, Paris, https://doi.org/10.1787/0921466e-en.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1(3):385–401.

Schwarzer, R., & Jerusalem, M. (1995). Generalized self-efficacy scale. *J. Weinman, S. Wright, & M. Johnston, Measures in health psychology: A user's portfolio. Causal and control beliefs,* 35(37), 82-003.

Schweig, J., Wang, E. L., Lee, S., & Mihaly, K. (2025). *Teach For Pakistan Evaluation.* RAND.