

**Pre-Analysis Plan for
When Helpers Need Help:
Encouraging Mental Health Care Use
Among Health Workers**

MICHELLE ESCOBAR CARÍAS
Newcastle University

CARLOS SCHMIDT-PADILLA
University of California, Berkeley

September 3, 2026

Contents

1	Study Overview	1
2	Design	1
2.1	Population and eligibility	1
2.2	Randomization	1
2.3	Arms	2
2.4	Timeline and data collection	3
3	Outcomes	3
3.1	Primary endpoints	3
3.2	Secondary outcomes	4
3.3	Construction rules	4
4	Estimation	4
4.1	Primary specification (intention to treat)	4
4.2	Targeting	5
4.3	Inference	5
4.4	Attrition, delivery, and balance	5
4.5	Dose-response (exploratory)	5
5	Hypotheses	5
6	Multiple Testing	6
7	Heterogeneity	6
8	Power	7
9	External Validity	7
10	Ethics	7
11	Deviations and Amendments	8

1 Study Overview

Depression and anxiety are widespread and largely untreated, and low demand is part of the reason. Take-up stays low even where care is free. The conditions themselves may suppress demand, since hopelessness and low motivation raise the perceived cost of seeking help, so the workers in greatest need can face the strongest barriers (Breza et al. 2026). We ask how to raise demand for mental-health care among public hospital workers in El Salvador, and whether the workers who come forward are the ones who need it most. We offer 1,057 workers up to four short, free, and confidential counseling sessions over WhatsApp, and we randomize the invitation. Treatment 1 adds a wellbeing profile of peers of the same sex and age. Treatment 2 adds a first call from a counselor, so the worker no longer has to make the first move. Administrative records give take-up and the number of sessions completed, and on both we estimate an average effect and a targeting effect, the difference between the effect on workers who screened positive at baseline and the effect on those who did not. An endline survey six months later measures wellbeing, emotion regulation, and self-reported professional functioning.

The service is part of SanaMente, a mental-health program that Glasswing International, a Salvadoran non-governmental organization, runs for public hospital staff. Between 2022 and 2025 the program gave a short group training, the *Formación General*, to roughly 10,900 workers. That training was not randomized and is not evaluated here. We invited all of its graduates to a baseline survey in early 2026, and response required active contact details and, in most cases, still working in a public hospital. The 1,521 who responded and consented form the baseline sample, spread across more than 30 hospitals in all 14 departments. The survey showed that the psychological burden among them is high. Thirty-nine percent reported at least mild anxiety ($GAD-7 \geq 5$), 29 percent low wellbeing ($WHO-5 < 50$), and 17 percent high burnout, and half crossed at least one of these thresholds. Only 2 percent had raised a mental-health concern at their last medical visit, yet 91 percent said they would take part in a follow-up. Among our respondents, stated demand is high and revealed demand is near zero. Of those who consented, the 1,057 still working in a public hospital and with no prior follow-up from the program are the sample we randomized.

Status and timestamp. This plan was finalized after random assignment and after outreach had begun, but before we received or examined any administrative take-up record or any endline data.

2 Design

2.1 Population and eligibility

We invited all roughly 10,900 eligible graduates of the *Formación General* to the baseline survey, fielded March to May 2026. It returned 1,561 records, of which 1,521 carry informed consent. Eligibility uses baseline covariates only, and we applied it before assignment, in this order: (i) consented at baseline to be invited to the service (1,381); (ii) currently employed in a public hospital (-162); (iii) no self-reported prior Glasswing follow-up (-161); (iv) no match, by national ID or normalized name, to the implementer’s internal rosters of prior follow-up, complementary training, or interventionist training (-1); (v) no duplicate within the pool (-0). The randomized sample is 1,057 workers.

2.2 Randomization

Assignment is individual with equal 1:1:1 allocation, blocked on sex crossed with a decade age band (≤ 29 , 30–39, 40–49, 50+), giving eight blocks; within a block each arm receives $\lfloor n/3 \rfloor$ participants and the remainder is handed to distinct arms at random, so within-block arm sizes differ by at most one. Blocking on sex and age also guarantees balance within the cells whose average wellbeing is reported back

Table 1: Trial snapshot.

Design	Individually randomized controlled trial, three arms, equal 1:1:1 allocation
Randomized sample	$N = 1,057$ (351 control, 354 Treatment 1, 352 Treatment 2), the full eligible pool
Population	Public hospital workers who completed the Formación General (2022–2025), answered the baseline survey, consented to be invited to the service, still work in a public hospital, and had no prior Glasswing follow-up
Service (all arms)	Up to four individual wellbeing sessions with a trained lay counselor (<i>intervencionista</i>), taken either as a WhatsApp call of 10 to 15 minutes or as an exchange of WhatsApp messages, at the worker’s choice; free, confidential, voluntary at every step
Control	WhatsApp message describing the service and giving the counselor’s number; the participant initiates contact
Treatment 1	Control message plus a one-screen wellbeing profile of peers of the same sex and age group; the participant initiates contact
Treatment 2	Control message plus a first proactive call from a counselor, who initiates contact
Randomization	Individual, blocked on sex $\times$ age band (≤ 29 , 30–39, 40–49, 50+; eight blocks); seeded R script archived with its log and balance record
Data	Program administrative and monitoring records (all participants, continuous); endline survey about six months after outreach, repeating the baseline instruments
Primary endpoints	P1 take-up (completed at least one session); P2 intensity (sessions completed, 0 to 4); both administrative. On each we estimate an average effect and a targeting effect by baseline screening status
Registration	AEA RCT Registry, AEARCTR-0019171 , registered August 11, 2026; this plan uploaded after assignment and after outreach began, before any outcome data were examined
Ethics	Universidad Francisco Gavidia ethics committee, approval no. 002-2026

to Treatment 1. We generated the assignment with a self-contained R script and a fixed seed, archived with its output, run log, and covariate-balance record; re-running it reproduces the assignment exactly. A joint orthogonality test (multinomial logit of arm on all baseline covariates: demographics, occupation, physical health, care-seeking, and every baseline mental-health scale) gives $p = 0.77$ without and $p = 0.58$ with block fixed effects ([Bruhn and McKenzie 2009](#)). Participants are not told which arm they are in; the arm is revealed only implicitly through the outreach they receive.

2.3 Arms

All arms are offered the identical service: up to four weekly individual sessions with a trained lay counselor, taken either as a WhatsApp call of 10 to 15 minutes or as an exchange of WhatsApp messages, at the worker’s choice, covering psychoeducation on stress, anxiety, trauma, and grief with practical emotion-regulation strategies. The service is free and confidential, the employer is never informed, and participation is voluntary at every step. The arms differ only in the outreach.

- *Control (general information)*. A WhatsApp message describing the sessions and giving the counselor’s number. The participant must reply or call to begin; non-responders receive a reminder after three days and a final reminder after seven.
- *Treatment 1 (peer wellbeing information)*. The control message plus a one-screen profile card for the participant’s sex and age group (women 20–40, women 41+, men 20–40, men 41+), built from the baseline survey: a headline share of the group with at least one condition (61, 42, 51, and 38 percent

respectively) and three components, anxiety symptoms ($\text{GAD-7} \geq 5$), low wellbeing ($\text{WHO-5} < 50$), and burnout or poor sleep affecting daily life (burnout mean ≥ 3 or sleep impairment ≥ 8), followed by a normalizing statement and the same call to action. The card is referential to the group, not an individual score or diagnosis. The participant still initiates contact.

- *Treatment 2 (proactive contact)*. The control information plus a first proactive call from a counselor, who offers to begin the service. Initiation shifts from the participant to the counselor.

The estimand for Treatment 1 is the effect of adding social information about peers to a self-initiated invitation; for Treatment 2 it is the effect of counselor-initiated contact. Their difference contrasts the two candidate barriers the baseline points to, the friction of initiating contact and what a worker knows about their own need, holding the service fixed.

2.4 Timeline and data collection

The baseline survey ran from March to May 2026. Randomization followed in July, and outreach began shortly after, delivered in monthly waves. Administrative records accrue continuously. The endline survey comes about six months after outreach and repeats the baseline instruments.

3 Outcomes

3.1 Primary endpoints

We construct both primary endpoints from the implementer’s outreach and monitoring records, matched to the randomized sample by program code. They are defined for all 1,057 randomized participants regardless of response, so there is no attrition on the primaries and estimates are intention to treat. Family-wise error rates are controlled within each confirmatory family ([Section 6](#)).

Table 2: Primary endpoints

	Endpoint	Definition (administrative records)
P1	Take-up	Indicator for completing at least one session with a counselor, by call or by message
P2	Intensity	Number of sessions completed, 0 to 4

Sessions come in two forms. The worker can take a session as a WhatsApp call, or as an exchange of messages with the counselor. The two are not symmetric. A worker who accepts the message format starts receiving content and keeps receiving it until they ask to stop, so accepting and receiving are close to the same event. A worker who accepts the call format must still take or place each call. Take-up therefore means something slightly different in the two groups. We will report the primary endpoints separately by form and for the two together. We will fix the rule for combining them before we examine outcomes by arm, state that rule in the paper, and show how the results move under the alternatives we considered. Format is chosen by the worker and not randomized, so any comparison between the two carries no causal claim.

We pre-specify three earlier steps as secondary outcomes: whether the outreach reached the worker, whether the worker responded to it, and whether a first session followed. These steps matter because a response and a session are not the same thing, at least for the workers who choose calls. We will use the administrative logs the counselors produce to measure engagement at each step.

3.2 Secondary outcomes

The endline survey, about six months after outreach, repeats the baseline instruments. Each scale is scored by its standard rules, as a continuous score and as a binary indicator at the conventional cutoff, and outcomes are grouped into four families:

- *A. Mental health and wellbeing.* WHO-5 wellbeing (< 50 low); GAD-7 anxiety (≥ 5 at least mild; ≥ 10 clinical); PHQ-2 depression (≥ 3); PROMIS sleep-related daytime impairment (≥ 8 on the 0–16 sum); Copenhagen Burnout Inventory, short form (mean ≥ 3 high). Family endpoint: an index over the five scales.
- *B. Emotion regulation and coping.* DERS-SF total; Mini-COPE adaptive (active coping, emotional and instrumental support) and maladaptive (denial, behavioral disengagement) subscales. Family endpoint: an index over the three, with the DERS-SF and the maladaptive subscale reversed first.
- *C. Professional functioning.* SEPCQ-27 patient-centered self-efficacy; closeness to colleagues (Inclusion of Other in the Self). Family endpoint: an index over the two.
- *D. Help-seeking.* Disposition to seek emotional support; reported use of mental-health services outside the program. Family endpoint: an index over the two. Mental-health stigma (four items, sum 4–20; elevated = above the baseline median of 10) belongs to this family but is reported on its own rather than folded into the index, since it is a barrier to help-seeking rather than a form of it.

3.3 Construction rules

Each family has a summary index. Components are first signed so that a higher value is better, then standardized against the control-group endline distribution. The index is the inverse-covariance-weighted average of its components (Anderson 2008), which weights each component by the information it adds rather than by an equal share. We report the first principal component and the equal-weighted mean (Kling, Liebman and Katz 2007) as alternative constructions. Components enter an index only when they measure the same construct; where they do not, we report them separately. Scale scores follow each instrument’s partial-response rules; an index is computed when at least half of its components are observed; missing values are never coded as zero; primary endpoints are never imputed. All cutoffs above are the ones used in the baseline report and in the Treatment-1 profile cards, fixed before any outcome data existed.

4 Estimation

4.1 Primary specification (intention to treat)

For participant i in randomization block b , with treatment indicators $T1_i$ and $T2_i$ and block fixed effects δ_b ,

$$Y_i = \alpha + \tau_1 T1_i + \tau_2 T2_i + \beta Y_{i,0} + \mathbf{X}_i' \boldsymbol{\gamma} + \delta_b + \varepsilon_i, \quad (1)$$

where τ_1 and τ_2 are the intention-to-treat effects of each treatment relative to control and the contrast $\tau_2 - \tau_1$ is tested from the same regression. We report every outcome under two specifications: unadjusted, with block fixed effects only ($\beta = 0$, $\boldsymbol{\gamma} = 0$), and adjusted, adding the baseline value of the outcome where one exists together with a set of baseline covariates $\mathbf{X}_i$. We will fix $\mathbf{X}_i$ before we examine outcomes by arm and state it in the paper. The primary endpoints have no baseline analog, so their adjusted specification carries covariates only. We estimate (1) by OLS throughout, as a linear probability model for the binary P1 and as a linear model for the count P2. For endline outcomes with a baseline value, (1) is ANCOVA, chosen over difference-in-differences for power (McKenzie 2012).

4.2 Targeting

Raising take-up is worth little if the workers drawn in are not the ones who need care, so we pre-specify a targeting analysis on the same two endpoints. Let S_i indicate that worker i screened positive at baseline on at least one of WHO-5 below 50, PHQ-2 at or above 3, GAD-7 at or above 5, burnout at or above 3, or sleep impairment at or above 8. Half of the baseline sample meets this definition. We estimate

$$Y_i = \alpha_0 + \alpha_1 S_i + \sum_{k=1}^2 \left(\beta_k Tk_i + \theta_k Tk_i \times S_i \right) + \mathbf{X}_i' \boldsymbol{\gamma} + \delta_b + \varepsilon_i, \quad (2)$$

where Y_i is P1 or P2.

The regression produces three numbers for each arm k : the effect among workers who screened negative (β_k), the effect among those who screened positive ($\beta_k + \theta_k$), and the gap between them (θ_k). We report all three. The gap is the targeting effect. It is positive when the arm raises take-up more among workers who screened positive than among those who did not, which is the case where the outreach reaches the workers who need the service. We also test whether the two arms differ in this respect (θ_1 against θ_2). The remaining coefficient, α_1 , compares screened-positive and screened-negative workers within the control group. It says whether workers with greater need come forward on their own when nothing is added to the invitation. As robustness we repeat the regression with S_i replaced by the continuous baseline mental-health index, and by each mental-health measure on its own, entered both as its threshold indicator and as its continuous score.

4.3 Inference

Standard errors are heteroskedasticity-robust at the individual level, the level of randomization; the service is delivered one to one, so there is no group structure to cluster on. Counselor identity is recorded for every contact and session, and counselor-clustered standard errors are a pre-specified robustness check. We also report randomization inference for the primary tests.

4.4 Attrition, delivery, and balance

The primary outcomes are administrative and as such, they do not suffer from attrition. We expect endline survey attrition of 20 to 30 percent and report it by arm with baseline predictors. Where it differs across arms, we bound secondary effects with Lee trimming bounds (Lee 2009) and report inverse-probability-weighted estimates as robustness. Nothing is imputed. We report outreach delivery by arm, and a baseline balance table with joint orthogonality tests.

4.5 Dose-response (exploratory)

For the secondary mental-health outcomes we pre-specify an exploratory dose-response: two-stage least squares of the endline outcome on sessions completed, instrumented by the assignment indicators. A session is not the same object in the two delivery formats, so we read the dose-response for each format as well as for both together. It should be read with caution in any case, since Treatment 1's information may affect wellbeing directly. The intention-to-treat estimate is the confirmatory one.

5 Hypotheses

- **H1 (proactive contact raises take-up).** Treatment 2 raises P1 and P2 relative to control ($\tau_2 > 0$). The baseline gap between stated interest (91 percent) and revealed demand (2 percent) indicate that

initiation costs, not willingness, are the binding constraint.

- **H2 (peer information raises take-up).** Treatment 1 raises P1 and P2 relative to control ($\tau_1 > 0$). It can work through two channels. The first is normalization: seeing that distress is common among workers of the same sex and age lowers the perceived stigma of seeking support. The second is information about oneself: a worker who has not thought of their own state as something to act on may recognize it in the profile of people like them, and revise what they believe about their own need.
- **H3 (friction dominates information).** The proactive-contact effect exceeds the information effect ($\tau_2 > \tau_1$). Stated willingness was already high at baseline, so the cost of taking the first step is the more likely binding constraint.
- **H4 (targeting).** Treatment 1 improves targeting ($\theta_1 > 0$). A peer profile that names the group’s symptoms gives a worker a benchmark against which to read their own state, and it should move those who recognize themselves in it. We do not sign θ_2 . A proactive call lowers the cost of starting for everyone, so it may draw in workers with little need as readily as those with much, and the sign is the object of the test.
- **H5 (mental health, secondary).** Both treatments improve the family-A index at six months ($\tau_1, \tau_2 \geq 0$). We expect intention-to-treat effects to be small because partial take-up dilutes them, and we expect anxiety and wellbeing to respond before burnout.
- **H6 (emotion regulation and coping, secondary).** Both treatments improve the family-B index at six months ($\tau_1, \tau_2 \geq 0$). The sessions teach emotion-regulation strategies directly, so this family sits closest to the content of the service. If the service works at all, an effect should show here first.
- **H7 (professional functioning, secondary).** Both treatments improve the family-C index at six months ($\tau_1, \tau_2 \geq 0$). This family is the most distal, since it requires a change in the worker’s own state to carry through into how they work, so we expect the smallest effects of the four. Both measures are self-reported, so an effect on them is evidence about how workers see their own practice, not necessarily about the care patients receive.
- **H8 (help-seeking, secondary).** Both treatments raise the disposition to seek emotional support and lower stigma. We expect the stigma effect to be larger under Treatment 1, whose content targets it directly.

Registered pattern (exploratory): any effect on mental health should be concentrated among participants who completed sessions; the dose-response exhibit states whether it is.

6 Multiple Testing

We deal with multiple hypothesis testing through the indices in the first instance. Where we report sub-indices or groups of outcomes, we report p -values adjusted within that group by the Romano-Wolf step-down procedure ([Romano and Wolf 2005](#)), alongside the unadjusted values.

7 Heterogeneity

Pre-specified moderators, each tested as an interaction with the two treatment indicators in (1) on the primary endpoints and corrected within the set: (i) sex; (ii) age band (under 30, 30 to 49, 50 and over); (iii) occupation (physician, nurse, other); (iv) baseline stigma above the sample median; (v) time since the Formación General (24 months or less versus more). Baseline mental-health need is absent from this list because it defines the targeting analysis of [Section 4.2](#), which is confirmatory. The baseline motivates (i), (ii), and (iii), where men, physicians, and the youngest workers stood out. Stigma (iv) is the channel through which we expect the information arm to work. We report heterogeneity as hypothesis-generating

unless it survives correction.

8 Power

Minimum detectable effects follow from one standard formula: $MDE = 2.8 \times SE$ of the arm difference, with $2.8 = z_{0.975} + z_{0.80}$ (80 percent power, 5 percent two-sided tests), for a pairwise contrast between two arms of about 352 participants. For a binary endpoint $SE = \sqrt{2p(1-p)/n}$, where the control rate p is unknown in advance and is bracketed in Table 3; for an endline outcome under ANCOVA, $SE = \sqrt{1-\rho^2} \sqrt{2/n}$ with an assumed baseline–endline correlation $\rho = 0.5$ and 20 percent attrition ($n \approx 282$ per arm). The trial is powered for moderate take-up effects, roughly five to ten percentage points, and for moderate effects on endline outcomes, not for small ones; subgroup interactions are detectable only if large. The targeting effect is an interaction, and the roughly even split of the sample at the screening threshold is the best case for it. For scale, Breza et al. (2026) report a targeting effect of 13 percentage points on appointment scheduling for their information arm and 23 percentage points for their combined arm, so effects of that size sit at or above our detection threshold.

Table 3: Minimum detectable effects for a pairwise arm contrast.

Endpoint	Assumption	MDE
P1 take-up (binary; $n \approx 352$ per arm)	control rate 5 percent	4.6 pp
	control rate 10 percent	6.3 pp
	control rate 20 percent	8.4 pp
	control rate 30 percent	9.7 pp
Endline outcome (ANCOVA, $\rho = 0.5$; $n \approx 282$ per arm)	continuous	0.20 SD
Targeting effect θ_k on P1 (interaction)	control rate 10 percent	≈ 12.6 pp
Binary-moderator interaction on an endline outcome	$\rho = 0.5$	≈ 0.41 SD

9 External Validity

The randomized sample is selected on reachability and on stated willingness to be contacted, and the direction of that selection is unknown. Workers already interested in support are more likely to have answered, which would put measured take-up above what an untargeted invitation to the full graduate pool would produce. Withdrawal and disengagement are also symptoms of the conditions we measure, so the least well may be the least likely to have answered us at all, which would understate both the burden and the unmet need in the pool. We do not attempt to sign this selection or to correct for it. Randomization is within the responding group, so the treatment contrasts are internally valid, and we report take-up levels as descriptive of this group rather than of the graduate population.

10 Ethics

The protocol was approved by the ethics committee of Universidad Francisco Gavidia, El Salvador (approval no. 002-2026). Informed consent was obtained at the baseline survey, where respondents also consented to be invited to the service; every outreach message states that participation is voluntary, free, and confidential, that the employer is never informed, and that the participant may decline or stop at any point. Counselors follow the implementer’s protocols for distress and referral. Glasswing International holds the identifiable data. We work with de-identified analysis files linked by program code.

11 Deviations and Amendments

We upload this plan after assignment and after outreach began, and before we received or examined any outcome data. Amendments before any administrative or endline outcome data are seen are permitted as dated registry addenda with justification. Once we have seen outcome data we make no further changes. We label any departure in the paper a deviation and report it alongside the registered specification. All outputs distinguish confirmatory (registered) from exploratory results.

References

- Anderson, Michael 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.
- Breza, Emily, Kevin Carney, Vijaya Raghavan, Kailash Rajah, Thara Rangaswamy, Gautam Rao, Frank Schilbach, Sobia Shadbar, and James Stratton.** 2026. “Financial Incentives, Health Screening, and Selection into Mental Health Care: Experimental Evidence from College Students in India.” Working paper. AEA RCT Registry AEARCTR-0015379.
- Bruhn, Miriam, and David McKenzie.** 2009. “In Pursuit of Balance: Randomization in Practice in Development Field Experiments.” *American Economic Journal: Applied Economics*, 1(4): 200–232.
- Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz.** 2007. “Experimental Analysis of Neighborhood Effects.” *Econometrica*, 75(1): 83–119.
- Lee, David S.** 2009. “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects.” *Review of Economic Studies*, 76(3): 1071–1102.
- McKenzie, David.** 2012. “Beyond Baseline and Follow-up: The Case for More T in Experiments.” *Journal of Development Economics*, 99(2): 210–221.
- Romano, Joseph P., and Michael Wolf.** 2005. “Stepwise Multiple Testing as Formalized Data Snooping.” *Econometrica*, 73(4): 1237–1282.