Pre-Analysis Plan for

The Economic Consequences of Improving Sleep Among the Urban Poor Through CBT-I

Michelle Escobar Carías
Newcastle University

Idaliya Grigoryeva
University of California San Diego

Juan Pablo Franco
University of Melbourne

Sean P. A. Drummond
Monash University

September 3, 2026

Contents

1 Study Overview

This trial estimates the effect of CBT-I (cognitive behavioral therapy for insomnia), a treatment specific to insomnia, on the wellbeing and labor-market outcomes of low-income urban workers in Kibera, Nairobi, in partnership with the Busara Center for Behavioral Economics. The nearest experiment, in Chennai, did not target insomnia: it documented that low-income urban workers sleep less than the recommended amount, increased their sleep, and found no downstream effects, and its own results suggest that what matters may be sleep quality and not quantity (Bessone et al. 2021; Rao et al. 2021). CBT-I is effective at improving sleep efficiency without necessarily increasing sleep duration (Morin et al. 2006; Trauer et al. 2015; van Straten et al. 2018). Across twenty randomized trials, sleep efficiency improves by almost 10 percentage points (95% CI 8.09 to 11.73), while the estimate for total sleep time, 7.6 minutes, is positive, but not statistically distinguishable from zero at the 5% significance level (95% CI -0.51 to 15.74). (Trauer et al. 2015). Among people with insomnia it is common to stay in bed for a long time trying to fall asleep, and that is itself the source of anxiety, the point at which people start worrying and feeling hopeless; part of the intervention therefore reduces time in bed. The intervention increases sleep efficiency and develops a healthier relationship with sleep and sleep-related behaviors; it does not increase the amount of sleep. There are almost no studies of CBT-I in developing countries and almost no objective sleep data from Africa, so the trial also collects new objective sleep data. The first stage we expect is a reduction in insomnia symptoms and a rise in sleep efficiency, which is expected to have downstream effects on cognition, on how people interact with others, and on productivity.

The design is one treatment and one (active) control, whose compensation and time committed to the study are comparable. The study lasts six weeks from baseline to endline: (1) the first week is baseline data collection, including the baseline survey and the device setup (no intervention content); (2) the intervention consisting of the four weekly sessions, with a message exchange between sessions to check in and remind participants of the last week's session content and the next meeting time; and then (3) one more week of endline data collection. Control participants come to the lab on the same weekly schedule to complete surveys, with no insomnia or CBTi content. Participants are adults aged 18 to 45 who report insomnia symptoms and are working or looking for work, and who are not primary caretakers of children under 3 years old. Additional inclusion and exclusion criteria are detailed in Section 2.1. We plan the 6- and 12-month follow-ups to estimate long-run effects, conditional on additional funding. This plan was finalized after screening closed and randomization was complete, before any baseline or outcome data had been examined. The screening data were used for eligibility, for randomization, and for descriptive context only.

2 Design

Table 2 in the appendix summarizes the design in one table.

2.1 Sample

The trial targets people with insomnia, for whom CBT-I has been shown to be effective in other contexts. There are no data on the prevalence of insomnia in this population, and we estimate that prevalence from the screening and recruitment survey, which covers a large sample of residents of the low-income Kibera settlement. Participants are adults aged 18 to 45 who are employed or searching for employment, are not the primary caregiver of a child under 3, do not do regular night work (between midnight and 5 am), own a personal phone with an active WhatsApp number, and said at screening that they were willing to take part, with no more than one adult per household eligible. The sample of eligible participants is then further restricted to those with the Insomnia Severity Index (ISI) score of 10 or above, the conventional cutoff for detecting insomnia cases in a community sample (Morin et al. 2011). We anticipate extending recruitment to ISI 9 or ISI 8 (the sub-threshold insomnia cutoff, Bastien, Vallières and Morin 2001), if the eligible pool at 10 proves too small, since availability to attend weekly sessions matters as well as the score. Each eligible respondent is assigned to a single arm, and invited to the weekly sessions based on their expressed availability. Enrollment is defined as completing the Week-0 baseline, and we will report the conversion rates from screening to enrollment and corresponding attrition rates in the paper.

2.2 Randomization

We randomize at the individual level using a script with a fixed random seed, archived with its output. Based on the assigned arm, participants are invited to the corresponding group session. Sessions of both arms are spread across the same times of day and days of the week, so session timing does not differ systematically by arm. Once the pool of eligible participants is defined, strata are the cells of gender (2) $\times$ a screening-ISI split (2).[1] Within each stratum, participants are randomly ordered and paired, with one member of each pair assigned to treatment by a fair draw, so arms balance exactly within stratum; an unpaired participant from an odd-sized stratum is pooled and paired across strata by the same rule. We record which threshold each participant entered under, and the strata enter the estimation as fixed effects (Bruhn and McKenzie 2009).[2]

We collect enough data in the recruitment survey to determine eligibility and to randomize, so participants are assigned to a treatment arm and invited to a specific session group. The Week-0 baseline session is identical between the treatment and control subjects, and their treatment arm assignment will become clear only at the first weekly session or visit after the baseline. Attrition after the baseline therefore cannot be driven by knowledge of the arm.

Each treatment or control weekly session is split half and half into two separate rooms, and the sleep tracking devices rotate weekly between the rooms. From week 1, each participant's room assignment is fixed for the duration of the study, and participants are offered incentives to consistently

[1] The split is at the median of the eligible pool, which was 13 in the first batch.

[2] In the event eligibility is expanded to cutoffs 9 and 8 and randomized in separate batches, strata within these batches will be created by gender alone.

show up to their assigned group. Group, room, and facilitator identity are recorded for every session and visit, and we will control for them as additional covariates. We will report balance at baseline for enrolled and invited participants in the main paper.

2.3 Arms

Treatment: group CBT-I. The protocol is an adaptation of a CBT-I protocol developed and proven effective in high-income countries (Drummond et al. 2022) to a low-income population, with most changes driven by the necessary adjustments to a different living situation: some CBT-I components change sleep routines and the way the bedroom is used, while for our participants there is no separate bedroom for sleep, with shared living quarters and the bed also serving as the chair or day-use couch with limited other furniture. The protocol comes from our sleep-scientist co-author Sean Drummond, who developed a version of CBT-I in a high-income context and has experience adapting it to other settings.

Treated participants come to weekly sessions, watch pre-recorded videos, and have facilitator-led discussions in groups of about ten with lay, trained facilitators. The four sessions cover sleep restriction therapy, with a time-in-bed prescription computed from the baseline diary subject to a five-hour floor; stimulus control, adapted to shared sleeping spaces; sleep hygiene; and sleep and stress. Between sessions a message exchange checks in, reminds participants of the meeting, and repeats the two or three points discussed the previous week, nudging them to try the agreed upon adjustments if they have not yet done so.

Active control: participation matched on lab time, payment and group interaction. Payment is held constant between treatment and control arms: control participants come to the lab on the same weekly schedule as the treatment group and receive the same attendance compensation. They come primarily to complete surveys, and the weekly visits include no practice on the tasks measured at endline. Each visit opens and closes with a short group discussion, mimicking the format of the treatment groups but free of any new content. Control participants are told that the study is about how they live their lives. We ask them not to change their routines in response to the study. No control survey module at the four visits between baseline and endline is incentivized.

Time in the lab and the payments offered to treatment and control participants are the same, so both arms commit the same time to the study and are offered the same compensation. What varies are the four weekly sessions between baseline and endline: treatment participants receive cognitive behavioral therapy for insomnia in groups, and control participants primarily fill out surveys and take part in group discussions on general life topics. We limit potential contamination or spillovers by restricting the participants to one person per household and by spacing the households surveyed across the neighborhood, to reduce the probability of friends and relatives being included in the same study. Under this protocol, any remaining contamination between arms will only attenuate the intent-to-treat estimate, so we do not apply a correction.

2.4 Data collection

Data collection runs in five waves: screening; the Week-0 baseline, completed before participants know their arm; four weekly sessions or visits; the Week-5 endline; and remote follow-ups at 6 and 12 months.[3] Screening questions include demographics, employment, the ISI, a sleep quality-versus-quantity vignette (one of three versions, randomized), and the belief and norm items that provide the reference distributions for the incentivized beliefs at baseline. The baseline survey includes the following modules: household descriptives, a set of cognitive tasks, multiple-price lists eliciting economic preferences, sleep (including incentivized belief items), physical and mental health, and employment modules, each specified in Section 3. In Weeks 2 and 4, participants in both arms answer the insomnia and mental-health scales (ISI, GAD-7, PHQ-8). In Week 3, both arms repeat the work and earnings module from baseline. These are reported as midline outcomes. Separately, text reminders and call check-ins between sessions monitor sleepiness and wellbeing during the sleep restriction phase.

Participants complete the main survey twice throughout the study, once during the baseline period and once at endline. The Week-5 endline repeats the baseline instrument set, with several additions: (1) one additional real-life productivity task, a bracelet assembly task followed by questions on perceived performance and confidence, (2) the willingness-to-pay module, priced for another person (both arms, Section 3.9), (3) participants' own quality-versus-quantity choice, in the version seen at screening (both arms), and (4) a program-feedback module (treatment arm only). The long-run follow-ups, if funded, will administer the ISI, PHQ-8, GAD-7, the labor module, and the willingness-to-pay module.

There are two instruments, from which we collect data daily: the sleep diaries and the sleep-tracking devices from Week 0 to Week 5. (1) All participants will fill out a daily sleep diary on arm-specific forms that share every common item verbatim: minutes napped, hours of paid work, the seven nightly sleep questions, and a six-category disturbance checklist with an open 'other' line. The diary, thus, contains the self-reported record of when participants sleep, which we will compare with the device data in both arms. The control-arm diary adds four daytime routine items to complement the sleep questions to reduce the perceived focus on sleep topics among control participants. Participants receive the diary at baseline and start filling it in that week, so that by the time they come back the following week there is a full week of baseline sleep data for everyone. Payments are conditioned on data completion, such as wearing the device, and never on behavior. (2) Sleep trackers will contain data from every other week for each participants, with daily records for however many nights the participants were the trackers. Participants are offered incentives to wear devices at night, measures by valid nights with data recording in their weekly data exports.

There are fewer watches than participants, and because participants come in once a week, the watches rotate weekly between the two rooms of each session, so every participant wears one in alternating weeks. In any week half of each session's participants wear a device, in both arms, and with rotation

[3] The follow-up waves will be fielded if we are able to secure additional funding; otherwise, all main outcomes are those collected at the Week-5 endline.

we have objective sleep data on all participants. The rooms are the same size, and room membership is fixed from the baseline session and balanced within arm. The same watch is meant to return to the same participant on each wear week. Replacements are recorded, and device model is included as a covariate for device-measured outcomes. The data are downloaded during each weekly visit, which is at the limit of data storage capacity on these devices.

3 Outcomes and Predictions

We will estimate the impact of CBT-I on sleep, mental health, cognition, and economic outcomes. The primary outcomes that we are interested in are the insomnia score, device-measured sleep, the mental-distress measures, and cognition and productivity measured in the lab. These capture the main causal chain: treatment improves sleep quality and reduces insomnia, which then affects mental distress, the effort people put in, and their productivity.

Sleep measures include a self-reported component and an objective sleep-tracker one. The sleep-tracker measures include wake after sleep onset and sleep efficiency (Table 1). Labor-market measures outside the lab, economic preferences, the diaries, and beliefs are secondary. Secondary outcomes are the eight families below and the components of each primary index, and they are intended primarily as mechanisms for effects on the primary outcomes. Family-level outcomes, estimating equations, and variable definitions are fixed here. Component-level and mechanism analyses are pre-specified as secondary (Banerjee et al. 2020). We will run the specifications in Section 5 for every outcome variable unless stated otherwise. The primary timepoint is the Week-5 endline, and each family's prediction is directional. The 6- and 12-month follow-ups, if funded, add durability outcomes. For some secondary outcomes the prediction is a null, and a null would support the mechanism.

Table 1: Primary outcomes (Romano-Wolf adjusted jointly).

	Family	Primary outcome (Week-5 endline)
P1	Sleep, subjective	ISI total score (all participants)
P2	Sleep, objective	Device-measured wake after sleep onset (WASO) and sleep efficiency, each averaged over the valid nights of each participant's last wear week
P3	Mental health	Mental-health index: standardized symptom index of GAD-7 (anxiety) and PHQ-8 (depression); PSS-10 is secondary
P4	Cognition	Cognition index: standardized index of attention, working memory, mental arithmetic, and constrained optimization (PVT, n-back accuracy, arithmetic score, knapsack performance)
P5	Productivity	Performance on a novel task: standardized bracelet count at endline (correct bracelets in ten minutes)

Notes: Six primary tests, since P2 contributes two outcomes. ISI is the Insomnia Severity Index; PVT is the psychomotor vigilance task.

3.1 Sleep (P1, P2)

Measures. Our first primary outcome is the Insomnia Severity Index (ISI), an established survey measure of insomnia. The ISI questionnaire is a self-reported screening survey of perceived insomnia severity, validated against sleep diaries and polysomnography in clinical samples (Bastien, Vallières and Morin 2001) and against independently rated clinical improvement in community and clinical samples (Morin et al. 2011). It is included as one of the instruments named in the recommendations for a standard research assessment of insomnia (Buysse et al. 2006).

ISI assessment survey includes questions on how difficult the respondent finds falling asleep and staying asleep, whether the person wakes too early, and how much the problem affects daily life. Each question asks about the last 14 days, and each item is scored from "0" (no difficulty and not affected) to "4" ("very severe" difficulty or very affected), with higher values indicating greater severity. Scores on each question are summed, defining the insomnia severity score. This is the primary sleep-related outcome. The ISI total score is the sum of seven items scored 0 to 4 (0 to 28), where 0 represents the best outcome and 28 represents the worst outcome. Scores of 10 or above identify insomnia cases in a community sample with 86% sensitivity and 88% specificity, and a reduction of 8.4 points corresponds to moderate independently-rated improvement after treatment (Morin et al. 2011). The screening ISI enters only through the strata randomization.

We capture the participant ISI score at several points of the study: the recruitment survey (it is one of the eligibility criteria), the baseline and endline survey, and the midline Week-2 and Week-4 short surveys. We will use their baseline score as a control in regressions with controls.

The objective measures of sleep come from the wrist-band devices. Consumer wrist trackers detect sleep more reliably than wakefulness, which compresses measured efficiency toward its ceiling (Haghayegh et al. 2019; de Zambotti et al. 2018).

Wake after sleep onset is the minutes awake between falling asleep and the final awakening, and sleep efficiency is the share of the sleep record spent asleep, both calculated from the device's exported sleep record. Each outcome is the participant's mean over valid nights of their last wear week, which is a program week, balanced across arms by construction. We will winsorize WASO at the 1st and 99th percentiles of the distribution of all valid recorded nights.

The treatment also prescribes time in bed, the denominator of the efficiency ratio. Consolidation moves both outcomes; a change in time in bed with no change in wakefulness moves only efficiency. Valid nights and wear weeks are defined in Section 4. We aggregate the available nights into weekly averages, and the analysis compares the baseline week with the program weeks, which reduces the noise in any one week and the day-of-the-week effects of a single week.

The diary measures the same nightly constructs by self-report: sleep-onset latency is how many minutes it took the respondent to fall asleep; the number of awakenings throughout the night; wake-after-sleep onset (WASO) is the cumulative duration of these awakenings; time in bed measures the

interval from getting into bed to getting out of bed; total sleep time is that interval net of the time before trying to sleep, latency, WASO, and the interval from final waking to rising; and diary efficiency is the ratio of time asleep to time in bed. Some people over-report their own sleep and some under-report it, and our expectation is that under-reporting will be common among insomniacs. The gap between objective and subjective sleep reports is therefore itself an outcome: the diary-minus-device WASO difference, averaged over the participant's wear weeks, positive when the participant over-reports wakefulness, which is the sleep-misperception measure.

The secondary sleep outcomes are: the diary versions of WASO and efficiency; device-measured awakenings, total sleep time, time in bed, and regularity (the within-week standard deviation of sleep onset and of total sleep time); diary bedtime regularity and minutes napped; the ISI at Weeks 2 and 4; and the diary-minus-device WASO gap.

Analysis and predictions. Our primary first-stage measure will be insomnia severity on the ISI, with device-measured wake after sleep onset and sleep efficiency as the objective first stage. **H1 (first stage).** Treatment improves sleep quality: insomnia severity falls (P1), and device-measured wake after sleep onset falls while sleep efficiency rises (P2). Treatment does not raise total sleep time or time in bed. Because participants spend less time in bed, we expect both to fall slightly: if the treatment works through consolidation, we expect a negative effect for total sleep time and for time in bed compared to baseline. Downstream effects are attributed to sleep quality only if this first stage holds. The baseline covariate for the device outcomes is the Week-0 diary value of the same construct, available for every participant. In secondary analyses, we will further decompose treatment effects on sleep into two channels: consolidation (awakenings and regularity) and quantity (total sleep time and time in bed). We will also explore the impact of the treatment on self-reported sleep from the diary; the diary measures will be used to shed light on how participants perceive the changes in sleep, as compared to the objective device measures. **H2 (sleep misperception).** Baseline insomnia severity widens the diary-minus-device gap in reported wakefulness, a cross-sectional association; treatment reduces the gap.

3.2 Mental health (P3)

We measure participants' subjective well-being via three key survey questionnaires: (1) Generalized Anxiety Disorder scale (GAD-7) as a screener for anxiety; (2) Patient Health Questionnaire (PHQ-8) as a screener for depression; and (3) Perceived Stress Scale (PSS-10) as a measure of perceived stress. The GAD-7 and the PHQ-8 use the past two weeks and a four-point frequency running from not at all to nearly every day, the GAD-7 covering nervousness, uncontrollable worry, trouble relaxing, restlessness, irritability, and fear, and the PHQ-8 covering low mood, loss of interest, sleep, energy, appetite, self-worth, concentration, and psychomotor change; the PSS-10 uses the past month and asks how often the respondent felt that demands exceeded what they could control or cope with.

The Generalized Anxiety Disorder scale (GAD-7) is a brief self-report scale developed by the authors of the PHQ to identify probable cases of generalized anxiety disorder and to grade its severity,

validated in fifteen primary care clinics against independent diagnoses made by mental health professionals (Spitzer et al. 2006) and standardized in the general population (Löwe et al. 2008). Each item asks how often a symptom of anxiety occurred over the last 14 days, and each item is scored from "0" (not at all, a better outcome) to "3" (nearly every day, a worse outcome). Scores on each question are summed, defining anxiety severity. The outcome measure is the total summed score captured at endline (Week 5), with the Week-0 score as the baseline covariate and the Week-2 and Week-4 scores reported as midline outcomes. The outcome is thus measured on a scale from 0 to 21, where 0 represents the best outcome and 21 represents the worst outcome. Swahili versions have been validated in Kenyan samples (Nyongesa et al. 2020; Odero et al. 2023).

The Patient Health Questionnaire (PHQ) is a self-administered version of the PRIME-MD diagnostic instrument for common mental disorders (Kroenke, Spitzer and Williams 2001). We administer the eight-item form, the PHQ-8, which omits the ninth item of the PHQ-9 on thoughts of self-harm and is established as a measure of current depression in population-based studies (Kroenke et al. 2009). Each question asks how often a symptom of depression occurred over the last 14 days, and each item is scored from "0" (not at all, a better outcome) to "3" (nearly every day, a worse outcome). Scores on each question are summed, defining the severity of depressive symptoms. The outcome measure is the total summed score captured at endline (Week 5), with the Week-0 score as the baseline covariate and the Week-2 and Week-4 scores reported as midline outcomes. The outcome is thus measured on a scale from 0 to 24, where 0 represents the best outcome and 24 represents the worst outcome. The PHQ-9 has been validated in Kenyan samples, including in Swahili (Omoro et al. 2006; Monahan et al. 2009; Mwangi et al. 2020; Odero et al. 2023).

The Perceived Stress Scale (PSS) measures the degree to which a respondent appraises the situations in their life as unpredictable, uncontrollable, and overloading (Cohen, Kamarck and Mermelstein 1983). We administer the ten-item form, the PSS-10 (Lee 2012). Each question is about how often each experience occurred for the respondent over the last 30 days, and each item is scored from "0" (never) to "4" (very often), with the four positively worded items reverse-scored so that a higher score always represents a worse outcome. Scores on each question are summed, defining perceived stress. The outcome measure is the total summed score captured at endline (Week 5), with the Week-0 score as the baseline covariate. The outcome is thus measured on a scale from 0 to 40, where 0 represents the best outcome and 40 represents the worst outcome.

Mental Distress Score. The primary outcomes here are a combined index of GAD-7 and PHQ-8 measuring participants' level of mental distress (a higher score implies higher distress) and the converted binary mental distress indicator, which classifies a participant as distressed at the conventional threshold of 10 on either scale (Kroenke et al. 2009; Manea, Gilbody and McMillan 2012). PSS-10 and the individual scales are secondary, as are the Week-2 and Week-4 measures of GAD-7 and PHQ-8. The PHQ-8 item on sleep moves with the treatment's own first stage, so the index is also constructed from the seven remaining PHQ-8 items as a pre-specified robustness check.

Analysis and predictions. Our hypothesis is that CBT-I reduces insomnia and, thereby, reduces mental distress, in particular anxiety. Depression may take longer than five weeks to move. **H3 (mental health).** Treatment reduces mental distress (P3). Anxiety responds by endline; depression responds with a lag.

3.3 Cognition (P4)

Measures. We will estimate a cognition index (P4) based on the following cognitive tasks administered on tablets at the Week-0 baseline and the Week-5 endline. Each task is incentivized, and the score is the participant's own observed performance. They measure cognitive functions and skills that prior research on sleep and insomnia links to sleep loss:

1. **Sustained attention**, measured by the Psychomotor Vigilance Task (PVT), the standard measure of the attention deficits caused by poor sleep (Basner and Dinges 2011).

2. **Working memory**, measured by a spatial n-back: squares light up one at a time, and the participant reports whether the current square is the one that lit up n steps earlier. We run it at loads of one, two, and three squares (Kirchner 1958; Jaeggi et al. 2010).

3. **Applied numeracy**, measured by mental arithmetic set in monetary scenarios, adapted from Duquennois (2022).

Scoring the PVT. In the PVT a stimulus appears on the screen at random intervals and the participant presses a key as soon as they see it. Each trial yields a reaction time (RT) in milliseconds (1000 milliseconds in a second), and we score the trial as 1000/RT, which converts it into a response speed per second: a reaction time of 250 milliseconds scores 4, one of 500 milliseconds scores 2, so a higher score means a faster response.[4] The participant's PVT score is the mean of these trial scores. The n-back score at each load is the share of scored trials answered correctly. Mental arithmetic runs in two blocks, both set in monetary scenarios, one framed with financial strain (debt, money-lending, shortfalls) and one in neutral spending terms, with block order randomized; the P4 component is the number correct across both blocks, and the strain-framed minus neutral difference is the secondary outcome behind H5. The knapsack score is the value ratio: for each instance, the value of the submitted bundle relative to the best attainable value, averaged over instances. P4 is the standardized index of the four components; each component is also secondary.

The PVT matches one of the performance tasks from the Chennai experiment (Bessone et al. 2021), enabling us to perform a direct comparison of the two sleep interventions with economic outcomes. It is also the least effortful of the four tasks, measuring instantaneous attention. That makes it a useful complement if the mechanism is effort, as we can compare the effect on the PVT with the effects on

[4]Response speed is the recommended summary for this task. Mean reaction time is dominated by a small number of very slow responses, and Basner and Dinges (2011) find that speed-based and lapse-based metrics separate sleep-deprived from rested participants far more sharply, advising against mean reaction time as a primary measure.

the tasks that require more effort. We do not expect an effect on the PVT, but do expect effects on more effortful tasks.

Perceived performance. After the arithmetic task, and after the bracelet task at endline, participants report how many items they think they answered correctly and how confident they are. We expect participant with poorer sleep to have more negative self-perceived performance. Where an objective count exists, the perceived-minus-actual gap is a secondary outcome. It is the cognitive counterpart of the diary-minus-device gap in Section 3.1 is a secondary outcome. We do not ask these questions on the knapsack or the n-back, where the participant has no natural way to count their own successes. Post-task ratings of effort, relative performance, and enjoyment follow every task and are specified in Section 3.6.

Analysis and predictions. H4 (cognition). Treatment improves cognitive performance (P4). *Pre-specified secondary within H4:* (1) gains rise with task difficulty: no effect at the lowest n-back load, growing across the three loads. A uniform improvement at every load would indicate engagement. *Pre-specified secondary within H4:* (2) treatment improves how accurately participants judge their own performance. Where an objective count exists, on the arithmetic task and on the bracelet task at endline, we test the perceived-minus-actual gap and its absolute value, and we predict that the absolute gap falls. **H5 (financial-strain arithmetic).** Financial-strain framing lowers arithmetic performance: performance is lower on the strain-framed block than on the neutral-money block; the gap is larger at a lower baseline socioeconomic index (Section 7) and worse baseline sleep; a strain-first block order lowers performance on the neutral block that follows.

3.4 Performance & Productivity (P5)

Measures. P5 measures performance on effort-inducing tasks which mimic work-setting-related activities with measured accuracy: a manual task of bracelet assembly task, digital (tablet-based) constrained-optimization task, and task follow-through outside the lab measured by weekly time diary completion.

1. **Novel Manual Productivity Task** – bracelet assembly:

   The participant follows a pattern given in a picture and assembles as many bracelets as they can in ten minutes, paid a piece rate. The score is the number of correctly assembled bracelets: incorrect ones are counted and not paid, so accuracy affects earnings. Bracelet assembly is administered only at endline, so participants attempt it for the first time after the program has run. Neither arm has practiced it, so this task is not subject to potential (differential) learning effects compared to the cognitive tasks in P4.

2. **Constrained optimization**, measured by the knapsack task: the participant chooses among items with given values and weights to fill a bag of fixed capacity (Murawski and Bossaerts 2016). We run it at an easy and a hard level. With the knapsack, we will be able to measure participants' planning ability. The trial-level data records click order, working selections, and timestamps, which we will use to compute the time before the first item is selected, which is planning latency,

how many times a person revises a partial solution, and how sharply performance falls as instance complexity rises.

3. **Follow-Through Outside the Lab – Weekly Sleep Diary Completion** – we use the weekly diary completion that needs to be done outside the lab to capture follow-throughou on an uncued task as sleep loss impairs self-initiated actions more than it impairs performance on a task someone is actively prompting you through. We will record whether the participants brought the time diaries at the assigned times, how many days of the week were recorded, and how complete those daily records were.

Participants work in different types of jobs and occupations, so no single work task would be relevant to all of them. Bracelet assembly represents the kinds of tasks that participants involved in manual work do, and their performance captures attention to detail, focus, and manual skills overall. Constrained optimization is relevant for participants engaged in their own businesses and in self-employment, being related to accounting and planning.

Perceived performance is recorded alongside the bracelet count, as specified in Section 3.3: while the enumerator takes the box away and checks it, the participant reports how many bracelets they think are correct and how confident they are.

Analysis and predictions. P5 includes the bracelet counts and errors at endline, performance on the knapsack, and diary completion score. We will test the effects on task-specific productivity, as well as a combined index of P5 outcomes. **H6 (economic returns):** Treatment raises incentivized task output (P5) – treated participants assemble more correct bracelets at endline and make fewer mistakes; their have a higher performances on knapsack, and higher diary completion.

3.5 Labor-market activity

Measures. The work module repeats the baseline questions: how many days and hours the participant worked in the past seven days, whether those without work tried to find a new job, and, for everyone, income-generating activity. The Week-3 work module repeats the same survey items in both arms. The family has seven components, adjusted within family:

1. employment status;

2. weekly hours, days worked in the last seven days times hours on a typical day;

3. total personal earnings in the last week, collected in six bands and analyzed as band midpoints in inverse-hyperbolic-sine form;

4. the number of job-search methods used in the last seven days, asked of those searching for work and set to zero for the rest;

5. workdays missed to poor health in the last two weeks;

6. any paid work on a given day, from the daily diary;

7. hours worked on a given day, from the daily diary (Section 5.2).

Analysis and predictions. H7 (labor-market activity). Treatment raises labor-market activity; week five endline effects are expected to be small. If the follow-ups are funded, the labor components at 6 months are the durability outcomes.

3.6 Cost of output

Measures. Poor sleep can leave output unchanged while the effort behind it rises: a worker holds performance steady by working harder. Three measures capture the costs.

Post-task effort is a single item on a five-point scale, asked after each task, following the cognitive-load scale of Paas (1992). Each task contributes one component.

Daytime impairment is the sixteen-item PROMIS sleep-related impairment scale, which asks about functioning during the day and is scored as the raw sum.

The effort reservation wage is elicited with an incentivized price list, run after the participant has done both a low-load and a high-load version of the n-back, so they know from experience what each level of the task costs them. Each row offers a lower rate for the low-load version against a fixed rate for the high-load one, and the participant chooses row by row. The outcome is the effort premium, the gap between the two rates at the row where the participant switches, higher meaning more effort-averse. It is coded at the first switch, with multiple switchers flagged and an indicator for each censored case. The elicitation follows the fixed-versus-variable choice design of Dohmen and Falk (2011).

Analysis and predictions. H8 (cost of output). Treatment lowers the cost of producing output, even where output does not rise: post-task perceived effort, daytime impairment, and the effort reservation wage fall. The framework predicts that workers defend output by spending costly effort: output can stay flat while its cost falls.

3.7 Beliefs and norms

Measures. The belief module has two parts. The first is what participants believe about other people: how much others in the community sleep, and how others regard sleep and help-seeking. The second is what they believe about sleep itself, and about their own capacity to change it.

Beliefs about the community are incentivized. Six items ask the participant how many adults out of ten sleep at least seven hours, regularly give up sleep, believe seven hours is important, would look down on someone seeking help for sleep, would look down on someone seeking help for mental health, and chose the interrupted night in the vignette. Participants are paid for an exact match against the true community value on each item, which we compute from the recruitment survey, a large screening sample drawn from the same population before the trial began. The outcome is the absolute distance

between the guess and that value, so a smaller distance means a more accurate belief. Because we also measure how much each participant sleeps, a person's own behavior and their perception of what others do can both be compared with what others report.

The stigma gap uses two of those items. The same "out of ten" question is asked about looking down on someone who seeks help for sleep and about someone who seeks help for mental health, and the gap between the two answers is the measure. One argument for delivering a sleep program is that take-up may exceed that of a mental-health program because sleep attracts less stigma, and this gap tests that argument.

Beliefs about sleep. Process beliefs: the six-item adapted Dysfunctional Beliefs and Attitudes about Sleep module (DBAS-6), on a five-point agreement scale, tested as the six-item mean and as the single fixed-hours item, "I must get at least 7 hours of sleep to be healthy"; and sleep self-efficacy, the mean of three items on importance, confidence, and commitment. Returns beliefs: nine agreement items on the perceived returns to sleep, tested as their mean.

The quality-versus-quantity vignette is fielded at screening in three randomized versions and repeated at endline in the version the participant saw at screening. Versions are pooled with version indicators, and the outcome is the share choosing the shorter but uninterrupted night, with the screening choice as the baseline covariate.

Norms are the "out of ten" items above, one unincentivized item on whether it is acceptable to prioritize sleep even at the cost of socializing, working, or chores, and three agreement statements on stigma and help-seeking. Five-point agreement items are analyzed as agree/disagree binaries (agree = 4 or 5 of 5) in the main tables, with means and alternative cuts in an appendix.

Analysis and predictions. H9 (beliefs). Treatment corrects process beliefs (DBAS, the fixed-hours item) and raises self-efficacy, more than it changes perceived returns; the vignette choice shifts toward the consolidated night; endline willingness to pay for others to receive the program is higher in treatment than in control (Section 3.9).

3.8 Economic preferences

Measures. Risk aversion and time preference from incentivized multiple price lists: a five-row risk list, a sure amount against a coin flip, and a six-row time list, an amount today against a larger amount in one month. One of the two lists is drawn and one row of it is paid. The outcome is the switching point, with never-switchers coded at the boundary and multiple-switchers identified and excluded from the switching-point measure (retained in a consistency indicator).

Analysis and predictions. H10 (preferences). Treatment reduces impatience (the time-list switch row); effects on risk aversion are two-sided (no directional prediction).

3.9 Demand for the program

Measures. One question is whether people with job uncertainty care about their own sleep at all. Subjective questions answer it in part; the willingness-to-pay module answers it with a choice that is implemented. Because control participants will not have seen the program and treated participants will have, the endline does not ask what a participant would pay for their own program; it asks what they would be willing to give up so that someone else receives it. The object is the program materials, the recorded session videos shared by link and the printed booklet, which can be given to another person. Each participant, in both arms, chooses row by row between cash for that person and the materials for that person, and the switch point is the cash value placed on the materials for someone else. The recipient is a real person screened for the trial, eligible, not enrolled, who agreed to be contacted again. The choice is incentive-compatible because it is implemented: after all endline measurement, three participants are drawn at random, one row of each of their lists is drawn, and the recipient receives that row's cash or the materials. The endline treatment-control contrast is the effect of having experienced the program on its valuation for others, and the level in both arms speaks to the general demand for sleep interventions in this population. The list has seven rows.

Analysis and predictions. The prediction is stated with H9. The cash equivalent is analyzed in levels.

3.10 Sleep behaviors

Measures. Diary-based behaviors (bedtime regularity and naps, Section 3.1) and survey-based behaviors from the baseline and endline survey: cups of caffeinated tea or coffee after noon in the past week, energy drinks after noon in the past week, and eating within two hours of bedtime and eating heavy or spicy food within two hours of bedtime, the last two as often-or-always binaries.

Analysis and predictions. The expected direction is a fall in evening stimulant use and in eating near bedtime.

3.11 Social support and affect

Measures. Social support on the Multidimensional Scale of Perceived Social Support (MSPSS), twelve items on a five-point scale; affect (13 binary items recalled for the previous day), ten negative and three positive, analyzed as the count of negative items endorsed with the positive count as a second test.

Analysis and predictions. H11 (mood and social support). Treatment improves affect and perceived social support (MSPSS).

3.12 Financial strain

Measures. Borrowing in the past month: we measure whether the participant tried to borrow, whether they were successful, the number of times, and the amount. We also measure whether the participant holds personal savings, and the savings band.

Analysis and predictions. H12 (financial stress). Treatment eases financial strain: borrowing incidence and amounts fall and savings rise. Effects are more likely in the longer term than at the Week-5 endline.

4 Variable Definitions

We apply the following four rules to every outcome.

- **Indices.** We use standardized indices to put components on a common scale. Indices are equal-weighted means of components standardized against the control-group endline distribution, signed so that higher is better, with the exception of the mental-health index (P3), where a higher score indicates more mental distress; the baseline value of an index is built with the same moments; inverse-covariance (Anderson) weighting is the pre-specified robustness construction (Anderson 2008; Kling, Liebman and Katz 2007).

- **Missing values.** A task the participant never attempted is treated as missing. Within an attempted scored task, an unanswered item earns no credit. Survey and scale items left unanswered are missing. A scale score is the sum of its items. Where items are blank, the sum of the answered items is rescaled to the full number of items, provided at least half the items were answered; otherwise the score is missing. An index is computed when at least half of its components are observed.

- **Validity filters.** Device, diary, and task records pass through validity filters before entering any outcome. Filters are set without reference to treatment assignment and applied identically across arms. Records failing a filter are treated as missing on that outcome and are never imputed.

- **Winsorizing and outliers.** Device-measured WASO is winsorized at the 1st and 99th percentiles of the control-arm night distribution (Section 3.1); no other outcome is winsorized. Earnings (band midpoints) and the amount borrowed are analyzed in inverse-hyperbolic-sine form, with levels reported alongside; the coefficients can be interpreted as the standard log transformation away from zero. Bounded list variables (the effort premium, the cash equivalent, and the switch rows) are analyzed in levels.

5 Estimation and Inference

5.1 Primary specification (ITT)

Our experiment includes one treatment arm and one control arm. All treatment effects are estimated from the random assignment, with the treatment indicator as the main independent variable. The main regressions are ANCOVA regressions at the individual level, one observation per participant, which control for the baseline value of the outcome (where it exists) and for other individual baseline characteristics (such as age or gender). For participant i in randomization stratum s, endline outcome

$Y_{i,1}$ and baseline value $Y_{i,0}$:

$$Y_{i,1} = \alpha + \tau T_i + \beta Y_{i,0} + X_i'\gamma + \delta_s + \varepsilon_i, \tag{1}$$

where T_i is assignment, δ_s are strata fixed effects, and X_i is a vector of variables measured at the baseline; it always includes the randomization-batch indicator and, for device-measured outcomes, the rotation-half indicator (which alternating weeks the participant wears a device) and device model. For each regression, we will specify which other variables are considered, if any. We will run one specification with the demographic control variables, and one without them; the specification without them keeps the strata fixed effects only ($\beta = 0$, $\gamma = 0$). The ANCOVA specification was selected based on the power calculations in McKenzie (2012), in which the ANCOVA specification achieves higher power than difference-in-differences when outcomes are not highly autocorrelated. τ is the intent-to-treat effect.

5.2 High-frequency panel specification

The trial produces two high-frequency panels: the nightly device panel (wear weeks) and the daily diary panel (all participants, all days). For each high-frequency outcome, treatment and control means are plotted across the study window by week. The Week-0 segment is pre-treatment in both arms, so the two trajectories are expected to coincide there up to noise. The estimator behind the plots, for a night- or day-level outcome y_{it} of participant i on study day t:

$$y_{it} = \alpha + \tau T_i + \delta_s + \lambda_{w(t)} + \varepsilon_{it}, \tag{2}$$

where $\lambda_{w(t)}$ are week-of-study fixed effects and δ_s is as in Equation 1. This unadjusted form is reported for every panel outcome, alongside an adjusted form adding the participant's Week-0 mean of the outcome and covariates appropriate to the outcome, listed with each exhibit. For all panel regressions, we will cluster standard errors at the participant level, with room-clustered standard errors reported as a robustness. Replacing τT_i with week-by-week interactions $\sum_w \tau_w (T_i \times \lambda_w)$ gives the event-study coefficients plotted with the raw trajectories; the pooled τ over Weeks 1 to 5 is the summary panel treatment effect. The panel specification depends on data quality: if the nightly or daily records are too sparse or too noisy, the measures are aggregated to weekly means before estimation. The protocol predicts that time in bed falls early in the program before consolidation gains appear. Device-panel regressions include only device-wear nights and the rotation-half indicator; diary-panel regressions cover all days. Daily paid work is analyzed with Equation 2 as two outcomes, both pre-specified secondary components of the labor-market activity family: an indicator for any paid work that day, and hours worked, which will be counted as zero if the participant does not work that day, since a day with no paid work is a true zero (inverse-hyperbolic-sine form).

5.3 Inference

Standard errors are heteroskedasticity-robust at the individual level, which is the level at which treatment was assigned. Because both arms are delivered in fixed rooms, outcomes may be correlated within a room, so room-clustered standard errors are reported for robustness.

5.4 Compliance, adherence, and treatment-on-the-treated

Participants who do not attend the baseline are not part of the trial. Everyone who completes the baseline is in the analysis sample, whatever they attend afterwards, and is invited through the program. A participant who misses a weekly session is invited to attend a session of the same arm later in the week; this is allowed only a few times, and participants are offered an attendance bonus for attending most sessions of their assigned group on time. Attendance is recorded for every session, and in the final analysis the number of sessions attended is the measure of treatment intensity. Take-up is defined as attending at least three of the four sessions. TOT is estimated by 2SLS, instrumenting sessions attended (0 to 4, and structurally zero in the control arm) with assignment. It is interpreted as the effect per CBT-I session, under the assumption that assignment affects outcomes only through the content of those sessions. Adherence is measured by device-based time-in-bed contraction and wake-time regularity on wear weeks, their diary equivalents for all participants, session attendance, and the time-in-bed prescription recorded per participant.

5.5 Attrition & Missing data

Participants attend on fixed session days and times, and participants who choose morning sessions may differ from those who choose evening sessions, so attrition is expected to vary by session day and time as well as by arm. When a participant misses a session, the check-in records the reason, and the participant keeps their place in the rotation, with a spare watch covering the participant who would have received theirs that week.

We will report the following on attrition: First, attrition rates by arm with the arm-difference test, baseline predictors of attrition, and attrition by scheduled session day, with session-day indicators available to the adjusted specifications. Second, where attrition differs across arms, Lee bounds on the primary effects, trimming the better-retained arm's most extreme outcomes until retention is equal across arms (Lee 2009). Third, inverse-probability-weighted estimates as a robustness check, with weights taken from baseline covariates. Fourth, for device-measured outcomes, the device-retention rate covering loss, damage, and non-wear.

We will not be imputing any missing values, but will aggregate the higher-frequency data over weeks to reduce the daily missings, and aggregate weeks to pull in the full sample once everyone has works the sleep trackers for a week.

6 Multiple Testing

- We will adjust p-values for multiple hypothesis testing within groups of primary outcomes using the Romano-Wolf step-down procedure (Romano and Wolf 2005). Both unadjusted and adjusted p-values are reported.

- Within each family, secondary components are adjusted with the same Romano-Wolf step-down procedure, over the components listed in the family's subsection.

- Heterogeneity interactions (Section 7) are adjusted within their own set, with the same procedure.

- Coefficients and standard errors are reported throughout; inference on the primary outcomes uses the adjusted p-values.

7 Heterogeneity

To test heterogeneity, we will include additional interactions, and we state for each specification how the interaction was constructed. We will estimate additional specifications to explore heterogeneous treatment effects on the primary outcomes by gender, by baseline insomnia severity and the baseline mental-distress measures, by the baseline socioeconomic index, by employment status and type, by family composition, and by baseline beliefs about sleep, using interactions and subsamples where appropriate.

8 Power

Minimum detectable effects are computed for 80% power, 5% two-sided tests, 1:1 allocation, an enrollment target of 300 with 15% endline attrition, and a group-delivery design effect at the room level, with a room of about eleven participants and a within-room intraclass correlation of 0.02 to 0.05.

Randomization is individual. Delivery is not: from Week 1 each participant sits in a fixed room with the same facilitator and the same group for the rest of the study, in both arms. Outcomes correlate within rooms, so the calculation applies a room-level design effect.

P1 to P4 are measured at both baseline and endline and are estimated by ANCOVA with an assumed baseline-endline correlation of 0.6. On these assumptions the full-sample minimum detectable effect is roughly 0.31 to 0.34 standard deviations, the lower figure at an intraclass correlation of 0.02 and the upper at 0.05.

P5 is measured only at endline, so it is estimated without a baseline value of its own and the ANCOVA gain does not apply. Its minimum detectable effect is larger by a factor of $1/\sqrt{1-0.6^2} = 1.25$, roughly 0.38 to 0.43 standard deviations over the same range. A treatment effect on bracelet output smaller than about four-tenths of a standard deviation is therefore unlikely to be detected.

Subgroup interactions are detectable only if large.

9 Ethics and Cost-Effectiveness

The trial holds approval from the Meru University Research Ethics Committee (MIRERC), a NACOSTI research permit, and the UC San Diego IRB. Written informed consent is obtained at screening and again at the Week-0 baseline. Adverse events and study-discontinuation reasons are recorded under the approved distress and adverse-event procedure and reported for both arms regardless of statistical significance.

Cost-effectiveness is reported in brief. The cost of the intervention per participant is set against the gains in sleep and mental health, valued as public-health studies value a reduction in insomnia or in depression and anxiety symptoms, and it is compared with the cost of other interventions, such as general CBT programs and the sleep interventions in Chennai (Bessone et al. 2021), and with what participants are willing to give up so that others receive the program (Section 3.9).

10 Amendments and Deviations

Any change to this plan after it is filed and before endline data are seen will be made as a dated addendum in the registry, with the reason. Once endline data have been seen, we will not make further changes to this pre-analysis plan. We will report which analyses are in line with or deviate from the PAP.

Appendix A: Design Summary

Table 2: Trial design.

Design	Individually randomized controlled trial, two arms, 1:1 allocation
Population	Adults 18 to 45 with insomnia symptoms, employed or actively job-searching, recruited in Nairobi informal settlements (primarily Kibera), one adult per household; screening ISI ≥ 10, extendable to ISI 9 and 8 by whole batches
Treatment	Four weekly 45-minute group CBT-I sessions (lay facilitators, video modules)
Control	Active control: weekly lab visits on the same schedule (no insomnia or CBT content)
Randomization	Individual, 1:1, in batches after screening closes, by a script with a fixed random seed; strata are gender $\times$ a screening-ISI split (gender only where a batch cannot support the split); batch indicator in the adjusted specifications; household-linked participants share one arm; concealed until the Week-0 baseline is complete
Objective sleep	Watches rotating weekly between the two fixed rooms of each session group, Weeks 0 to 5; every participant wears in alternating weeks; device model fixed within participant and a covariate
Waves	Screening; Week-0 baseline; two interim contacts and one work module during the four weekly sessions; Week-5 endline; 6- and 12-month remote follow-ups
Primary outcomes	Five, listed in Table 1, giving six primary tests
Registration	AEA RCT Registry, AEARCTR-0018309; ethics approvals from MIRERC (Kenya), NACOSTI, and the UC San Diego IRB

Notes: A summary of Section 2, whose prose is the registered statement. ISI is the Insomnia Severity Index.

Appendix B: Outcome Dictionary

Table 3: Outcome dictionary: primary (P) and secondary (S) outcomes.

Outcome	Status	Wave	Source	Construction and direction
ISI total (P1)	P	BL, EL, 6m, 12m	Survey	Sum of seven items scored 0 to 4 (0 to 28); rescaled if at least half the items are answered, else missing; lower is better
Wake after sleep onset (P2)	P	Weeks 0 to 4	Device (all participants)	Minutes awake between sleep onset and final awakening in the main sleep record; mean over valid nights of the last wear week; winsorized 1/99 on control-arm nights; lower is better
Sleep efficiency (P2)	P	Weeks 0 to 4	Device (all participants)	Minutes asleep over the duration of the main sleep record; mean over valid nights of the last wear week; higher is better
Mental-health index (P3)	P	BL, EL	Survey	Standardized symptom index of z(GAD-7) and z(PHQ-8); higher = more distress; robustness version drops the PHQ-8 sleep item

Table 3, continued

Outcome	Status	Wave	Source	Construction and direction
Cognition index (P4)	P	BL, EL	Incentivized tablet tasks	z-index of PVT mean response speed (1000 divided by reaction time in milliseconds, so higher is faster), mean n-back proportion correct over three loads, arithmetic number correct, knapsack value ratio (submitted over optimal value, averaged over instances); higher is better
Performance on a novel task (P5)	P	EL	Incentivized output task	Bracelet count (correct bracelets in ten minutes), standardized; task is new to participants at endline, so there is no baseline measurement; higher is better
GAD-7; PHQ-8	S	BL, interim, EL, 6m	Survey	Item sums, rescaled if at least half the items are answered, else missing; lower is better
PSS-10	S	BL, EL	Survey	Item sum with the four positive items reversed; lower is better
PHQ-8 at 6 months (durability)	S	6m	Survey	6-month administration; lower is better
Device awakenings, TST, TIB, regularity	S	Weeks 0 to 4	Device	Per wear week over valid nights; regularity is the within-week SD of sleep onset and of TST; TST and TIB predicted not to rise
Diary WASO, efficiency, latency, awakenings, bedtime regularity, naps	S	Weeks 0 to 5	Diary (all)	From the seven nightly items and the nap item, as mapped in Section 3.1
Diary-minus-device WASO gap (misperception)	S	Weeks 0 to 4	Diary × device	Per wear night, averaged; positive means over-reported wakefulness; lower is better
Labor-market activity (seven components)	S	BL, interim, EL, 6m; diary daily	Survey + diary	As defined in Section 3.5; earnings as band midpoints (IHS); Romano-Wolf-adjusted within family
Cost of output: perceived effort, PROMIS sleep-related impairment, effort premium	S	BL, EL	Survey + MPL	Effort item per task; PROMIS sixteen-item sum; effort premium, high-load minus low-load rate at the n-back switch, higher = more effort-averse
Task-level components (PVT lapses and inverse median RT, n-back by load, arithmetic by block, knapsack by difficulty level)	S	BL, EL	Incentivized tablet tasks	Unattempted tasks treated as missing
Perceived performance and confidence; post-task effort, relative-performance, and enjoyment ratings	S	BL, EL	Tasks + survey	Perceived-minus-actual gaps where objective counts exist
DBAS-6 mean; fixed-hours item; self-efficacy	S	BL, EL	Survey	Five-point agreement items; self-efficacy as the three-item mean; Likert items reported as agree/disagree binaries
Returns beliefs (nine items)	S	BL, EL	Survey	Mean of nine agreement items
Community beliefs (six incentivized items)	S	BL, EL	Survey	Absolute distance between the "out of 10" guess and the screening truth; lower is better

Table 3, continued

Outcome	Status	Wave	Source	Construction and direction
Vignette choice	S	Screening, EL	Survey	Share choosing the shorter uninterrupted night, pooled with version indicators; screening choice as the baseline covariate
Norms and stigma	S	BL, EL	Survey	The "out of 10" norm items, the injunctive item, and three stigma and help-seeking statements
Risk aversion; impatience	S	BL, EL	MPL	Switch rows of the five-row risk list and the six-row time list; never-switchers at the boundary; multiple-switchers excluded
Willingness to pay for others to receive the program	S	EL, follow-ups	Incentivized MPL	Cash equivalent: bracket midpoint at the switch, censored at both ends, with indicators; higher = values the package more
Sleep behaviors	S	BL, EL	Survey	Caffeinated drinks and energy drinks after noon (counts); eating near bedtime (often-or-always binaries)
MSPSS social support; affect	S	BL, EL	Survey	MSPSS mean of twelve items; affect as the count of negative items endorsed
Financial strain: borrowing items; savings	S	BL, EL, 6m	Survey	Amount borrowed in inverse-hyperbolic-sine form; borrowing predicted to fall and savings to rise

Notes: BL is the Week-0 baseline, EL the Week-5 endline, 6m and 12m the remote follow-ups, and interim the two mid-program contacts. MPL is a multiple price list, IHS the inverse hyperbolic sine, TST and TIB total sleep time and time in bed, SD the standard deviation. Where a rule here and its family subsection differ, the subsection governs. Indices follow Section 4.

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.

Banerjee, Abhijit, Esther Duflo, Amy Finkelstein, Lawrence F. Katz, Benjamin A. Olken, and Anja Sautmann. 2020. "In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plans for RCTs in Economics." National Bureau of Economic Research Working Paper 26993.

Basner, Mathias, and David F. Dinges. 2011. "Maximizing Sensitivity of the Psychomotor Vigilance Test (PVT) to Sleep Loss." *Sleep*, 34(5): 581–591.

Bastien, Célyne H., Annie Vallières, and Charles M. Morin. 2001. "Validation of the Insomnia Severity Index as an outcome measure for insomnia research." *Sleep Medicine*, 2(4): 297–307.

Bessone, Pedro, Gautam Rao, Frank Schilbach, Heather Schofield, and Mattie Toma. 2021. "The Economic Consequences of Increasing Sleep among the Urban Poor." *Quarterly Journal of Economics*, 136(3): 1887–1941.

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.

Buysse, Daniel J., Sonia Ancoli-Israel, Jack D. Edinger, Kenneth L. Lichstein, and Charles M. Morin. 2006. "Recommendations for a Standard Research Assessment of Insomnia." *Sleep*, 29(9): 1155–1173.

Cohen, Sheldon, Tom Kamarck, and Robin Mermelstein. 1983. "A Global Measure of Perceived Stress." *Journal of Health and Social Behavior*, 24(4): 385–396.

de Zambotti, Massimiliano, Aimée Goldstone, Stephanie Claudatos, Ian M. Colrain, and Fiona C. Baker. 2018. "A validation study of Fitbit Charge 2 compared with polysomnography in adults." *Chronobiology International*, 35(4): 465–476.

Dohmen, Thomas, and Armin Falk. 2011. "Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender." *American Economic Review*, 101(2): 556–590.

Drummond, Sean P. A., et al. 2022. "ARISE Clinician Manual: Cognitive Behavioral Therapy for Insomnia (I-CBT-I)." Clinical protocol manual; basis for the 4-session group adaptation used in this study.

Duquennois, Claire. 2022. "Fictional Money, Real Costs: Impacts of Financial Salience on Disadvantaged Students." *American Economic Review*, 112(3): 798–826.

Haghayegh, Shahab, Sepideh Khoshnevis, Michael H. Smolensky, Kenneth R. Diller, and Richard J. Castriotta. 2019. "Accuracy of Wristband Fitbit Models in Assessing Sleep: Systematic Review and Meta-Analysis." *Journal of Medical Internet Research*, 21(11): e16273.

Jaeggi, Susanne M., Martin Buschkuehl, Walter J. Perrig, and Beat Meier. 2010. "The Concurrent Validity of the N-Back Task as a Working Memory Measure." *Memory*, 18(4): 394–412.

Kirchner, Wayne K. 1958. "Age Differences in Short-Term Retention of Rapidly Changing Information." *Journal of Experimental Psychology*, 55(4): 352–358.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica*, 75(1): 83–119.

Kroenke, Kurt, Robert L. Spitzer, and Janet B.W. Williams. 2001. "The PHQ-9: Validity of a brief depression severity measure." *Journal of General Internal Medicine*, 16(9): 606–613.

Kroenke, Kurt, Tara W. Strine, Robert L. Spitzer, Janet B.W. Williams, Joyce T. Berry, and Ali H. Mokdad. 2009. "The PHQ-8 as a measure of current depression in the general population." *Journal of Affective Disorders*, 114(1-3): 163–173.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies*, 76(3): 1071–1102.

Lee, Eun-Hyun. 2012. "Review of the Psychometric Evidence of the Perceived Stress Scale." *Asian Nursing Research*, 6(4): 121–127.

Löwe, Bernd, Oliver Decker, Stefanie Müller, Elmar Brähler, Dieter Schellberg, Wolfgang Herzog, and Philipp Yorck Herzberg. 2008. "Validation and Standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the General Population." *Medical Care*, 46(3): 266–274.

Manea, Laura, Simon Gilbody, and Dean McMillan. 2012. "Optimal Cut-Off Score for Diagnosing Depression with the Patient Health Questionnaire (PHQ-9): A Meta-Analysis." *Canadian Medical Association Journal*, 184(3): E191–E196.

McKenzie, David. 2012. "Beyond Baseline and Follow-up: The Case for More T in Experiments." *Journal of Development Economics*, 99(2): 210–221.

Monahan, Patrick O., Enbal Shacham, Michael Reece, Kurt Kroenke, Willis Owino Ong'or, Otieno Omollo, Violet Naanyu Yebei, and Claris Ojwang. 2009. "Validity/Reliability of PHQ-9 and PHQ-2 Depression Scales Among Adults Living with HIV/AIDS in Western Kenya." *Journal of General Internal Medicine*, 24(2): 189–197.

Morin, Charles M., Geneviève Belleville, Lynda Bélanger, and Hans Ivers. 2011. "The Insomnia Severity Index: Psychometric Indicators to Detect Insomnia Cases and Evaluate Treatment Response." *Sleep*, 34(5): 601–608.

Morin, Charles M., Richard R. Bootzin, Daniel J. Buysse, Jack D. Edinger, Colin A. Espie, and Kenneth L. Lichstein. 2006. "Psychological and behavioral treatment of insomnia: update of the recent evidence (1998–2004)." *Sleep*, 29(11): 1398–1414.

Murawski, Carsten, and Peter Bossaerts. 2016. "How Humans Solve Complex Problems: The Case of the Knapsack Problem." *Scientific Reports*, 6: 34851.

Mwangi, Paul, Moses K. Nyongesa, Hans M. Koot, Pim Cuijpers, Charles R. J. C. Newton, and Amina Abubakar. 2020. "Validation of a Swahili Version of the 9-Item Patient Health Questionnaire (PHQ-9) Among Adults Living with HIV Compared to a Community Sample from Kilifi, Kenya." *Journal of*

Affective Disorders Reports, 1: 100013.

Nyongesa, Moses K., Paul Mwangi, Hans M. Koot, Pim Cuijpers, Charles R. J. C. Newton, and Amina Abubakar. 2020. "The Reliability, Validity and Factorial Structure of the Swahili Version of the 7-Item Generalized Anxiety Disorder Scale (GAD-7) Among Adults Living with HIV from Kilifi, Kenya." *Annals of General Psychiatry*, 19(1): 62.

Odero, Sabina Adhiambo, Paul Mwangi, Rachel Odhiambo, Brenda Mumbua Nzioka, Constance Shumba, Eunice Ndirangu-Mugo, and Amina Abubakar. 2023. "Psychometric Evaluation of PHQ-9 and GAD-7 Among Community Health Volunteers and Nurses/Midwives in Kenya Following a Nation-Wide Telephonic Survey." *Frontiers in Psychiatry*, 14: 1123839.

Omoro, Sylvester A. O., Jesse R. Fann, Ernest A. Weymuller, Isaac M. Macharia, and Bevan Yueh. 2006. "Swahili Translation and Validation of the Patient Health Questionnaire-9 Depression Scale in the Kenyan Head and Neck Cancer Patient Population." *The International Journal of Psychiatry in Medicine*, 36(3): 367–381.

Paas, Fred G. W. C. 1992. "Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics: A Cognitive-Load Approach." *Journal of Educational Psychology*, 84(4): 429–434.

Rao, Gautam, Susan Redline, Frank Schilbach, Heather Schofield, and Mattie Toma. 2021. "Informing sleep policy through field experiments." *Science*, 374(6567): 530–533.

Romano, Joseph P., and Michael Wolf. 2005. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica*, 73(4): 1237–1282.

Spitzer, Robert L., Kurt Kroenke, Janet B.W. Williams, and Bernd Löwe. 2006. "A brief measure for assessing generalized anxiety disorder: The GAD-7." *Archives of Internal Medicine*, 166(10): 1092–1097.

Trauer, James M., Mary Y. Qian, Joseph S. Doyle, Shantha M. W. Rajaratnam, and David Cunnington. 2015. "Cognitive Behavioral Therapy for Chronic Insomnia: A Systematic Review and Meta-analysis." *Annals of Internal Medicine*, 163(3): 191–204.

van Straten, Annemieke, Tanja van der Zweerde, Annet Kleiboer, Pim Cuijpers, Charles M. Morin, and Jaap Lancee. 2018. "Cognitive and behavioral therapies in the treatment of insomnia: A meta-analysis." *Sleep Medicine Reviews*, 38: 3–16.