Effect of Light-Touch Psychological Interventions on Economic and Psychological Outcomes: Amendment to the Pre-Analysis Plan*

Anett John† and Kate Orkin‡

October 31, 2020

Abstract

This document supplements the original pre-analysis plan (available at https://www.socialscienceregistry.org/trials/2850/history/27566) submitted in March 2018. It pre-specifies the analysis of long-term follow-up data, collected between July-December 2020, between 30-36 months after the interventions were conducted.

*Supported by grant NIH UH2 NR016378. We thank Jane Dougherty, Magdalena Larreboure, Daniel Mellow, Winnie Mughogho, Moritz Poll, Clemence Pougué-Biyong, and the Busara Center for Behavioral Economics for research assistance; Johannes Haushofer, Clair Null, Michael Kremer, and the WASH Benefits Kenya team for collaboration and advice; and Nava Ashraf, David Laibson, Xavier D'Haultfoeuille, Alessandro Iaria, and Pedro Rey-Biel for comments.

†Department of Economics, University of Birmingham, UK.

‡Department of Economics, University of Oxford, UK.

1 Long-Term Effects: Data Collection

1.1 Sampling Strategy

Between July and December 2020, we conducted a long-term follow-up survey of our entire study population. We attempted to reach all 3750 women who had been recruited for the study between October 2017 and January 2018. At baseline, they were aged between 18-35, and lived in Bungoma and Kakamega counties. Notably, our sampling strategy includes both the pure control group as well as non-compliers in the three active treatment arms, neither of whom attended the light-touch psychological interventions. We also attempt to reach those who attrited from the previous endline survey (10 weeks after interventions) and the unannounced chlorine visit (12 weeks after interventions).

1.2 Survey Design

The survey was originally planned as an in-person survey, with a starting date of April 15, 2020. Given the sanitary crisis due to the Covid-19 pandemic, in-person surveying became impossible, and the research team decided to conduct the survey by phone. Phone surveys were conducted by the Busara Center for Behavioral Economic, with enumerators working from home. The survey comprised questions on key outcomes from the previous endline: chlorination of drinking water, children's diarrhea, savings behavior, and labor supply. We elicited long-term effects on potential psychological mechanisms including time preferences, self-efficacy, and salience. To better understand the observed effects on diarrhea, we chose to include new questions on water sources, perceived water safety, general hygiene practices, sanitation, and child mortality. To improve our understanding of mechanisms for the visualization intervention ("Imagining the Future"), we add a measure of the *as-if discounting* proposed by Gabaix and Laibson (2017), which hypothesizes that impatience is linked to imperfect foresight. Finally, we test for demand effects using the methodology of de Quidt et al. (2018). The questionnaire was administered in Kiswahili using SurveyCTO, which took approximately 30 minutes. Participants received KES 100 for their participation, paid via mPesa at the end of the survey.

1.3 Attrition

To minimize the delay between calls to participants in the same village, we call participants by village, with villages sequenced in a random order. We follow a strict protocol of (initially) seven call attempts to the respondent's primary contact number, at three different times of the day and different days of the week. If unsuccessful, we follow a series of specified steps, including calling second and third phone numbers where available. We contact the respective village elders to obtain additional phone numbers. If we find that overall response rates are below 80 percent, we will consider additional measures, including i) sending field officers to villages to obtain more phone numbers, and ii) following an intensive tracking approach for a random subsample (see e.g. Molina-Millan and Macours (2017)). All such measures will depend on the ongoing sanitary situation, and on approval from the respective public health authorities. Furthermore, if we find that response rates are differential across treatment groups, we correct for non-response bias using the method of Behaghel et al. (2015) (see Section 3.4.2).

2 Description of Outcomes

We use a variety of outcomes previously used in the short-term endline survey. In addition, we add the following outcomes:

2.1 Water Treatment and Water Safety

We elicit self-reported water treatment in three different ways:

- *General water treatment*: As in the baseline and endline survey, we ask participants "What is the main way you treat your water before drinking it?" In the baseline and endline surveys, the answer options were prompted: this question was included in the z-tree module, and participants could see the answer options {*boil, add chlorine, strain through a cloth, use a water filter, let it stand and settle, no treatment*} on the tablet screen. The phone survey required a different format. We decided to not prompt answers at all. Participants responded in an open format, and enumerators coded their answers into the appropriate treatment category. To assure that participants did not feel prompted or reminded to give certain answers, this was the first mention of water treatment in the survey. We additionally asked for the frequency of water treatment (always, most of the time , occasionally, never).

- *Chlorine treatment by collection*: Participants were asked how often they or another household member collected drinking water from each of their two stated main water sources, in the last 7 days. We then ask to how many of these collected water batches they added chlorine (from a dispenser or bottle), as well as how many batches were boiled after collection. Summing over sources, we obtain the proportion of water (in units of collected batches) that was chlorinated or boiled.

- *Current drinking water*: We finally ask only about the last water that was collected, and thus the water that is currently consumed. In an open, unprompted format, participants report whether and how this last water batch is treated.

We add a "verification" of self-reports by asking participants what is written on a chlorine dispenser, as well as what is written on a WaterGuard bottle. We do this indirectly, by asking how many turns of the dispenser valve are required to treat a 20L jerrycan of water, and what the slogan of WaterGuard is. Both answers are written in large letters on the respective chlorine source. The questions are in open format without answer prompts. Verification is considered successful if at least one question is answered correctly.

To measure the perceived safety of water, we asked participants to rate the safety of each of their stated main water sources, on a scale from "very safe" to "very unsafe."

2.2 Sanitation and Hygiene

Participants were asked in which situations they wash their hands (unprompted, coded by enumerators into appropriate categories). They further reported whether soap or other cleaning products are available for handwashing, and whether they wash fruit and vegetables before eating. We also elicit which type of toilet the household has access to, whether there is a water source for handwashing near the toilet, and what is done to dispose of young children's stools.

2.3 Diarrhea and Mortality

We generate a comprehensive roster of all children in the household (both those recorded in previous survey rounds, and any new children). For each child individually, we ask about the number of independent episodes of diarrhea over the last three months (identical to the question used in the 10-week endline). To reduce problems of recall, we add a question about diarrhea in the last 7 days.

While we expect to be underpowered to see effects on mortality, we include questions on any children who died since the beginning of the study in 2017. This includes children recorded in previous survey rounds, as well as newly born children (live births only). For any deceased children, we ask about gastrointestinal illness in the period before they died.

2.4 Savings

We made some adjustments to the savings questions used in the 10-week endline survey. The previous survey asked how often they put money aside for a given savings purpose, and how much each time. The variable "Amount saved regularly (per week, KES)" was obtained by converting to weekly frequency, and summing across savings purposes. In the follow-up survey, we ask how often they put money aside in a given savings *place* (safe hiding place, bank account, ROSCA...), and how much each time. This may be more intuitive to participants who save for unspecified purposes. The same variable as before, "Amount saved regularly", is obtained by converting to weekly frequency, and summing across savings places. We additionally include questions about the current balance in each savings place.

2.5 Labor Supply

The 10-week endline survey measured labor supply in the last three months using a rate-based approach, asking about hours worked per day, days worked per week, and weeks worked in the last three months. Following the concerns of Arthi et al. (2018) about recall bias in rate-based approaches, we shorten the recall period to 7 days and simplify the question: Participants directly reported how many hours they had worked in the last 7 days, and how much they earned (cash or in-kind). Types of work include work on own farm or in own business, or on others' farm or business, but not household chores.

2.6 Time Preference and Utility Forecasting

Time and budget constraints prevented us from re-running the full incentivized elicitation of time preferences over effort from the 10-week endline survey. Instead, we use two measures from the Global Preference Survey (GPS) developed by Falk et al. (2018). In addition, we develop a new measure to capture the imperfect utility forecasting which leads to *as-if discounting* in the model of Gabaix and Laibson (2017).

- *Time preferences, quantitative:* We include the "staircase procedure" from Falk et al. (2018), which is a series of five interdependent binary choices. In each question, participants choose between receiving 300 KES today, and a larger amount in 12 months (up to KES 644, stakes from GPS for Kenya). The delayed payment increased or decreased depending on the respondent's choice in the previous question, thus allowing to "zoom in" on the point of indifference. Given potential concerns of trust and uncertainty for payments with a 12-month delay, all choices were hypothetical (as in the original GPS).

- *Time preferences, qualitative:* We also included the qualitative time preference measure from the GPS, which asks participants to rate their agreement with the statement "I am willing to give up something that is beneficial for me today in order to benefit more from that in the future" on a 7-item Likert scale. This survey item has shown the highest correlation with incentivized choice tasks in the survey validation of Falk et al. (2016).

- *Utility forecasting*: Following the model of Gabaix and Laibson (2017), agents who exhibit imperfect forecasting of future utility will behave *as if* they discount the future, even if they are perfectly patient. Agents who spend more time thinking about intertemporal trade-offs, and who are more experienced in forecasting utility, will generate less noisy expectations, and thus display less as-if discounting. We develop survey measures based on this prediction, loosely inspired by the Plymouth Sensory Imagery Questionnaire (Andrade et al. (2014)) and the Spontaneous Use of Imagery Scale (SUIS, see Reisberg et al. (2003)). Participants rate their agreement with the following items on a 7-point Likert scale:

  - [*Vividness of mental imagery*] "Please try to form an image in your mind, of yourself and your family in the future. Please imagine where you will be, and what you will look like, in one year's time. I will give you a moment to imagine this. [*Pause.*] Please now rate your mental image on the following scale: 1 (no image at all) to 7 (image as clear and vivid as real life)."
  - [*Spontaneous use of imagery - important*] "When I have to make an important decision, I try to paint a clear picture of the consequences of that decision."
  - [*Spontaneous use of imagery - small*] "Even in the small everyday decisions that I make in my daily life, I try to paint a clear picture of the consequences of those decisions."

2.7 Self-Efficacy

As in previous surveys, we measure self-efficacy using the Generalized Self-Efficacy (GSE) scale (Schwarzer and Jerusalem, 2010). Given time constraints, we drop the two additional reversed items which we had included in the endline survey, and revert to the original scale of 10 items used by (Schwarzer and Jerusalem, 2010).

2.8 Salience

To test whether our treatments increase the salience of chlorination in the long term, we include the salience task used in the 10-week endline survey: Participants listen to lists of nine words read out by the enumerator. Afterwards, they repeat back any words they remember. Each list contains one word related to water chlorination, one related to savings, and one related to farm investment, along with six "filler" words. The outcome of interest is whether participants remember the chlorine [savings] word, *conditional* on the total number of words remembered. The position of the chlorine, savings, and farm words varied across the lists, thus allowing us to control for order effects. The word lists are shown in Table 2 of the original PAP. We made the following adjustments to adapt to the shorter phone survey format:[1] Instead of reading out all three lists to each participant, each participant was read one randomly selected list. We selected 100 participants who were paid KES 15 for each word remembered. The incentives were announced at the start of the task, and winners were informed and paid at the end of the survey.

2.9 Experimenter Demand Effects

We consider the possibility that our obtained measures of chlorination are not participants' *natural* choice, but are biased by unobservable experimenter demand effects. In the terminology of de Quidt et al. (2018), instead of the natural action $a(\zeta)$ that a participant would take in environment ζ, we may measure the action $a^L(\zeta)$ (where L stands for latent demand effects). At the end of the survey, we repeat the elicitation of *Water treatment by collection* (Section 2.1), but include random "demand treatments", which make the experimenter's hypotheses explicit. Specifically, participants are randomized to group A [group B], and informed: "We hypothesize that people who participated in this study and received the same treatment as you will give higher [lower] responses to these questions than others." We then repeat the

[1] This task is conceptually similar to the word search task in Lichand and Mani (2020), which was also conducted over the phone.

elicitation of *Water treatment by collection.* We obtain $a^+(\zeta)$ and $a^-(\zeta)$, the actions under explicit experimenter demand. Assuming that any latent demand effects in $a^L(\zeta)$ are less extreme than our demand manipulation, $a^+(\zeta)$ and $a^-(\zeta)$ can be used as bounds on demand-free behavior, and used to calculate "demand-robust" confidence intervals for our estimated treatment effects (demand treatment wording and methodology fromde Quidt et al. (2018)).

2.10 Behavior before Covid-19 Crisis

The ongoing Covid-19 sanitary crisis may have impacted respondents' health and economic behavior. Since our randomization was stratified on village of residence, exposure to the crisis should be orthogonal to assignment to treatment. However, we were concerned that heterogeneity in response to the crisis would increase the variance of our outcomes of interest, and make us underpowered to detect responses to our treatments. We thus elicit pre-Covid behavior for two key outcomes: *General water treatment* (Section 2.1) and *Amount saved regularly* (Section 2.4). Instead of referring directly to Covid-19 (which may prime anxiety in responses) or referring to absolute dates (which are not habitually used by all respondents), we refer to the time before the nationwide curfew. The introduction of a nationwide 7pm-curfew on 25 March 2020 marks a salient start to the Covid-19 crisis in Kenya, and is easily remembered by all Kenyans. We elicit pre-Covid behavior by asking respondents after the respective survey section, "Did you treat your water [save] in the same way before the curfew started, and as often as now?" If they do not reply in the affirmative, we repeat the respective survey section, but refer to the time before the curfew. The resulting measures *General water treatment (pre-Covid)* and *Amount saved regularly (pre-Covid)* can be used to replace the post-Covid measures in robustness checks.

3 Econometric Approach

3.1 Main Specification

Our estimation of treatment effects follows the econometric specification in the original PAP. In particular, we continue to run two main specifications:

1. A comparison of the three active treatment groups, which compares the two psychological treatments to the active control group (see Section 2.1 in original PAP). Regressions control for the baseline value y_{i0} of the respective outcome variable where available. The sample is restricted to those who are in these

three active treatment groups and who participated at least in the baseline survey and the first intervention session ("compliers").

2. A comparison of all four treatment groups, which uses the pure control group as a reference category (see Section 2.2 in original PAP). All participants who are reached in the follow-up survey are included in this specification. Since baseline outcomes are not observed for the pure control group and for non-compliers in the active treatment groups (beyond variables collected in the recruitment census), this specification does not control for baseline values y_{i0}.

In addition, we run a subgroup analysis based on treatment assignment in the WASH Benefits study, again as described in the original PAP (Section 2.2). We investigate whether the effect of our psychological interventions on key health outcomes differs by whether or not the village was randomly assigned to receive a chlorine dispenser.

3.2 List of Outcome Variables

As in the original PAP, we divide outcomes into behaviors, psychological mechanisms, and tests for alternative mechanisms. The former two are enumerated below while the latter are described in Section 3.4.3. Within the behavior and psychological mechanism groups, we list primary, secondary and exploratory variables of interest. We apply the multiple hypothesis testing described in section 3.3. We adjust for multiple hypothesis testing within outcome groups (behaviors and psychological mechanisms) and hierarchical categories (primary, secondary and exploratory), but not across hierarchical categories or across outcome groups. Variables marked with * are available in the 10-week endline survey, the rest are measured only in the long-term follow-up survey. Heterogeneity analysis is described in Section 3.4.1.

1. Economic and Health Behaviors
   (a) Primary:
      i. General water treatment: Indicator for chlorine *
      ii. Number of diarrhea episodes per child under 15 in last 3 months *
   (b) Secondary:
      i. Amount saved regularly (frequency converted to weekly, KES) *
      ii. Total hours of work in last 7 days

(c) Exploratory

i. Health

- Chlorine treatment by collection: Proportion of collected water batches that were chlorinated, last 7 days
- Current drinking water: Indicator for adding chlorine
- Verified chlorination: Interaction of self-reported water treatment with verification questions
- Number of diarrhea episodes per child under 5 in last 3 months *
- Proportion of children taken for healthcare check-up in last 3 months *
- Handwashing and hygiene index: Index of
  - Number of situations in which they report washing hands
  - Indicator: Always washes fruit and vegetables
  - Indicator: Water source for handwashing within 2m of toilet
- Mortality: Any children died since 2017 (indicator) *

ii. Savings

- Indicator: Amount saved regularly is positive *
- Number of ROSCAs *
- Indicator: Respondent saves for productive investments *
- Savings balance (KES)

iii. Labor supply

- Total earnings from work, last 7 days (cash and in-kind)

2. Psychological Mechanisms

(a) Primary:

i. Generalized Self-Efficacy (GSE) scale, z-scored*

ii. Utility forecasting: Vividness rating

(b) Secondary:

i. Utility forecasting: Experience index

- Index of: *Spontaneous use - important decisions* and *Spontaneous use - small decisions*

ii. Time Preferences, quantitative (staircase method): Discount factor

iii. Time Preferences, qualitative: Rating

3.3 Multiple Hypothesis Testing (MHT) Correction

We use a stepdown procedure to adjust p-values for the false discovery rate (FDR) among a group of outcomes, and will report the resulting "q-values" Anderson (2008). Indices are constructed following Anderson (2008). As discussed above, we adjust for multiple hypothesis testing within outcome groups (behaviors and psychological mechanisms) and hierarchical categories (primary, secondary, exploratory), but not across. We consider the effects of our two active interventions to be theoretically distinct and therefore do not correct across them.

3.4 Other Analyses

3.4.1 Heterogeneous treatment effects by water safety

We test whether treatment effects on water chlorination are heterogeneous by perceived water safety. In other words, we hypothesize that individuals who perceive their main drinking water source to be safe, are less likely to respond to our treatments than those who believe their water source to be unsafe. Since perceived water safety is only observed during the follow-up survey, we need to account for potential endogeneity to treatment status. We address this using two methods:

1. We estimate predicted water safety following Abadie et al. (2018). We regress water safety reported during the follow-up on observed baseline characteristics in the active control group. Available baseline characteristics include choice of water source, self-reported water treatment, and a range of demographics. We then extrapolate these predictions to the treatment groups, and use predicted water safety as a heterogeneity variable in estimating treatment effects on chlorination. We avoid overfitting of predicted values in the active control group using leave-one-out estimation (Abadie et al., 2018). Because the pure control group did not participate in the baseline survey, this method can only be used for comparisons between the three active treatment arms.

2. An alternative method is to assign an average water safety rating to each water source by village. The follow-up survey elicits 13 different types of water sources. We assume that participants in the same village who report the same water source type are a) referring to the same water source, or b) referring to water sources with comparable water safety. Under this assumption, we can assign to each water source type the average safety rating given by control group participants in that village. Since choice of water source is itself potentially endogenous to treatment, we test whether treatment assignment predicts the

average safety rating of the chosen water source. In other words, we test whether treated participants move to water sources which are rated as safer by the control group (such as bottled water, which is safer but more expensive). If we find that water source safety is not significantly related to treatment status (for instance, because limited sources are available within each village), we use the obtained average ratings as a heterogeneity variable in estimating effects on chlorination.

3.4.2 Selective attrition

We test for differential attrition by treatment group using the approach specified in the original PAP (Section 2.5, equations 3 and 4). If we find that attrition differs significantly across groups, we will provide robustness checks that control for sample selection bias resulting from non-response, using the methodology of Behaghel et al. (2015). Using this method, only individuals who are reached after a specified level of effort (e.g., number of phone calls, additional inquiries to obtain updated phone numbers) are included in the treatment groups with the higher response rates. Hardest-to-reach respondents are excluded to equalize response rates in all groups.

3.4.3 Alternative mechanisms

We examine whether our treatments influenced chlorination behavior through mechanisms other than time preferences or self-efficacy, using the following measures:

1. **Salience:** We test for the possibility that our treatments differentially increased salience of water chlorination in the long term, using the methodology outlined in the original PAP (Section 2.5, equations 6 and 7). In contrast to the endline survey, each participant now recites one randomly selected word list, rather than all three. This affects the number of observations, but not the econometric specification. As is the original design, word order effects are controlled for by using three different lists, varying the word order, and including list fixed effects.

2. **Experimenter Demand Effects**: We test for experimenter demand effects using the randomized demand treatments described in Section 2.9.

   (a) We first test the basic hypothesis of no demand effects, $a^-(\zeta) = a^+(\zeta)$. Assuming monotonicity of latent demand effects, $a^+(\zeta) \geq a^L(\zeta) \geq a^-(\zeta)$, as well as bounding, $a^+(\zeta) \geq a(\zeta) \geq a^-(\zeta)$ (assumptions 1 and 2 in de Quidt et al. (2018)), equality of the bounds $a^+(\zeta)$ and $a^-(\zeta)$ would imply that the observed $a^L(\zeta)$ in our survey is equal to the natural action $a(\zeta)$.

(b) In case we find significant demand effects $a^-(\zeta) \neq a^+(\zeta)$, we will provide "demand-robust" confidence intervals for our estimated treatment effects, using the methodology in de Quidt et al. (2018).

References

Abadie, Alberto, Matthew M. Chingos, and Martin R. West, "Endogenous Stratification in Randomized Experiments," *The Review of Economics and Statistics*, 2018, *100* (4), 567–580.

Anderson, Michael L, "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*, 2008, *103* (484), 1481–1495.

Andrade, Jackie, Jon May, Catherine Deeprose, Sarah-Jane Baugh, and Giorgio Ganis, "Assessing Vividness of Mental Imagery: the Plymouth Sensory Imagery Questionnaire," *British Journal of Psychology*, 2014, *105* (4), 547–563.

Arthi, V, K Beegle, J De Weerdt, and A Palacios-Lopez, "Not your average job: Measuring farm labor in Tanzania," *Journal of Development Economics*, 2018, *130*, 160–172.

Behaghel, Luc, Bruno Crepon, Marc Gurgand, and Thomas Le Barbanchon, "Please Call Again: Correcting Nonresponse Bias in Treatment Effect Models," *The Review of Economics and Statistics*, 2015, *97* (5), 1070–1080.

de Quidt, Jonathan, Johannes Haushofer, and Christopher Roth, "Measuring and Bounding Experimenter Demand," *American Economic Review*, 2018, *108* (11), 3266–3302.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde, "Global evidence on economic preferences," *The Quarterly Journal of Economics*, 2018, *133* (4), 1645–1692.

_ , _ , Thomas J Dohmen, David Huffman, and Uwe Sunde, "The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences," *IZA Discussion Paper*, 2016, *9674*.

Gabaix, Xavier and David Laibson, "Myopia and Discounting," *NBER Working Paper No. 23254*, 2017.

Lichand, Guilherme and Anandi Mani, "Cognitive Droughts," *University of Zurich Working Paper*, 2020, *341.*

Molina-Millan, Teresa and Karen Macours, "Attrition in Randomized Control Trials: Using tracking information to correct bias," *IZA Discussion Paper*, 2017, *10711.*

Reisberg, Daniel, David G Pearson, and Stephen M Kosslyn, "Intuitions and introspections about imagery: The role of imagery experience in shaping an investigator's theoretical views," *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 2003, *17* (2), 147–160.

Schwarzer, Ralf and Matthias Jerusalem, "The General Self-Efficacy Scale (GSE)," *Anxiety, Stress, and Coping*, 2010, *12*, 329–345.