Pre-Analysis Plan: Temperature and Economic Choices

Ingvild Almås, Maximilian Auffhammer, Tessa Bold, Ian Bolliger,
Aluma Dembo, Solomon Hsiang, Shuhei Kitamura, Edward Miguel, and Robert Pickmans

September 12, 2017

List of Tables

List of Figures

Abstract

This document describes the design and analysis plan for an experiment to be held during Fall 2017 (and possibly Spring 2018) at both the Busara Center for Behavioral Economics in Nairobi, Kenya, and the Experimental Social Science Laboratory (Xlab) at the University of California, Berkeley. In this lab experiment, we aim to study the causal effect of temperature on individual decision-making. Through a series of lab modules, we will study the causal effect of

temperature on standard economic choices and cognitive performance. In particular, we will be testing the effect of temperature variation on productivity, cognitive ability, pro-social behavior, trust, trustworthiness, cooperation, destruction, time preferences, and risk preferences.

1 Introduction

There is an extensive literature documenting a relationship between high temperature, conflict and poor economic performance, yet to date, little is known about the causal effect of temperature on individual decision-making (Burke et al., 2009; Hsiang et al., 2013). This question, important in and of itself, is made even more salient given the predicted rise in global temperature and variation in weather patterns. Related is a nascent literature addressing how environmental factors and neurobiology influence economic choices, which includes work on the psychology of poverty, hunger and stress (Mani et al., 2013; Bushman et al., 2014; Haushofer and Fehr, 2014). In this research project, we aim to study the causal effect of temperature on individual decision-making. To this end, we plan to conduct a sequence of modules at both the Busara Center for Behavioral Economics in Nairobi, Kenya, and the Experimental Social Science Laboratory (Xlab) at the University of California, Berkeley.

2 Research strategy

The experiments and tests will be designed to study the causal effect of temperature[1] on standard economic choices and cognitive performance. In particular, we will employ versions of standard laboratory experiments to test the effect of temperature variation on productivity, cognitive ability, pro-social behavior, trust, trustworthiness, cooperation, destruction, time preferences, and risk preferences. Lastly we will study beliefs about others responses and beliefs about others behavior in a few modules involving strategic behavior.

The participants will be recruited through the Busara Center in Nairobi, Kenya, and the Xlab at the University of California, Berkeley. All experiments will be conducted in these two locations.

2.1 Sampling

2.1.1 Sampling Frame

We will conduct the experiment in both Nairobi, Kenya at the Busara Center, as well as in Berkeley, California at the Xlab. The target population for the study in Nairobi is university students enrolled at the University of Nairobi. Meanwhile, the target population for the study in Berkeley is university students enrolled at the University of California, Berkeley. In both places, the sample will be recruited by the Busara Center and the Xlab in their respective locations, and recruitment will be aided through the use of advertisements.

2.1.2 Statistical Power

In order to determine the relationship between the minimum detectable effect (MDE) and sample size, we have developed a series of "MDE-sample size graphs" for the Berkeley sample, the Kenya sample, and the combined sample.[2] Participants are randomly sampled and are allocated across

[1] In particular, we will be examining environmental temperature, and are not referring to an individual's internal body temperature.

[2] In calculating the MDE, we have maintained several parameters constant. Namely, $\alpha = 0.05$, power $(1 - \beta) = 0.80$, and false discovery rate = 0.05. The effect was calculated for a two-sample t test, assuming equal sample sizes across treatment arms as well as equal variances across treatment arms. Standard errors are clustered using early pilot data run for an earlier version of the experiment held in November 2016 and January 2017 at the Busara Center.

different temperature treatments in equal proportions. In order to adjust for multiple hypothesis testing, we plan to correct for the false discovery rate in our analysis. For the current power calculations, we assume that the proportion of true null hypotheses (pi_0) is either 0.50 or 0.80.[3] We trace out lines assuming FDR adjustments, and bound them with lines showing the relationship between the MDE and sample size in the case where: 1) p-values are unadjusted for multiple inference; and 2) we adjust p-values using Bonferroni on 10 hypotheses (where 10 comes from the number of hypotheses on outcomes of central interest that would be included in our SUR framework for multiple hypothesis testing adjustments).[4] Complementing these graphs are "cost graphs" for the Berkeley sample and the Kenya sample, which display the expected cost given the sample size. To address possible participant no-shows, we plan to overrecruit at the Xlab and the Busara Center by 50 percent.

Figure 1 and Figure 2 display the MDE-sample size graph and cost graph,[5] respectively, for the Berkeley sample. If 800 students are recruited at the Xlab in a context of 50% overrecruitment and a target of $25 per hour on average for participants, then the MDE should be 0.223 if $pi_0 = 0.50$ and close to 0.264 if $pi_0 = 0.80$. Such recruitment is projected to cost $32,000 if the length of the experiment is 1.5 hours.

Figure 3 and Figure 4 display the MDE-sample size graph[6] and cost graph,[7] respectively, for the Kenya sample. Similar to above, if 800 students are recruited at the Busara Center in a context of 50% overrecruitment and a target of 900 KSh per hour on average for participants, then the MDE should be close to 0.223 if $pi_0 = 0.50$ and close to 0.264 if $pi_0 = 0.80$. Such recruitment is projected to cost $13,143.96 if the length of the experiment is 1.5 hours.

Finally, Figure 5 displays the MDE-sample size graph for the combined sample. The larger sample size leads to improvements in the MDE: if 1600 participants are recruited (800 from each site), then the MDE should be close to 0.158 if $pi_0 = 0.50$ and close to 0.188 if $pi_0 = 0.80$. The cost for the experiment in both sites considered together is projected to be $45,143.96 if the length of the experiment is 1.5 hours.

2.1.3 Assignment to Treatment

In the Busara Center, the administrators will use two different colored sets of similar sized laminated place cards (e.g red and green) numbered 1-6. All the cards are put into an envelope and shuffled. A card is randomly pulled out and handed to each participant. The color of the place card one receives determines the room that one goes to and the place card number determines the space one sits at in the lab room. In the Xlab, a random number generator will output a number for each participant that assigns each participant to a workstation in one of the two treatment groups. This assignment will serve as the basis for the treatment indicator (assigned to either the 22 °C group or the 30 °C group).

2.1.4 Attrition from the Sample

We do not expect people to attrit from the experiment once they have entered. The temperatures of the rooms to which people will be exposed to will be either 22 °C or 30 °C, which are well within the typical temperature distribution experienced during the summer. People will of course

[3]See 3.1.3 for details on multiple hypothesis testing adjustments.

[4]As expected, the lines accounting for FDR adjustments are always between the unadjusted and Bonferroni lines.

[5]At the Xlab, participants must be guaranteed on average a payment of $15 per hour, whether it come from fixed payments or game payouts. This assumption is maintained in this cost graph. At this stage we have not yet timed the experiment, so we have included graphs for versions where the experiment is 1 hour and where it is 1.5 hours.

[6]Because we do not expect any difference in the parameters between the sites, Figure 3 is the same as Figure 1.

[7]There are mandatory partcipation fees for university students (400 Kenyan Shillings) that are assumed to be paid to show-ups who are turned away because of overrecruitment. We have also assumed that 50% of participants (and overrecruited) show up early, earning a 50 Kenyan Shilling show-up fee.

be allowed to leave the experiment early, although this would entail forgoing payments earned from completing the experiment.

In order to prevent sample attrition, we will encourage participants to drink water and to use the restroom while they are in the waiting period (see 3.3 for details), which should mitigate any sort of discomfort that might be experienced as a result of the treatment. We will also build in a detailed debriefing at the pilot stage to ascertain the degree to which attrition is a problem, and possibly modify the experiment to address any concerns.

At this stage we do not expect attrition to be significant and so have not factored it into our MDE-sample size calculations.

2.2 Fieldwork

2.2.1 Instruments

Both the Xlab and the Busara Center will use:

- Computers or tablets loaded with modules
- Space heaters: to raise room temperature as needed
- Air conditioning systems: to cool room temperature as needed
- HOBO Temperature/Relative Humidity Data Logger: to measure temperature. We will also use this to measure relative humidity for robustness (see 3.1.5 for details)
- HOBO MX2302 External Temperature/RH Sensor Data Logger: to record outside temperature and relative humidity for robustness (see 3.1.5 for details)
- Braun Thermoscan 5 In-Ear Thermometer: for research assistants who are administering the test to record their own core temperatures as a proxy for core temperatures in others

2.2.2 Data Collection

Prior to the experiment, we will run a three-day pilot in which we will make sure that the modules are running smoothly and possibly adjust power calculations. We will run the pilot as we will the main experiment, including the use of recruited student participants. There will also be a survey to ascertain the degree to which participants knew what the experiment's true goal was.[8]

Adjustments may have to be made to modules, or the experiment set up generally, to account for any unforeseen outcomes. Any sort of adjustments will be noted afterwards. None of the data from the pilot will be used in the main analysis.

We expect the experiment (and hence, the entire data collection process) to take several months. Data from the experiment will be sent directly from participants' computers to the research assistants' computers, and then downloaded via .csv output from the oTree platform, through which the modules are run. The data will be kept anonymous (data separated from names) and hosted on a Box folder shared among the research team.

2.2.3 Data Processing

Data processing entails 1) cleaning the data, 2) managing the data, and 3) analyzing the data. We anticipate data processing to occur as results come in. The data will be anonymous and work involving the data will be kept either in a Box folder shared among the research team or brought locally onto private computers used by research team members.

[8] This will be administered at the end of a pilot session for half of the pilots, and midway for the other half.

2.3 Debriefing Questions

To gauge the extent to which participants might identify the treatment, during the pilot we will employ a battery of questions about the lab environment, of which only some are meant to ascertain the previously mentioned possibility. These questions will be useful if we have to revisit any of the parameters for the experiment. Questions will also have space for written comments. The debriefing questions are to be structured as follows:

1. How would you describe the comfort of your chair?
   (a) Comfortable
   (b) Uncomfortable
2. How would you describe the screen brightness?
   (a) Comfortable
   (b) Too bright
   (c) Too dim
3. How would you describe the air in the room?
   (a) Comfortable
   (b) Too dry
   (c) Too humid
4. How would you describe the temperature in the room?
   (a) Comfortable
   (b) Too cold
   (c) Too hot
5. How would you describe the lighting in the room?
   (a) Comfortable
   (b) Too bright
   (c) Too dim
6. How would you describe the space in the room?
   (a) Comfortable
   (b) Too empty
   (c) Too crowded
7. How would you describe the ambient noise in the room?
   (a) Comfortable
   (b) Too silent
   (c) Too noisy

8. What do you think this experiment was measuring (Please check any that apply)?
   [ ] Willingness to compete
   [ ] Screen brightness
   [ ] Willingness to collaborate
   [ ] Room temperature
   [ ] Ethnicity
   [ ] Room lighting
   [ ] Ambient noise
   [ ] Differences in generosity
   [ ] Room space
   [ ] None of the above
   [ ] Other

2.4 Future Research

For future work, we will also attempt to identify mechanisms by which temperature affects behavior. The literature on the psychology of poverty has found that stress and mental fatigue are key mechanisms for how poverty influences decision-making (Mani et al., 2013; Haushofer and Fehr, 2014).[9] We hypothesize that similar mechanisms explain how temperature influences behavior and will therefore test whether temperature causes stress (measured by saliva samples and other biomarkers). The set-up of the experiment will also allow us to make suggestive statements about whether temperature affects behavior via its effect on productivity and mental acuity.

3 Empirical Strategy

3.1 Empirical specification

3.1.1 Main specification

The object of our analysis is temperature on economic behavior, but we are also *a priori* interested in how effects may differ by gender or location. For our main specification, we will only want to control for gender or location. Thus, our main specification will regress the outcome variables of each individual module on 1) the treatment indicator (an indicator variable, with 1 for 30 °C and 0 for 22 °C); 2) an indicator for gender; 3) an indicator for country site; and 4) covariates on which treatment and control groups are unbalanced. This set of overall regressions will use the combined data from the Berkeley and Nairobi experiments. The specification is as follows:

$$Y_i^k = \beta_0^k + \beta_1^k T_i + \beta_2^k Gender + \beta_3^k Site + X_i' \Gamma + \varepsilon_i^k \quad (1)$$

where i refers to individual; s refers to site, k refers to the module outcome, Y_i is an outcome; T_i is the treatment indicator; *Gender* is an indicator for gender of the individual; *Site* is a fixed effect for experiment location; X_i is a vector of covariates that are unbalanced between treatment and control groups; and ε_i is an idiosyncratic error term.[10] We will cluster standard errors at the session level.

From the overall regressions we are primarily concerned with β_1^k; heterogeneity will be explored through other specifications (see discussion below).

[9] Related literature includes how hunger (Ashton, 2015) and sleepiness (Castillo et al., 2017) affect decision-making.

[10] Because the power calculations are based on a simple regression, the results from (2.1.2) provide a lower bound.

3.1.2 Heterogeneity

We are particularly interested in the interactions that temperature might have with gender and country ("site") in this experiment, and so will also run the following specification:

$$Y_i^k = \beta_0^k + \beta_1^k T_i + \beta_2^k Gender + \beta_3^k (T_i * Gender) + \beta_4^k Site + \beta_5^k (T_i * Site) + X_i^{'} \Gamma + \varepsilon_i^k \tag{2}$$

where i refers to individual; s refers to site, k refers to the module outcome, Y_i is an outcome; T_i is the treatment indicator; *Gender* is an indicator for gender of the individual; *Site* is a fixed effect for experiment location; X_i is a vector of covariates that are unbalanced between treatment and control groups; and ε_i is an idiosyncratic error term. From (2) we will pay particular interest to β_3^k and β_5^k. As above, we will cluster standard errors at the session level.

There may be some contexts where the triple interaction coming from temperature, gender, and country could be of interest, and so we will run the following specification:

$$\begin{aligned} Y_i^k = {} & \beta_0^k + \beta_1^k T_i + \beta_2^k Gender + \beta_3^k (T_i * Gender) + \beta_4^k Site + \\ & \beta_5^k (T_i * Site) + \beta_6^k (Gender * Site) + \beta_7^k (T_i * Gender * Site) + X_i^{'} \Gamma + \varepsilon_i^k \end{aligned} \tag{3}$$

where i refers to individual; s refers to site, k refers to the module outcome, Y_i is an outcome; T_i is the treatment indicator; *Gender* is an indicator for gender of the individual; *Site* is a fixed effect for experiment location; X_i is a vector of covariates that are unbalanced between treatment and control groups; and ε_i is an idiosyncratic error term. However, we note that β_7^k is not of central interest; we may be underpowered to detect this effect. Any findings will be reported in an appendix. As above, we will cluster standard errors at the session level.

3.1.3 Controlling for Multiple Inference

For our primary analysis, we plan to run multiple hypothesis testing adjustments to protect against the false discovery rate (FDR) on two different families of p-values, which will serve as the core of our analysis. To give the correct inference for β_1^k, the first family that adjustments will be performed on is the single set of p-values associated with β_1^k coming from (1) across all k outcomes. We calculate sharpened q-values over our set following Benjamini et al. (2006) to control the FDR. Rather than specifying a single q, we will report the minimum q-value at which each hypothesis is rejected, following Anderson (2008). Thus, in our analysis we will report both typical p-values and minimum q-values.

The second family to include adjustments comprises the sets of p-values associated with β_3^k and β_5^k coming from (2) for each k outcome. We perform adjustments in our analysis this way in order to understand heterogeneity for a given module outcome. We will calculate sharpened q-values within each set following Benjamini et al. (2006) to control the FDR and report the minimum q-value at which each hypothesis is rejected, following Anderson (2008). Similarly to above, we will report both typical p-values and minimum q-values.

For exploratory analysis, we will complement these adjustments with F-tests of joint significance. The first family described above will be complemented with an overall test using a Seemingly Unrelated Regression (SUR) approach, which consists of performing an F-test of on overall significance of β_1, stacking across all k from (1). From this result, we will be able to better deduce an overall temperature effect on economic behavior. Meanwhile, exploration of heterogenous effects will be complemented with a SUR approach as well, where we test the two joint null hypotheses on β_3^k and β_5^k. From these results, we will be able to better deduce overall temperature interaction effects on economic behavior. However, we emphasize that the FDR-adjustments will be of primary consideration.

3.1.4 Exact tests of the treatment effect

In addition to the approach laid out above, we plan to perform Monte Carlo approximations of exact tests of the treatment effect as a robustness exercise (Fisher, 1935). With randomization inference, we will be able to test the Fisherian sharp null hypothesis that $Y_i^{(t)} = Y_i^{(c)}$.[11]

We will calculate exact p-values for the treatment effects under the null hypothesis using a Fisher permutation test. To be more precise, we will take 10,000 permutations of the treatment indicator T_i and calculate the t-statistic for each m^{th} permutation. We hold fixed the treatment-control balance by reallocating realized treatment assignments.

We will adjust for multiple inference by also providing minimum q-values to control the FDR, using the p-values derived from the exact tests. Furthermore, we will use an omnibus test for overall experimental significance implemented by Young (2016).[12]

3.1.5 Additional data

Even though the primary motivation for this paper is to understand the effect of temperature on economic choices, there may be day-to-day idiosyncratic factors that influence temperature's effect on economic choices. Our reliance on an indicator for temperature difference across treatment arms could be complemented with investigation on actual temperature variation or temperature changed measured directly by the human body. To this end, we will also collect information that will allow us, in a fine-grain way, to explore the effect of temperature (and possibly weather, more broadly speaking) on behavior. These variables include:

- Temperature outside of the laboratories
- Relative humidity inside of the laboratories
- Relative humidity outside of the laboratories
- Actual room temperature
- Core temperature experienced by the experimenters, as measured by thermometer (data to be collected only during the pilot)

3.2 Balancing Checks

3.2.1 Balance between treatment and control

We will perform Student's t-test to check balance between treatment and control groups. The variables that will be checked for balance across treatment arms include: 1) the participant's self-reported weight; 2) the participant's self-reported height; 3) age; 4) gender; 5) in-state residency status (Xlab) or ethnicity (Busara Center); 6) father's occupation status; 7) mother's occupation status; 8) combined parental income; 9) father's highest level of education; and 10) mother's highest level of education. Any variable that is found to be unbalanced across treatment group at the $\alpha = 0.05$ level will be included in the regression at the level where the unbalance was found. These questions will be included in the demographic survey module.

[11] This hypothesis is more restructive than a null hypothesis of no average treatment effect.

[12] Specifically, the ombnibus test calculates the p-value by calculating the Wald statistic for each realization of the treatment assignment (i.e. over 10,000 permutations) and then calculates the p-value of joint significance by looking at how many times the Wald statistic was larger than the original Wald statistic.

3.2.2 Balance between attritors and non attritors

To check for balance between attritors and non attritors, we will investigate if there is a statistically significant difference in the rates of attrition between treatment arms at the $\alpha = 0.05$ level. In the case of such a difference between attritors and non attritors, then Lee (2009) bounds will be implemented. We will perform this check by site as well as overall.

3.3 Before the modules

Prior to beginning the modules in the testing rooms, participants will enter a "waiting period" of twenty minutes in order to guarantee sufficient exposure to treatment prior to beginning the first module. During this waiting period, participants will fill out any necessary paperwork and be told the guidelines of the experiment. Participants will also have time to use the restroom or grab a drink of water.

4 Modules

Below is a detailed description of the modules and discussion of how the analysis will be carried out. The modules are listed in the order in which they will occur in the experiment. Table 1 lays out outcomes of primary interest and exploratory interest (the latter specificaly noted) for each module, as well as the definitions of these outcomes.

Table 1: Modules and outcomes of interest

Modules	**Outcomes**	**Definitions**
1) Production phase	a) *prodlevel*	a) total points earned (absolute, not normalized)
2) Real effort dictator game	a) *dictatorshare*	a) share of the total that is allocated to the other participant
3) Risk preferences	a) *risk*	a) categorical variable indicating coin choice from A
	b) *transitivity*	b) indicator of transitivity violation using both A & B
	c) *FOSDviolation* (exploratory)	c) indicator of choice of coin 7 in A
4) Time preferences	a) *beta* b) *delta*	a) aggregate estimate of beta b) aggregate estimate of delta
5) Trust game	a) *sharesent* b) *sharesentback* (exploratory)	a) share of the total amount sent b) share sent back
6) Public Goods game	a) *cooperation* b) *beliefs* (exploratory)	a) the amount put into the fund b) indicator variable that marks as 1 if the individual guesses correctly about another's contribution, and 0 otherwise
7) Cognitive Ability - Raven's	a) *puzzles*	a) percentage of puzzles chosen correctly
8) Joy of Destruction	a) *destroyed*	a) percentage of gift cards destroyed
9) Cognitive reflection	a) *sharecorrect*	a) share of six questions answered correctly
	b) *answerintuitive* (exploratory)	b) question answered intuitively

Table 1 – continued from previous page

Modules	Outcomes	Definitions
	c) *timespentoverall* (exploratory)	c) time spent on all questions
	d) *timespent* (exploratory)	d) time spent on a question
11) Demographic survey questions	n/a	n/a
12) Charity	a) *donation*	a) absolute amount chosen to donate

4.1 Production phase

The production phase serves two purposes. First, it enables us to measure the effect that temperature has on productivity. Second, it provides the necessary work effort to create real effort stakes in the dictator game.

4.1.1 Specifications

In the production phase, participants are engaged with a slider task, where they place a slider on an assigned number from 1 - 100 using the touchscreen (or mouse). Participants receive a point if the number is correct, 0 otherwise. Final earnings from production phase are either "high" (weakly above median) or "low" (below median) productivity. The median is calculated within treatment cohort. Thus there are three pairs: one high, one low, and the median pair (the last of which is randomly assigned to either high or low).

The primary outcome of interest is total points earned (absolute, not normalized):

$$Y_i = prodlevel_i$$

Production phase: main specification

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 Site + X_i^{'}\Gamma + \varepsilon_i \tag{4}$$

Production phase: heterogeneity

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 (T_i * Gender) + \beta_4 Site + \beta_5 (T_i * Site) + X_i^{'}\Gamma + \varepsilon_i \tag{5}$$

4.1.2 Hypotheses

From (4), we test the null hypothesis that:

- Temperature does not affect productivity ($H_{PP1} : \beta_1 = 0$)

From (5), we test the null hypothesis that:

- Temperature does not affect productivity differentially by gender ($H_{PP2} : \beta_3 = 0$)
- Temperature does not affect productivity differentially by site ($H_{PP3} : \beta_5 = 0$)

4.1.3 Multiple Inference Corrections

Recall we will only perform multiple hypothesis adjustments for primary outcomes, grouping hypotheses in sets and perform adjustments along those sets. In terms of FDR adjustments, H_{PP1} will belong to the set of main hypotheses that will have adjustments on that set, and adjustments will be performed on the set of hypotheses H_{PP2} and H_{PP3}. These results will be part of the core of the analysis.

For more exploratory analysis, recall we will also perform tests of overall significance on primary outcomes. Thus, H_{PP1} will belong to a set of overall hypotheses that will be tested via SUR. The hypotheses on β_3 and β_5 coming from (5) for primary outcomes (H_{PP2} and H_{PP3}) will be featured in sets of hypotheses of β_3^k and β_5^k, where the joint null hypotheses will also be tested via SUR. Exploratory analysis will also be served by randomization inference and products leading from such, to be tested in a similar way described above.

4.2 Real effort dictator game

The purpose of this module is to study whether temperature affects pro-social behavior.

4.2.1 Specifications

We will use the production phase as a determinant of earnings in order to establish clear entitlements. We will match participants with equal productivity (either high or low) and give them the information about what each of them have earned in the production phase (i.e., we give them a clear suggestion that the fair outcome is an equal split). Participants are matched in pairs and asked how much of the joint earnings they want to transfer to the other participant. All participants will act as dictators and they will know that there is a fifty percent chance that their decision will be implemented. (See Cherry et al. 2005 for reference).

The primary outcome of interest is the share of the total that is allocated to the other participant.

$$Y_i = dictatorshare_i$$

Real effort dictator game: main specification

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 Site + X_i^{'}\Gamma + \varepsilon_i \tag{6}$$

Real effort dictator game: heterogeneity

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 (T_i * Gender) + \beta_4 Site + \beta_5 (T_i * Site) + X_i^{'}\Gamma + \varepsilon_i \tag{7}$$

In order to investigate whether stakes may matter for behavior in this module, we will, as a robustness check, include variants of the above specifications that include an indicator for being part of the "high" or "low" group ($highlow_i$) (see 4.1.1). In this specification we will also include a variable for production level ($prodlevel_i$), as any effect from being in the "high" or "low" group may be due to productivity and not stakes per se. If this robustness check leads to far more precise estimates, then we may use estimates from (8) and (9) rather than (6) and (7), respectively, for our main results.

Real effort dictator game: main specification (robustness check)

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 Site + \beta_4 highlow_i + \beta_5 prodlevel_i + X_i^{'}\Gamma + \varepsilon_i \tag{8}$$

Real effort dictator game: heterogeneity (robustness check)

$$\begin{aligned} Y_i = {} & \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 (T_i * Gender) + \beta_4 Site + \beta_5 (T_i * Site) + \\ & \beta_6 highlow_i + \beta_7 prodlevel_i + X_i^{'}\Gamma + \varepsilon_i \end{aligned} \tag{9}$$

4.2.2 Hypotheses

From (6) or (8), we test the null hypothesis that:

- Temperature affects share allocated ($H_{DG1} : \beta_1 = 0$)

From (7) or (9), we test the null hypothesis that:

- Temperature does not affect share allocated differentially by gender ($H_{DG2} : \beta_3 = 0$)
- Temperature does not affect share allocated differentially by site ($H_{DG3} : \beta_5 = 0$)

4.2.3 Multiple Inference Corrections

Recall we will only perform multiple hypothesis adjustments for primary outcomes, grouping hypotheses in sets and perform adjustments along those sets. In terms of FDR adjustments, H_{DG1} will belong to the set of main hypotheses that will have adjustments on that set, and adjustments will be performed on the set of hypotheses H_{DG2} and H_{DG3}. These results will be part of the core of the analysis.

For more exploratory analysis, recall we will also perform tests of overall significance on primary outcomes. Thus, H_{DG1} will belong to a set of overall hypotheses that will be tested via SUR. The hypotheses on β_3 and β_5 coming from (7) or (9) for primary outcomes (H_{DG2} and H_{DG3}) will be featured in sets of hypotheses of β_3^k and β_5^k, where the joint null hypotheses will also be tested via SUR. Exploratory analysis will also be served by randomization inference and products leading from such, to be tested in a similar way described above.

4.3 Risk preferences

The purpose of this module is to look at the effect of temperature on risk preferences, as well as to measure quality of decision-making.

4.3.1 Specifications

In this module we will elicit risk preferences using choice over lotteries with equal probability. There will be two menus to chose from, each tracing out a different budget line: Budget (A) has a slope of -2, per Eckel and Grossman (2008). Budget (B) has a slope of -1. The intercepts are 2880 tokens and 2160 tokens respectively. For both menus A and B we include the risk neutral point (H = 0), the risk averse choice (H = T), one choice below the 45-degree line (H > T), the intersection point of A and B. Of the remaining seven points 3 fall above the intersection and 3 fall below intersection. The point below the 45-degree line has the same variance as the point between the 45 degree line and below the intersection of the two lines. The choice of values is designed to be used for revealed preference analysis, not for parametric analysis. Note that the second budget line has a slope of 1 so it has the same expected payout along the entire budget.

The primary outcomes of interest are a risk measure (categorical variable indicating coin choice from A) as well as transitivity violation (indicator of transitivity violation using both A & B). An exploratory outcome of interest is FOSD violation (indicator of choice of coin 7 in A).

$$Y_i = risk_i \text{ or } transitivity_i \text{ or } FOSDviolation_i$$

Risk preferences: main specification

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 Site + X_i^{'}\Gamma + \varepsilon_i \tag{10}$$

Risk preferences: heterogeneity

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 (T_i * Gender) + \beta_4 Site + \beta_5 (T_i * Site) + X_i^{'}\Gamma + \varepsilon_i \tag{11}$$

4.3.2 Hypotheses

From (10), we test the null hypothesis that:

- Temperature does not affect risk ($H_{RK1} : \beta_1 = 0$)
- Temperature does not affect transitivity violation ($H_{TV1} : \beta_1 = 0$)
- Temperature does not affect FOSD violation ($H_{FV1} : \beta_1 = 0$)

From (11), we test the null hypothesis that:

- Temperature does not affect risk differentially by gender ($H_{RK2} : \beta_3 = 0$)
- Temperature does not affect risk differentially by site ($H_{RK3} : \beta_5 = 0$)
- Temperature does not affect transitivity violation differentially by gender ($H_{TV2} : \beta_3 = 0$)
- Temperature does not affect transitivity violation differentially by site ($H_{TV3} : \beta_5 = 0$)
- Temperature does not affect FOSD violation differentially by gender ($H_{FV2} : \beta_3 = 0$)
- Temperature does not affect FOSD violation differentially by site ($H_{FV3} : \beta_5 = 0$)

4.3.3 Multiple Inference Corrections

Recall we will only perform multiple hypothesis adjustments for primary outcomes, grouping hypotheses in sets and perform adjustments along those sets. In terms of FDR adjustments, H_{RK1} and H_{TV1} will belong to the set of main hypotheses that will have adjustments on that set, and adjustments will be performed on the set of hypotheses H_{RK2} and H_{RK3} as well as on the set of hypotheses H_{TV2} and H_{TV3}. These results will be part of the core of the analysis.

For more exploratory analysis, recall we will also perform tests of overall significance on primary outcomes. Thus, H_{RK1} and H_{TV1} will belong to a set of overall hypotheses that will be tested via SUR. The hypotheses on β_3 and β_5 coming from (11) for primary outcomes (H_{RK2} and H_{RK3}, as well as H_{TV2} and H_{TV3}) will be featured in sets of hypotheses of β_3^k and β_5^k, where the joint null hypotheses will also be tested via SUR. Exploratory analysis will also be served by randomization inference and products leading from such, to be tested in a similar way described above.

4.4 Time preferences

The purpose of this module is to (1) identify patience, and the effect of temperature on patience, and (2) identify time inconsistency and the effect of temperature on time inconsistency.

4.4.1 Specifications

We will use an established protocol for eliciting so-called "beta-delta" preferences, namely a choice over temporal budgets (CTB) design following the design in (Andreoni et al., 2015). We will do this for the overall population by treatment, and then by the set of subgroups we are focusing on, which is location and gender. Each person will be shown 2 budget lines for today vs. 3 weeks and 2 budgets for 3 vs. 7 weeks.

The primary outcomes of interest are time inconsistency (aggregate estimate of beta) and discounting (aggregate estimate of delta) for the overall population and by subgroup. To estimate these parameters, we will use Nonlinear Least Squares (NLS) estimation, as in Andreoni et al. (2015). We will use structural estimation to get aggregate estimates of both parameter estimates and standard errors, and standard errors will be used in the t-test.

$$Y = beta \text{ or } delta$$

In order to accomodate the structure of the outcome variables we utilize a simple specification.

Time preferences: main specification

$$\mu_T = \mu_C \tag{12}$$

where T refers to treatment and C refers to control

Time preference: heterogeneity

$$\mu_{T,male} - \mu_{C,male} = \mu_{T,female} - \mu_{C,female} \tag{13}$$

for heterogeneity by gender, and again T refers to treatment and C refers to control

$$\mu_{T,USA} - \mu_{C,USA} = \mu_{T,Kenya} - \mu_{C,Kenya} \tag{14}$$

for heterogenetity by site, and again T refers to treatment and C refers to control

4.4.2 Hypotheses

From (12), we test the null hypothesis that:

- Temperature does not affect beta ($H_{BE1} : \mu_T = \mu_C$)
- Temperature does not affect delta ($H_{DE1} : \mu_T = \mu_C$)

From (13), we test the null hypothesis that:

- Temperature does not affect beta differentially by gender ($H_{BE2} : \mu_{T,male} - \mu_{C,male} = \mu_{T,female} - \mu_{C,female}$)
- Temperature does not affect delta differentially by gender ($H_{DE2} : \mu_{T,male} - \mu_{C,male} = \mu_{T,female} - \mu_{C,female}$)

From (14), we test the null hypothesis that:

- Temperature does not affect beta differentially by site ($H_{BE3} : \mu_{T,USA} - \mu_{C,USA} = \mu_{T,Kenya} - \mu_{C,Kenya}$)
- Temperature does not affect delta differentially by site ($H_{DE3} : \mu_{T,USA} - \mu_{C,USA} = \mu_{T,Kenya} - \mu_{C,Kenya}$)

4.4.3 Multiple Inference Corrections

Recall we will only perform multiple hypothesis adjustments for primary outcomes, grouping hypotheses in sets and perform adjustments along those sets. In terms of FDR adjustments, H_{BE1} and H_{DE1} will belong to the set of main hypotheses that will have adjustments on that set, and adjustments will be performed on the set of hypotheses H_{BE2} and H_{BE3} as well as on the set of hypotheses H_{BE3} and H_{DE3}. These results will be part of the core of the analysis.

Given the lack of a regression structure for the analysis of the Time Preference module, we will not include these hypotheses into the sets of overall significance testing. Exploratory analysis will also be served by randomization inference and products leading from such, to be tested in a similar way described above.

4.5 Trust game

We will use the trust game to study how temperature affects the share (of currency one starts with) sent (often referred to as "trust") as well as to study how temperature affects the share of currency sent back (often referred to as "trustworthiness").[13] The main motivation for studying this behavior is that it may be important for societies economic performance and social outcomes. The mechanisms behind the observed behavior may be several such as *efficiency, altruism, inequality, self interest*, and for sending back also *reciprocity*. Note that we cannot separate the potential mechanisms for the observed behavior using this experimental design. That will be left for future work.

4.5.1 Specifications

In the trust game, the participants are matched in pairs. They play the game twice, each time with a different partner. Participant A is given an initial amount X and the other participant, Participant B, is not given any endowment. Note that everyone plays as Participant A before playing as Participant B, because we deliberately want to give priority to the measurement of sending behavior over sending back behavior. Participant A decides how many tokens, Y, to pass on to Participant B. This amount is multiplied by 3. Participant B then decides how much Z, of 3*Y to send back to Participant A. Participant A's payment is X - Y + Z, and Participant B's payment is 3*Y - Z. See Johnson and Mislin (2011) for reference.

The primary outcome of interest is the share of the total amount sent.

$$Y_i = sharesent_i \text{ or } sharesentback_i$$

Trust game: main specification

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 Site + X_i^{'}\Gamma + \varepsilon_i \tag{15}$$

Trust game: heterogeneity

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 (T_i * Gender) + \beta_4 Site + \beta_5 (T_i * Site) + X_i^{'}\Gamma + \varepsilon_i \tag{16}$$

An exploratory outcome of interest is the share sent back. However, for this we slightly modify the specifications, and include as covariates a variable for how much the participant received ($sharesentto_i$) as well as a variable for how much the participant sent when they were Participant A ($sharesent_i$).

Trust game: main specification (for $sharesentback_i$)

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 Site + \beta_4 sharesentto_i + \beta_5 sharesent_i + X_i^{'}\Gamma + \varepsilon_i \tag{17}$$

Trust game: heterogeneity (for $sharesentback_i$)

$$\begin{aligned} Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 (T_i * Gender) + \beta_4 Site + \beta_5 (T_i * Site) + \\ \beta_4 sharesentto_i + \beta_5 sharesent_i + X_i^{'}\Gamma + \varepsilon_i \end{aligned} \tag{18}$$

[13] We will also have a survey question in which we ask participants to use a scale from 0 to 10 to rate how much they agree with the statement "As long as I am not convinced otherwise, I assume that people have only the best intentions." This will be used in the exploratory part of the analysis.

4.5.2 Hypotheses

From (15), we test the null hypothesis that:

- Temperature does not affect share sent ($H_{SS1} : \beta_1 = 0$)

From (16), we test the null hypothesis that:

- Temperature does not affect share sent differentially by gender ($H_{SS2} : \beta_3 = 0$)
- Temperature does not affect share sent differentially by site ($H_{SS3} : \beta_5 = 0$)

From (17), we test the null hypothesis that:

- Temperature does not affect share sent back ($H_{SB1} : \beta_1 = 0$)

From (18), we test the null hypothesis that:

- Temperature does not affect share sent back differentially by gender ($H_{SB2} : \beta_3 = 0$)
- Temperature does not affect share sent back differentially by site ($H_{SB3} : \beta_5 = 0$)

4.5.3 Multiple Inference Corrections

Recall we will only perform multiple hypothesis adjustments for primary outcomes, grouping hypotheses in sets and perform adjustments along those sets. In terms of FDR adjustments, H_{SS1} will belong to the set of main hypotheses that will have adjustments on that set, and adjustments will be performed on the set of hypotheses H_{SS2} and H_{SS3}. These results will be part of the core of the analysis.

For more exploratory analysis, recall we will also perform tests of overall significance on primary outcomes. Thus, H_{SS1} will belong to a set of overall hypotheses that will be tested via SUR. The hypotheses on β_3 and β_5 coming from (16) for primary outcomes (H_{SS2} and H_{SS3}) will be featured in sets of hypotheses of β_3^k and β_5^k, where the joint null hypotheses will also be tested via SUR. Exploratory analysis will also be served by randomization inference and products leading from such, to be tested in a similar way described above.

4.6 Public Goods game

Again, we would like to elicit how pro-social behavior – in particular cooperation – may be affected by temperature.

4.6.1 Specifications

We will conduct a standard public goods games with 3 players. Participants are randomly matched with two other participants. Participants are each endowed with 1200 tokens, and must decide how much of one's endowment to put into a shared fund. Choices are made simultaneously, and each token put into the fund is multiplied by 2. The shared fund is then split equally among the three of you. Each token not put into the fund will be the participant's to keep. Thus, earnings from this part of the experiment depend on the participant's choice and the choice of two other participants in the experiment. After selecting how much they will put into the fund, participants will be asked what they believe another player has put into the fund.

The primary outcome of interest is cooperation, defined as the amount put into the fund. An exploratory outcome of interest (possibly useful for discussion of the mechanism by which temperature might affect behavior) are beliefs, defined as an indicator variable that marks as 1 if the individual guesses correctly about another's contribution, and 0 otherwise.

$$Y_i = cooperation_i \text{ or } beliefs_i$$

Public Goods game: main specification

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 Site + X_i^{'}\Gamma + \varepsilon_i \tag{19}$$

Public Goods game: heterogeneity

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 (T_i * Gender) + \beta_4 Site + \beta_5 (T_i * Site) + X_i^{'}\Gamma + \varepsilon_i \tag{20}$$

4.6.2 Hypotheses

From (19), we test the null hypothesis that:

- Temperature does not affect cooperation ($H_{CN1} : \beta_1 = 0$)
- Temperature does not affect beliefs ($H_{BF1} : \beta_1 = 0$)

From (20), we test the null hypothesis that:

- Temperature does not affect cooperation differentially by gender ($H_{CN2} : \beta_3 = 0$)
- Temperature does not affect cooperation differentially by site ($H_{CN3} : \beta_5 = 0$)
- Temperature does not affect beliefs differentially by gender ($H_{BF2} : \beta_3 = 0$)
- Temperature does not affect beliefs differentially by site ($H_{BF3} : \beta_5 = 0$)

4.6.3 Multiple Inference Corrections

Recall we will only perform multiple hypothesis adjustments for primary outcomes, grouping hypotheses in sets and perform adjustments along those sets. In terms of FDR adjustments, H_{CN1} will belong to the set of main hypotheses that will have adjustments on that set, and adjustments will be performed on the set of hypotheses H_{CN2} and H_{CN3}. These results will be part of the core of the analysis.

For more exploratory analysis, recall we will also perform tests of overall significance on primary outcomes. Thus, H_{CN1} will belong to a set of overall hypotheses that will be tested via SUR. The hypotheses on β_3 and β_5 coming from (20) for primary outcomes (H_{CN2} and H_{CN3}) will be featured in sets of hypotheses of β_3^k and β_5^k, where the joint null hypotheses will also be tested via SUR. Exploratory analysis will also be served by randomization inference and products leading from such, to be tested in a similar way described above.

4.7 Cognitive Ability - Raven's

We will use Raven's matrices to measure cognitive ability. The test of cognitive ability will enable us to identify the effect of temperature on mental acuity and create destructible earnings for the joy of destruction task.

4.7.1 Specifications

As mentioned, we will use Raven's matrices to measure cognitive ability. The participants are **not** told how many they have completed correctly, and are not given their payment following this module. If the person gets 0 to 3 correct (out of 6), they get a low payout; otherwise they get a high payout. See (Frederick, 2005) for reference.

The primary outcome of interest is the percentage of puzzles chosen correctly.

$$Y_i = puzzles_i$$

Cognitive Ability - Raven's: main specification

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 Site + X_i^{'}\Gamma + \varepsilon_i \tag{21}$$

Cognitive Ability - Raven's: heterogeneity

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 (T_i * Gender) + \beta_4 Site + \beta_5 (T_i * Site) + X_i^{'}\Gamma + \varepsilon_i \tag{22}$$

4.7.2 Hypotheses

From (21), we test the null hypothesis that:

- Temperature does not affect percentage of puzzles chosen correctly ($H_{PZ1} : \beta_1 = 0$)

From (22), we test the null hypothesis that:

- Temperature does not affect percentage of puzzles chosen correctly differentially by gender ($H_{PZ2} : \beta_3 = 0$)
- Temperature does not affect percentage of puzzles chosen correctly differentially by site ($H_{PZ3} : \beta_5 = 0$)

4.7.3 Multiple Inference Corrections

Recall we will only perform multiple hypothesis adjustments for primary outcomes, grouping hypotheses in sets and perform adjustments along those sets. In terms of FDR adjustments, H_{PZ1} will belong to the set of main hypotheses that will have adjustments on that set, and adjustments will be performed on the set of hypotheses H_{PZ2} and H_{PZ3}. These results will be part of the core of the analysis.

For more exploratory analysis, recall we will also perform tests of overall significance on primary outcomes. Thus, H_{PZ1} will belong to a set of overall hypotheses that will be tested via SUR. The hypotheses on β_3 and β_5 coming from (22) for primary outcomes (H_{PZ2} and H_{PZ3}) will be featured in sets of hypotheses of β_3^k and β_5^k, where the joint null hypotheses will also be tested via SUR. Exploratory analysis will also be served by randomization inference and products leading from such, to be tested in a similar way described above.

4.8 Joy of Destruction

Here we would like to measure whether willingness to destruct increases with temperature (see Abbink and Sadrieh 2009).

4.8.1 Specifications

Participants will be informed that everyone has won different amounts of $1 amazon gift cards or air time vouchers worth 50 Ksh (item depending on location). They are then matched in pairs and told how much their partner has won X (Note that they will not know how much they themselves have won). They can destroy any number between 1 and X. The computer may also destroy some of the remaining vouchers (after flipping a virtual coin, if the coin is heads, it doesnt destroy anything, otherwise it destroys any number of the remaining cards). The lab assistant will destroy the total number of cards given by the computer's and participants choice. The other participant does not know whether the earnings were destroyed because of the computer or because of the other participant's decision. The idea here is that participants can partly hide their purposeful destruction behind random destruction. The game is conducted with gift cards/vouchers, so that actual destruction (and not reallocation) can take place.

The primary outcome of interest is a measure of destruction, defined as the percentage of gift cards or vouchers destroyed.

$$Y_i = destroyed_i$$

Joy of Destruction: main specification

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 Site + X_i^{'}\Gamma + \varepsilon_i \tag{23}$$

Joy of Destruction: heterogeneity

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 (T_i * Gender) + \beta_4 Site + \beta_5 (T_i * Site) + X_i^{'}\Gamma + \varepsilon_i \tag{24}$$

4.8.2 Hypotheses

From (23), we test the null hypothesis that:

- Temperature does not affect destruction ($H_{JD1} : \beta_1 = 0$)

From (24), we test the null hypothesis that:

- Temperature does not affect destruction differentially by gender ($H_{JD2} : \beta_3 = 0$)
- Temperature does not affect destruction differentially by site ($H_{JD3} : \beta_5 = 0$)

4.8.3 Multiple Inference Corrections

Recall we will only perform multiple hypothesis adjustments for primary outcomes, grouping hypotheses in sets and perform adjustments along those sets. In terms of FDR adjustments, H_{JD1} will belong to the set of main hypotheses that will have adjustments on that set, and adjustments will be performed on the set of hypotheses H_{JD2} and H_{JD3}. These results will provide the core of our analysis.

For more exploratory analysis, recall we will also perform tests of overall significance on primary outcomes. Thus, H_{JD1} will belong to a set of overall hypotheses that will be tested via SUR. The hypotheses on β_3 and β_5 coming from (24) for primary outcomes (H_{JD2} and H_{JD3}) will be featured in sets of hypotheses of β_3^k and β_5^k, where the joint null hypotheses will also be tested via SUR. Exploratory analysis will also be served by randomization inference and products leading from such, to be tested in a similar way described above.

4.9 Cognitive Reflection

We will use a Cognitive Reflection Test (CRT) that constitutes six survey questions to elicit potential treatment effects on cognitive reflection. These survey questions are all versions of standard such questions (see Frederick, 2005).[14]

4.9.1 Specifications

We include this module to study whether temperature affects cognitive reflection. We theorize that participants may answer a question incorrectly because 1) the questions may cognitively be difficult to work through, or 2) the questions are designed to be somewhat misleading, and so people might answer intuitively on questions answered incorrectly.

Thus, we will proceed analysis in two stages. In the first stage, we will first determine if the share of questions answered correctly is similar between treatment and control groups. If there is no difference, then we do not proceed to the second stage. In the second stage (more speculative), we will study whether there is a treatment effect on giving an intuitive answer to a CRT question answered incorrectly. Thus, we will pool responses to the questions answered incorrectly by individual, and have each response as values of the outcome variable, running our analysis with question fixed effects and individual fixed effects.

The primary outcome of interest is share of questions answered correctly. An exploratory outcome of interest is the probability that the incorrect question q was given the intuitive answer. For exploratory outcomes, we will also examine time spent in a similar way, where in the first stage we will examine time spent on all questions, and in the second stage we will examine time spent on each question answered incorrectly.

For the first stage:

$$Y_i = sharecorrect_i \text{ or } Y_i = timespentoverall_i$$

For the second stage:

$$Y_{iq} = answerintuitive_{iq} \text{ or } timespent_{iq}$$

Compared to other standard specifications, we will include quesion fixed effects and individual fixed effects in the second stage. In these specifications, Z_i is a vector of variables demarcating questions in the CRT and P_i is a vector of variables demarcating individuals in the observed setting.

Cognitive Reflection: main specification

For the first stage:

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 Site + X_i^{'}\Gamma + \varepsilon_i \tag{25}$$

For the second stage:

$$Y_{iq} = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 Site + X_i^{'}\Gamma + Z_q^{'}\Theta + P_i^{'}\Lambda + \varepsilon_{iq} \tag{26}$$

Cognitive Reflection: heterogeneity

[14] The survey questions are as follows: 1) Mary's mother had four children. The youngest three are named Spring, Summer, and Autumn. What is the oldest child's name?; 2) Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are in the class?; 3) If you flipped a fair coin 3 times, what is the probability that it would land "Heads" *at least* once?; 4) A bear loses 20% of its weight during hiberation. If it weighs 100 pounds *after* hibernation, how many pounds did it weight *before*?; 5) A notebook and a pen cost 22 dollars in total. The notebook cost 20 dollars more than the pen. How many dollars does the notebook cost?; and 6) If it takes 5 machines 5 minutes to make 5 bricks, how many minutes would it take 100 machines to make 100 bricks?

For the first stage:

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3(T_i * Gender) + \beta_4 Site + \beta_5(T_i * Site) + X_i^{'}\Gamma + \varepsilon_i \tag{27}$$

For the second stage:

$$\begin{aligned} Y_{iq} = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3(T_i * Gender) + \beta_4 Site + \beta_5(T_i * Site) + \\ X_i^{'}\Gamma + Z_q^{'}\Theta + P_i^{'}\Lambda + \varepsilon_{iq} \end{aligned} \tag{28}$$

4.9.2 Hypotheses

From (25), we test the null hypothesis that:

- Temperature does not affect share of questions answered correctly ($H_{SC1} : \beta_1 = 0$)
- Temperature does not affect time spent on all questions ($H_{TS1} : \beta_1 = 0$)

From (26), we test the null hypothesis that:

- Temperature does not affect probability that incorrect question was given the intuitive answer ($H_{PI1} : \beta_1 = 0$)
- Temperature does not affect time spent on a question answered incorrectly ($H_{TE1} : \beta_1 = 0$)

From (27), we test the null hypothesis that:

- Temperature does not affect share of questions answered correctly differentially by gender ($H_{SC2} : \beta_3 = 0$)
- Temperature does not affect share of questions answered correctly differentially by site ($H_{SC3} : \beta_5 = 0$)
- Temperature does not affect time spent on all questions differentially by gender ($H_{TS2} : \beta_3 = 0$)
- Temperature does not affect time spent on all questions differentially by site ($H_{TS3} : \beta_5 = 0$)

From (28), we test the null hypothesis that:

- Temperature does not affect probability that incorrect question was given the intuitive answer differentially by gender ($H_{PI2} : \beta_3 = 0$)
- Temperature does not affect probability that incorrect question was given the intuitive answer differentially by site ($H_{PI3} : \beta_5 = 0$)
- Temperature does not affect time spent on a question answered incorrectly differentially by gender ($H_{TE2} : \beta_3 = 0$)
- Temperature does not affect time spent on a question answered incorrectly differentially by site ($H_{TE3} : \beta_5 = 0$)

4.9.3 Multiple Inference Corrections

Recall we will only perform multiple hypothesis adjustments for primary outcomes, grouping hypotheses in sets and perform adjustments along those sets. In terms of FDR adjustments, H_{SC1} will belong to the set of main hypotheses that will have adjustments on that set, and adjustments will be performed on the set of hypotheses H_{SC2} and H_{SC3}. These results will be part of the core of the analysis.

For more exploratory analysis, recall we will also perform tests of overall significance on primary outcomes. Thus, H_{SC1} will belong to a set of overall hypotheses that will be tested via SUR. The hypotheses on β_3 and β_5 coming from (27) for primary outcomes (H_{SC2} and H_{SC3}) will be featured in sets of hypotheses of β_3^k and β_5^k, where the joint null hypotheses will also be tested via SUR. Exploratory analysis will also be served by randomization inference and products leading from such, to be tested in a similar way described above.

4.10 Demographic survey questions

The demographic survey questions include non-incentivized questions on trust,[15] physiological questions on body mass, demographics on the participant, and on the family's education level, income, and occupational status. As described previously, any variable (aside from the questions on trust) that is found to be unbalanced across treatment group will be included in the regression at the level where the unbalance was found.

4.11 Charity

In this module we want to measure how a participant's willingness to donate part of their earnings to charity varies with temperature, and furthermore, whether high temperature makes participants more or less likely to reveal in-group biases (based on ethnicity or residency status) with temperature.

4.11.1 Specifications

At the very end of the experiment, after the survey, we will give them their earnings and offer them to donate part of it to a charity. Participants are randomly allocated to charities on a list (7 in Busara, 6 in US) and can donate a percentage up to 40% of their earnings.

The primary outcome of interest is the absolute amount chosen to donate.

$$Y_i = donation_i$$

Compared to other standard specifications, we will include an indicator variable ($Ingroupdonation_i$) that takes a value of 1 if the individual donates to an organization for which she shares in-group status with, and 0 otherwise. In the main specification, $Ingroupdonation_i$ will act as a control. We will also interact this indicator variable with treatment in heterogeneity analysis. Furthermore, we will have earnings contributed ($earnings_i$) as a control variable.

Charity: main specification

$$\begin{aligned} Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 Site + \beta_4 Ingroupdonation_i + \\ \beta_5 earnings_i + X_i^{'}\Gamma + \varepsilon_i \end{aligned} \tag{29}$$

[15] Participants are asked to rate how well the statement "As long as I am not convinced otherwise, I assume that people have only the best intentions." describes them as a person. Participants will use a scale from 0 to 10 to answer, where 0 means does not describe me at all" and a 10 means describes me perfectly". See Dohmen et al. (2011) for reference.

Charity: heterogeneity

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Gender + \beta_3 (T_i * Gender) + \beta_4 Site + \beta_5 (T_i * Site) + \beta_6 Ingroupdonation_i + \beta_7 (T_i * Ingroupdonation_i) + \beta_8 earnings_i + X_i^{'}\Gamma + \varepsilon_i \tag{30}$$

4.11.2 Hypotheses

From (29), we test the null hypothesis that:

- Temperature does not affect amount donated ($H_{CH1} : \beta_1 = 0$)

From (30), we test the null hypothesis that:

- Temperature does not affect amount donated differentially by gender ($H_{CH2} : \beta_3 = 0$)
- Temperature does not affect amount donated differentially by site ($H_{CH3} : \beta_5 = 0$)
- Temperature does not affect amount donated differentially by sharing in-group status ($H_{CH4} : \beta_7 = 0$)

4.11.3 Multiple Inference Corrections

Recall we will only perform multiple hypothesis adjustments for primary outcomes, grouping hypotheses in sets and perform adjustments along those sets. In terms of FDR adjustments, H_{CH1} will belong to the set of main hypotheses that will have adjustments on that set, and adjustments will be performed on the set of hypotheses H_{CH2} to H_{CH4}. These results will be part of the core of the analysis.

For more exploratory analysis, recall we will also perform tests of overall significance on primary outcomes. Thus, H_{CH1} will belong to a set of overall hypotheses that will be tested via SUR. The hypotheses on β_3 and β_5 coming from (30) for primary outcomes (H_{CH2} and H_{CH4}) will be featured in sets of hypotheses of β_3^k and β_5^k, where the joint null hypotheses will also be tested via SUR. Exploratory analysis will also be served by randomization inference and products leading from such, to be tested in a similar way described above.

References

Abbink, K. and A. Sadrieh (2009). The pleasure of being nasty. *Economics Letters 105 (3)*, 306–308.

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

Andreoni, J., M. A. Kuhn, and C. Sprenger (2015). Measuring time preferences: A comparison of experimental methods. *Journal of Economic Behavior & Organization 116*, 451–464.

Ashton, L. (2015). Hunger games: Does hunger affect time preferences? *Working Paper*.

Benjamini, Y., A. M. Krieger, and D. Yekutieli (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika 93 (3)*, 491–507.

Burke, M. B., E. Miguel, S. Satyanath, J. A. Dykema, and D. B. Lobell (2009). Warming increases the risk of civil war in Africa. *PNAS 106 (49)*, 20670–20674.

Bushman, B. J., C. N. DeWall, R. S. P. Jr., and M. D. Hanus (2014). Low glucose relates to greater aggression in married couples. *PNAS 111 (17)*, 6254–6257.

Castillo, M., D. L. Dickinson, and R. Petrie (2017). Sleepiness, choice consistency, and risk preferences. *Theory and Decision 82*, 41–73.

Cherry, T. L., S. Kroll, and J. F. Shogren (2005). The impact of endowment heterogeneity and origin on public good contributions: evidence from the lab. *Journal of Economic Behavior & Organization 57*, 357–365.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011). Individual risk attitudes: Measurement, determinants, and individual consequences. *Journal of the European Economic Association 9 (3)*, 522–550.

Eckel, C. C. and P. J. Grossman (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior & Organization 68*, 1–17.

Fisher, R. A. (1935). *The Design of Experiments*, pp. 245. Edinburgh: Oliver & Boyd.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives 19 (4)*, 25–42.

Haushofer, J. and E. Fehr (2014). On the psychology of poverty. *Science 344*, 862.

Hsiang, S. M., M. B. Burke, and E. Miguel (2013). Quantifying influence of climate on human conflict. *Science 341*, 1212.

Johnson, N. D. and A. A. Mislin (2011). Trust game: A meta-analysis. *Journal of Economic Psychology 32*, 865–889.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies 76*, 1071–1102.

Mani, A., S. Mullainathan, E. Shafir, and J. Zhao (2013). Poverty impedes cognitive function. *Science 341 (6149)*, 976–980.

Young, A. (2016). Channelling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *Working Paper*.

Figure 1: Minimum Detectable Effect (MDE) curves for Berkeley sample

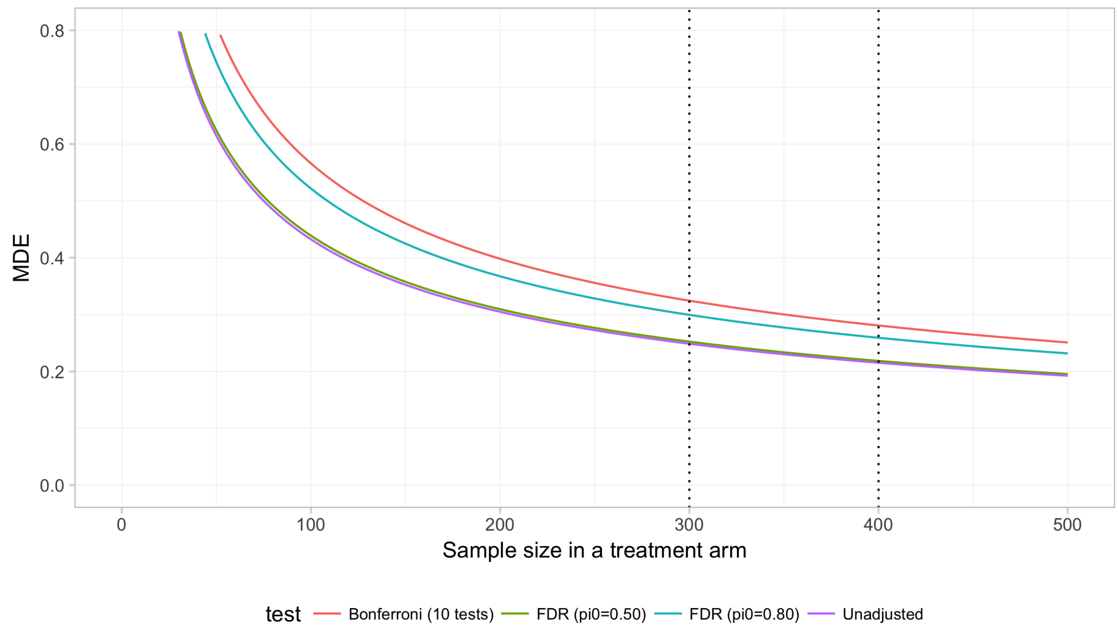

Figure 2: Total participants cost curves for Berkeley sample

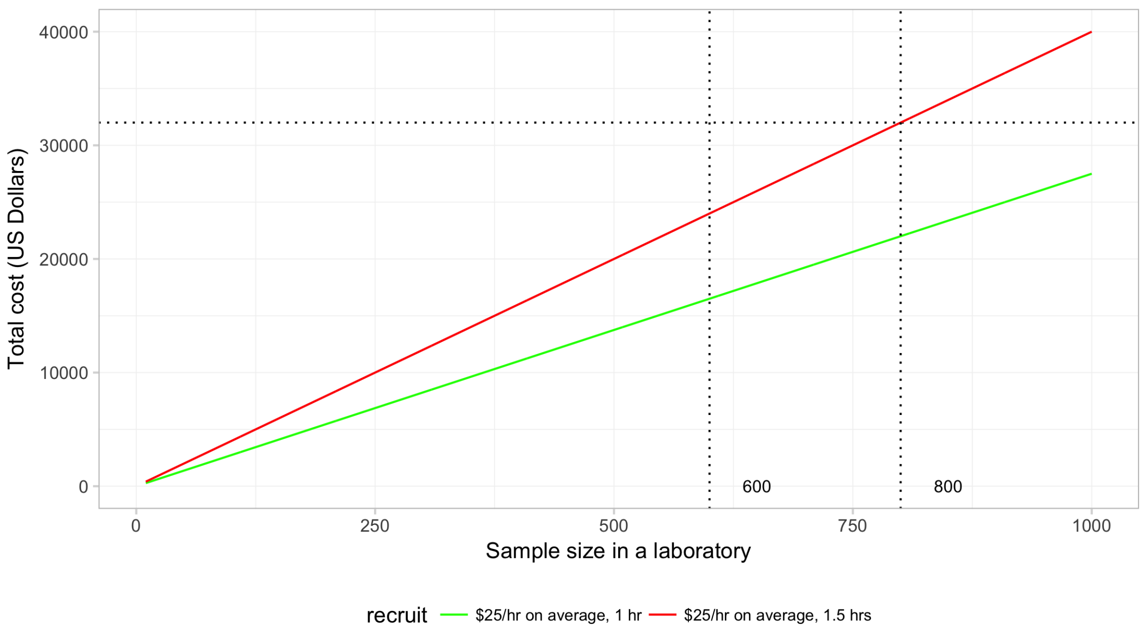

Figure 3: Minimum Detectable Effect (MDE) curves for Nairobi sample

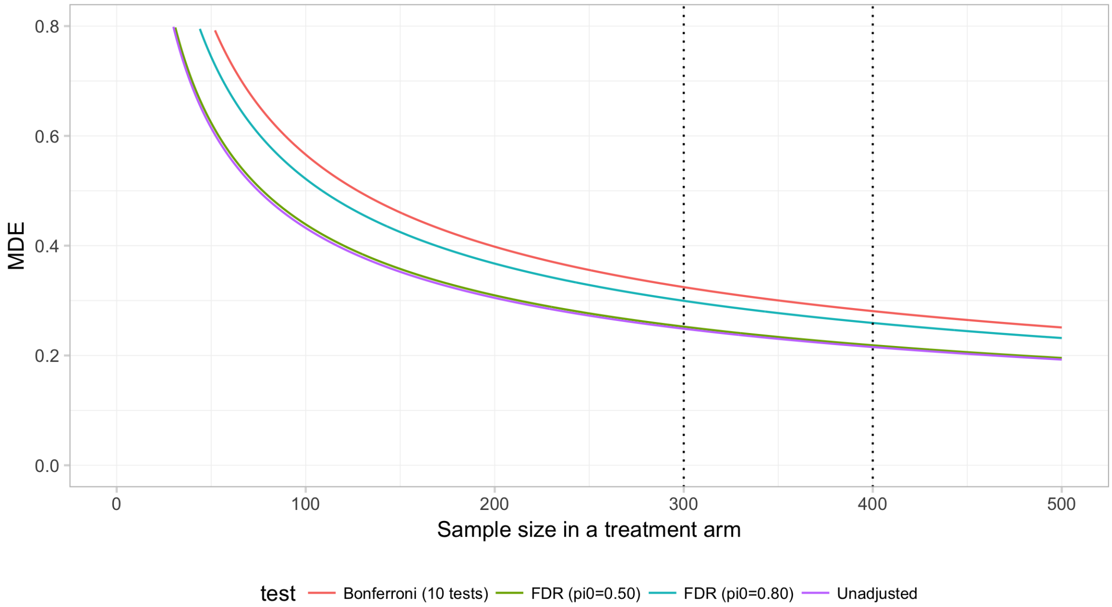

Figure 4: Total participants cost curves for Nairobi sample

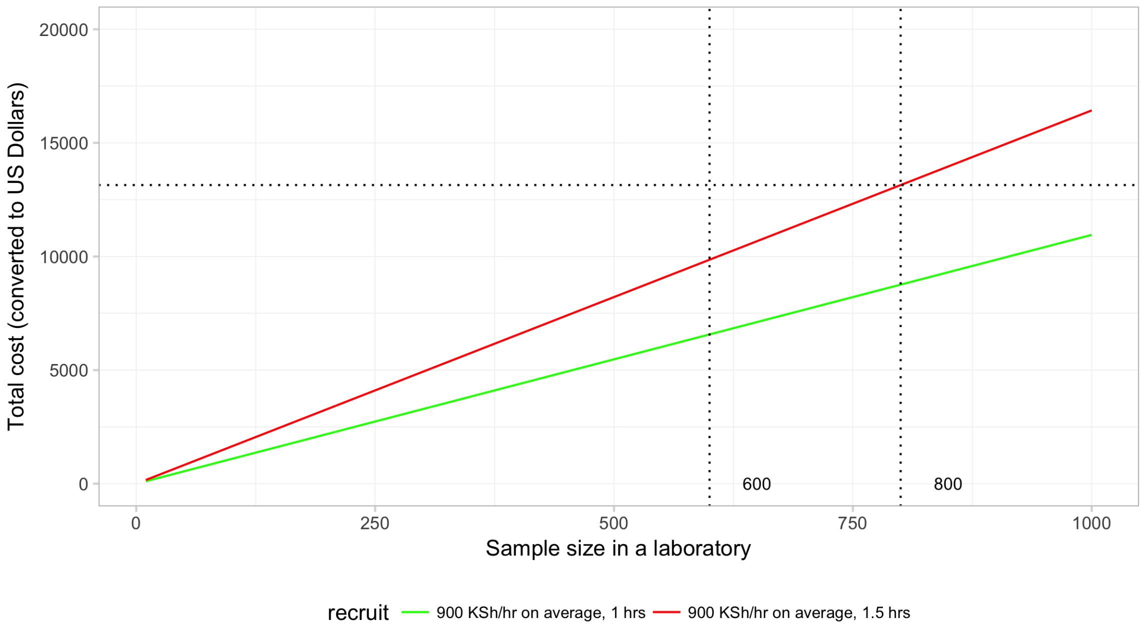

Figure 5: Minimum Detectable Effect (MDE) curves for overall sample