Towards Data-Leveraged Behavioral Policy Design for Alleviating Peak Electricity Demand

by

Ankit Pat

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2015

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The problem of managing peak electricity demand is of significant importance to utility providers. In Ontario, electricity consumption achieves its peak during the afternoon hours in summer. Electricity generation units are provisioned for these few days of the year, which is expensive. In the past, researchers have studied several approaches to curb peak electricity demand by providing consumers with incentives to reduce their load.

We study using non-cash (or behavioral) incentives to motivate consumers to set their thermostats a few degrees higher during the summer, thereby reducing aggregate peak demand. Such incentives exploit cognitive biases and find their foundations in behavioral economics and psychology. We mathematically model the effect of non-cash incentives using utility functions. To build an accurate utility model, we devise and conduct a large-scale survey to elicit consumers' behavioral preferences. At a high level, we propose an analytical Big-Data based approach to evidence-based policy design, where a mechanism design framework uses a data-driven utility model to inform incentive policies.

Acknowledgements

I am grateful to my advisors Prof. S. Keshav and Prof. Kate Larson for their constant guidance, support, research directions, and mentorship. This thesis would not have been possible without their critical inputs and supervision.

I would like thank my thesis readers Prof. Jesse Hoey and Prof. Pascal Poupart for their valuable feedback. I also thank Verena Tiefenbeck for sharing insightful details about psychometric survey design research. For their helpful comments and feedback, I am very thankful to Arthur Carvalho, John A. Doucette, Hadi Hosseini, Omid Ardakanian, Adedamola Adepetu, Rayman Preet Singh, Tommy Carpenter, Alimohammad Rabbani, Fiodar Kazhamiaka, Alan Tsang, Vijay Menon, and other members of the ISS4E and the Artificial Intelligence labs at University of Waterloo.

I am grateful to my friends Ashish Ranjan Hota, Karthik Velakur, Hemant Saxena, Neeraj Kumar, and David Szepesvari for always being helpful and fostering a motivating environment during my stay at Waterloo.

Dedication

To all who have positively influenced my life – my parents, teachers, friends, and a few metaphysical sources of inspiration.

Table of Contents

List of Figures

Chapter 1

Introduction

In this thesis we propose, study, and validate non-cash policies to reduce electricity usage during peak load hours in residential buildings. The basic approach is to craft energy policies that exploit psychological inclinations or biases in humans. By doing so, we argue that we can modify electricity usage behavior and help reduce consumption during peak load hours.

Electricity consumption in Ontario exhibits cyclic patterns of highs and lows during the day and the year. Usage typically increases as the day starts and progresses into the afternoon. It peaks sometime in the afternoon, then gradually decreases as the evening approaches. A second smaller spike is seen in the evening when people come back home from work, which then gradually decreases and attains a low point during the night. The exact times when such changes in the *load profile* occur, depends on several factors such as the day of the week, and the season. However, in Ontario, the general shape of the load profile is relatively constant [52], and the highest demand is observed during summer afternoons. Since these peak demand days do not occur at other times of the year, utility companies either have to maintain an over-provisioned generation infrastructure or purchase electricity from neighboring states or provinces (say) to meet this demand.

One approach to alleviate this problem, has been the employment of *demand response* strategies. Demand response is defined as: "Changes in electric usage by end-use customers from their normal consumption patterns in response to changes in the price of electricity over time, or to incentive payments designed to induce lower electricity use at times of high wholesale market prices or when system reliability is jeopardized" [19]. Among demand response strategies, the use of a time-varying price to control electricity usage in home and buildings is widely used. Essentially, the approach is to increase the price of electricity during periods where the utility companies seek to reduce demand. Several approaches such as time-of-use pricing, critical peak pricing, real-time pricing, and peak-time rebates are well known and are discussed by Newsham et al. in a comprehensive review [48]. A review of these approaches can also be found in the survey in [49].

An analysis of the use of prices to control electricity consumption reveals that it suffers from the following intrinsic problems:

- By having customers manually control demand in response to price signals, it puts the customer into a decision-making loop. This is onerous because most consumers are unaware of the power drawn by various appliances, and so do not know how best to reduce their demand.
- Due to the human in the control loop, prices cannot be changed very rapidly. For instance, time-of-use prices are typically changed every six months. Even real-time pricing changes only hourly. Grid peaks, however, can develop in a matter of seconds or minutes. Therefore, price-based control is not always effective.
- The typical savings from changing behavior amounts only to a few dollars per month. So, it is not sufficient to induce any significant change in usage behavior.

For these reasons, we believe that the traditional price-based approach to controlling electricity usage is inadequate. Instead, we study the use of non-cash (or behavioral) incentives for inducing a change in consumption pattern.

We are strongly motivated by two phenomena. First, electronic platforms and social media are bringing the world together. The social and geographical barriers that were previously required to be crossed to spread an idea are falling. This may provide the thrust that may be needed for non-cash incentives based on a desire for social recognition to be widely adopted! Second, *smart grids* make it feasible to communicate with and control electrical equipment without manual intervention. This enables the design of policies that do not need participants to be personally involved: only initial consent is needed. These two developments make behavioral policies inexpensive to implement, with the potential for huge gains.

We are also motivated by the fact that most prior work on policy-making based on behavioral changes has been experimental [14]. While such experimental studies do show the effectiveness of behavioral incentives, they conduct little analysis about quantitative results from the implementation of these policies. Specifically, researchers in the past have shown that behavioral incentives are effective and that behavioral policies should be exploited in policy-making [15]. In contrast, we quantify the value received by an individual when a particular policy is implemented. This helps us select the best set of policies to implement.

We focus on non-cash incentives for controlling peak demand during summer afternoons, by encouraging homeowners to set their thermostat a few degrees higher. We model behavioral incentives using utility functions and design an agent-based game theoretic model. We use this model to investigate agents' participation as a function of the design parameters. To build a realistic utility function, we conduct a survey of our target population to elicit preferences regarding proposed policies and discomfort anticipated due to higher thermostat settings. A major contribution of our work has been designing a survey questionnaire that provides us with data to experiment with a variety of schemes for non-cash incentives, while proposing policy measures with appropriate communication and control protocols.

In our work, we try to answer the following questions: How effective are non-cash incentives in the context of demand response? Which incentives or biases are better than others, and if so, for what sections of the population? How do we quantify the reduction in electricity that would be seen if we implement a certain behavioral policy? And, in general, how do we decide what kind of policies would work, before we spend a significant sum of money implementing them on the larger population? This thesis aims to answer these questions.

Specifically, our contributions are as follows:

- We design and recommend a set of behavioral policies geared towards increasing penetration of the *peaksaver* PLUS program in Ontario (this scheme is described in Chapter 2).
- We devise a game theoretic framework and carry out large-scale data-driven preference elicitation; leveraging these to design a computational (empirical) mechanism. We call this approach *Big Data Mechanism Design*.
- We propose a game theoretic model that accounts for an agent's overall utility received from behavioral biases and thermal discomfort.
- We design and conduct a large-scale survey on a crowdsourcing platform called CrowdFlower. The survey elicits respondents' preferences regarding behavioral biases and thermal discomfort.
- We propose and implement a novel post hoc survey cleaning methodology. This method can be used for selecting questions that best measure a particular subjective variable, if we have multiple questions that all aim to elicit data about the same variable.
- We evaluate and compare profits from inclusion or exclusion of various behavioral policies in the implemented policy set.
- We find that if the utility companies implement our proposed policies, they can benefit the equivalent energy savings of raising the thermostats of the entire population by $2°C$. The projected energy savings can be up to 78% higher if the utility companies raise thermostat settings in a customized way, based on each participant's preferences. We, therefore, recommend that our survey be carried out on a large scale to identify and target sections of the population accordingly.

The rest of the thesis is organized as follows. In the next chapter, we discuss the relevant background and related work. In Chapter 3, we present a game theoretic framework that captures the policy design process, and instantiate the model with the *peaksaver* PLUS program. To elicit agents' preferences, we design and conduct a survey – the details of which are covered in Chapter 4. In Chapter 5, we analyze the data from the survey, lay down the principles for interpreting the survey data, and draw inferences from our model based on the data. Finally, we discuss the conclusions of our study in Chapter 6.

Chapter 2

Preliminaries and Related Work

Our work explores behavioral biases and their incorporation into energy policies. In this chapter, we present concepts, definitions, and related work.

2.1 The *peaksaver* PLUS Program

We chose the *peaksaver* PLUS program to contextualize and instantiate our model for policy design. *peaksaver* PLUS is designed to reduce electricity demand. It is offered by several electricity distribution companies in Ontario, Canada; and is funded by the Ontario Power Authority. This program allows the system operator to send signals and control the household's central air conditioning system and electric water heater during peak demand hours using a radio-controlled programmable digital thermostat. It has been shown to reduce peak demand for the province of Ontario and also saves money for the consumer by reducing their electricity consumption. The programmable thermostat is installed for free in volunteering homes. Additionally, as a perk, the *peaksaver* PLUS program also provides participating homes with an in-home energy display device that provides real-time information about their rate of consumption, the peak price, and the monthly and daily projections of their electricity bill.

Typically, the *peaksaver* PLUS thermostat is activated during hot summer days, from

May 1 to September 30, on weekdays (never on weekends and holidays), between 12 noon and 7 p.m. The thermostat temperatures are generally set up to $2°C$ higher [6, 7]. Such adjustments are executed in batches of activation, each one of which may last for up to four hours. The *peaksaver* PLUS program has been shown to be very effective, with Enersource Hydro Mississauga Inc. claiming that the program can lower a household's electricity consumption by one to two kilowatt-hours [6]. Historically, between 2011 and 2014, *peaksaver* PLUS saved 410,345 kWh of peak load electricity, and was reported to have 54,451 participants in 2013 [45].

2.2 Preliminaries in Behavioral Economics

Much has been said and written about the irrationality of human behavior [16, 17, 43]. An entire branch of economics called Behavioral Economics studies the consequences of irrationality in human behavior and decision-making. Behavioral Economics advocates that human beings are irrational, and contradicts the dogma that humans are rational and always work in their best interests.

2.2.1 Cognitive Biases

Cognitive biases are the psychological phenomena that human irrationality can be attributed to [16, 17]. In the context of our study, we treat cognitive biases as non-cash incentives that may modify consumer behavior. Since cognitive biases are fundamental predispositions in the decision-making process of humans, in utility theory, we model cognitive biases as imparting some non-negative utility to a *biased* decision-maker. [1] Given this interpretation, cognitive biases can be regarded as incentives. Hereafter, we will be using the terms non-cash incentives, behavioral incentives, behavioral biases, and cognitive biases interchangeably.

[1] We can make this assumption because, in our model, an individual is never penalized due to the effect of these biases.

In this work, we investigate the usage of four biases in policy-making. Three of these are cognitive biases, and the fourth one is a motivational technique called *Gamification*. Although, in the literature, gamification is not modeled as a bias, for the sake of consistency in expression, from here on we will refer to gamification also as a bias. We define these biases as follows:

Status Quo Bias

Status quo bias is a cognitive bias that represents an agent's tendency to prefer the current state of affairs or the status quo to any change [35, 36, 53]. For example, behaviors such as students sitting at the same seats for lectures, and people not switching their default mobile ringtone are attributed to the status quo bias [43].

Commitment Devices

Commitment devices refer to an agent's commitment to abide by a future course of action so as to produce a desired result, usually to avoid being tempted into taking actions that might be undesirable. Quoting Bryan et al. from [21], for example, "Not keeping alcohol in the house," and "Only taking a fixed amount of cash when heading out to party for a night," are examples of commitment devices at work.

Social Norms

Social norms are standards set by the society that individuals who are a part of the society generally tend to follow [23, 51]. For example, conforming with dining etiquettes, tipping the waiter, queuing up patiently for one's turn, are all examples of social norms.

Gamification

Gamification refers to the introduction of fundamental elements of game design into non-game contexts [26]. In other words, gamification refers to turning non-game contexts into games. Some examples are – Facebook Likes, karma points on online forums like Reddit, and marks or grades in the academic system.

2.3 Discomfort Modeling

Our work investigates the awarding of incentives to increase participation in the *peaksaver* PLUS program. This program inherently causes additional thermal discomfort, because a system operator sends signals to control a participating household's air conditioning system. Therefore, it is important that we accurately model thermal discomfort.

2.3.1 Predicted Mean Vote and ASHRAE's Standard

Fanger's Predicted Mean Vote (PMV) model [27] is used to calculate a numerical value that represents the average human's comfort level in a certain thermal environment. PMV is computed as a function of air temperature, radiant temperature, air speed, and humidity of the environment, along with the person's clothing and physical activity level.

The ASHRAE Standard 55 [58] is a standardized scale used for measuring an environment's thermal comfort for human occupancy. According to the standard, if the PMV, calculated from the environmental and personal parameters, is between -0.5 and 0.5, the thermal environment is deemed to be comfortable.

2.4 Related Work

2.4.1 The *peaksaver* Program and Behavioral Energy Policies

Singla et al. proposed a payment-based incentive scheme called *smartset*, that offered monetary rewards to consumers for participating in the *peaksaver* program (a predecessor of *peaksaver* PLUS) [56]. In [59], Sugarman et al. investigated and compared end-user response to time-of-use pricing and the *peaksaver* program; showing that while the *peaksaver* program can be potentially very effective, in its current form it is not attractive to consumers. This, coupled with a dearth of research on the use of behavioral incentives for promoting adoption of the *peaksaver* PLUS program, motivates our present study.

Allcot et al., in their work [14], outline the importance of implementing behavioral biases in energy policies. In particular, they stress the relevance of *framing and psychological cues*, commitment devices, *default options* (status quo bias), and social norms, in energy policies. Experimental use of behavioral energy policies has also been adopted in several pilot programs initiated by a company called OPOWER, in the United States [15,40]. Our work extends on these propositions and conducts detailed quantitative analysis using data from consumers.

2.4.2 Psychometric Survey Design

To build a realistic and practical model of the degree to which irrational behavior can be exploited to mitigate thermal discomfort, we need to conduct psychometric surveys that can measure behavioral biases and the tolerance for thermal discomfort. This section presents previous work on psychometric scales that measure various aspects of human behavior and biases. Such standardized scales are universally accepted psychological research to measure specific aspects of human psychology and behavior. An established scale requires strict adherence to the exact set of questions as recommended by the developers of the scale. Of course, these questions may not necessarily be a perfect fit for the research question at hand, in which case, a fresh questionnaire may be designed to elicit the required information, as we did. We now discuss some standardized psychometric scales [2] that are related to eliciting behavioral preferences or patterns. [3]

[2] Developed by separate research groups, these scales are mutually incompatible. For example, some scales measure values on a single dimensional space, while others measure values in a multi-dimensional space.

[3] Further details on the psychometric scales, including example questions, can be found in Appendix B.

Social Comparison Scale

The social comparison scale [54] deals with two distinctive underlying dimensions of social comparisons: (a) comparisons of abilities referring to the question "How am I doing as compared to others?", and (b) comparisons of opinions referring to the question "How do my feelings/thoughts compare to those of others?"

HEXACO & the Big Five Inventory Scales

The HEXACO [41] and the Big Five Inventory [33] scales claim to measure fundamental personality traits along different dimensions. The Big Five Inventory measures personality traits along the five dimensions of Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness. The set of personality traits considered by the HEXACO scale is a superset of the Big Five Inventory, with the additional dimension of Honesty-Humility. This additional dimension represents traits like manipulative behavior for personal gain, temptation to break rules, lack of interest in lavish wealth and luxuries, feeling no special entitlement to elevated social status, etc.

Self-Monitoring Scale

Self-monitoring scale [42] measures the respondent's sensitivity to the expressive behavior of others, and their ability to modify self-presentation.

Self-Efficacy Scale

Jerusalem and Schwarzer devised the General Self-efficacy scale [55]. The scale consists of ten items and, as its name indicates, was created to assess a general sense of perceived self-efficacy.

Self-Control Scale

The self-control scales, proposed in [60], look quite relevant; particularly because commitment bias is known to be linked with the ability to overcome temptations [21].

Heuristics and Biases Scale

This scale [57] has questions that elicit information about Hindsight bias, Conjunction fallacy, Planning fallacy, Halo effect, Phenomenon of rare Events, Overconfidence bias, and Susceptibility to priming.

Need for a Scale for Biases

To the best of our knowledge, there are no existing psychometric scales for measuring preferences on social norms, commitment bias, status quo bias, and gamification. The social comparison scale is a well-cited measurement scale for social comparisons, but it does not directly measure peer pressure. The conscientiousness dimension on the HEXACO scale can be argued to be closely related to commitment bias, but without any strong grounds. The self-monitoring scale and self-efficacy scale are also not applicable for measuring the biases that we consider in our study. The heuristics and biases scale does measure some cognitive biases, but not the ones we need. Moreover, since using an established scale requires us to use its recommended questionnaire, these scales are not relevant in the context of our problem statement. Thus, we have a need to develop a standardized measurement scale for measuring social norms, commitment bias, status quo bias, and gamification. We do so in Chapter 3.

2.4.3 Big Data and Mechanism Design

In [64], Wellman discusses empirical game-theoretic analysis for games that do not have tractable computational or theoretical solution strategies. For such games, Wellman treats the game simulator as the primary source of input and performs strategic reasoning through a combination of simulation and game-theoretic analysis. Vorobeychik et al., in their work [61–63], study the broad topic of computational mechanism design where exact theoretical analysis is intractable. In particular, Vorobeychik et al. substitute players and their payoff profiles with simulation-based game models, provide Nash equilibrium approximation solution and investigate its convergence results for particular game types such as infinite simulation-based games.

Our work differs from simulation-based mechanism design in the fact that we do not obtain value functions from simulations; rather we investigate and elicit data regarding value functions from human participants using a survey. Depending on the nature of the data and results from our analyses, we make appropriate modeling assumptions about the agents' type distribution, and tractably solve our mechanism design problem.

Overall, there has been limited research on designing empirical mechanisms based on a large volume of elicited data from human participants. With big data analytics, policy makers can hugely benefit from data-based computational mechanisms that potentially have a greater modeling accuracy than theoretical mechanisms.

Chapter 3

Policy Design and Game Theoretic Framework

In this chapter, we delve deeper into the application of behavioral biases in policy design. To analyze the potential impact of various behavioral policies, we formally describe a game theoretic framework and model suitable utility functions.

3.1 Policy Design

In this section, we propose policy recommendations that leverage four behavioral biases. In particular, we suggest policies that incorporate the status quo bias, commitment devices, social norms, and gamification. The policies are designed to encourage participation in the *peaksaver* PLUS program, that is, authorizing the utility provider to set the air conditioner's temperature up to a couple of degrees higher during peak load hours, on at most five days in a year.

We realize that practical implementation of policies may involve several complications. For example, the *peaksaver* PLUS program requires the consumers to have thermostat controllers installed. Therefore, any practical implementation of such a policy would involve ensuring that the consumers have the necessary equipment installed. To overcome such impediments, one may, for instance, target the policy only towards occupants moving into newly constructed buildings or apartments, where one can ensure that the homes come pre-equipped with the thermostats. We acknowledge the fact that due to lack of field knowledge, our policy propositions may need to undergo several modifications before they are implementable. Therefore, our policy recommendations should be viewed more like a high-level framework, as opposed to a rigid set of rules. With this in mind, we make the following policy propositions for implementing each bias:

3.1.1 Status Quo Bias

The status-quo bias motivates the occupants to not make any changes to the existing thermostat infrastructure since it is the default option. Thus, as a policy measure, we propose automatic enrollment of new households into the *peaksaver* PLUS program, unless the household actively decides to opt out. Both pre-installation and automatic enrollment take advantage of the status-quo bias.

3.1.2 Commitment Devices and Social Norms

We recommend commitment devices be incorporated in our proposed policy measures, in conjunction with the establishment of favorable social norms. As commitment devices, we would encourage participants to publicly acknowledge their participation in the *peaksaver* PLUS program. As a part of the policy, we propose that tangible avenues be introduced that facilitate public visibility of such acknowledgments. The said avenues should be designed to cultivate a feeling of pride and well-being that is associated with making a good gesture towards the society and the environment. In particular, as a publicly visible indicator of participation, we propose awarding each participating home with a *peaksaver* PLUS *Volunteer Badge*. The badge would serve two distinct purposes. First, it would make participation public, thereby enforcing commitment on behalf of the consumer.

Secondly, it would also foster an environment that promotes the adoption of the program as a social norm. Additionally, we also realize that the media plays a crucial role in forming and molding public opinion. Therefore, recommend that the policy allows for active initiatives for setting up public billboards and sending newsletters, that advertise the growing participation in the neighborhood. Proceeding on similar lines, the policy may also aim at developing and popularizing mobile and web applications, to harness the power of online social media platforms like Facebook and Twitter. Other similar efforts may include events such as holding a special reception for the participants to meet the mayor of the city. Such events should further support the cause by feeding the social urge to be a part of an elite group, and disseminating quantitative information about the number of people participating in the program.

3.1.3 Gamification

In addition to the *peaksaver* PLUS Volunteer Badge, we recommend awarding a *green score* to each household, that could be calculated based on their energy consumption profile with respect to the *peaksaver* PLUS program. We intend the green score be revised periodically, say at the end of each year, to reflect how the household performed over the past year. This evaluative number implicitly implements a game scenario where one must reduce their energy consumption to achieve better reward points (green score). To foster competitiveness, we further recommend maintaining local leader boards of green scores.

3.2 Answering the Hows and Whys of Policy Making

Prior work on behavioral economics and non-cash incentives for behavior modification motivates the use of non-cash incentives in our proposed policy. However, most prior works only provide qualitative guidelines rather than quantitative insights. Given that millions of dollars are spent in making and implementing peak-reduction policies [24,30,31,47,50], it is important that we take well-informed policy decisions.

Design and implementation of any policy has several practical challenges. For example, we need to know how effective a particular policy will be. This would allow us to make smarter decisions which help in the design process. Also, we need to understand the target population. Since everybody is different, there are no guarantees that we would be able to come up with a one-size-fits-all policy for everybody. It is obvious that we should identify different subsections of our target population and tweak the policies so that they are tailor-made for each subpopulation.

Thus, our approach is to take a fresh approach to policy design. Specifically, we advocate the use of concrete data-driven analysis before moving to the implementation phase. We proceed by building a game-theoretic model based on a large crowd-sourced survey that provides us with agent preferences. Thus, we directly elicit the population's preferences and design a game-theoretic mechanism for policy selection. We believe that this approach can be used by policy makers even beyond the limited scope studied in this thesis.

3.3 Game Theoretic Model

A game-theoretic model is used to fine-tune both the design and implementation of our policy. We use a standard principal-agent game setting, where the principal represents the electricity supplier, and the agent is the consumer. The principal introduces and implements policies to reduce the peak demand on summer afternoons when the grid is experiencing very high demand. Therefore, the principal is interested in maximizing participation in the *peaksaver* PLUS program. To do so, we next model the policies proposed in Section 3.1. Note that the number of agents who participate in the program (say λ) goes beyond a threshold λ^*, the peak load reduction goal is met.

3.3.1 Principal's Goal and Problem Formulation

The principal wants to achieve the peak demand reduction target. This depends on the number of people participating in the program, and their level of thermal discomfort due to increased temperatures (if participants are too uncomfortable, they will leave the program). Therefore, the principal's goals are two-fold. First, it needs to decide which *policy set* from the set $\mathbb{I}$ of all possible policies, i.e. the power set over all incentives, maximizes electricity savings. [1] Second, it needs to identify sections of the population that would participate, for a certain ΔT increase in thermostat temperatures (0 temperature increase for non-participating agents). The net electricity savings would, therefore, be proportional to $\lambda \Delta T$ (the monetary savings are actually non-linearly increasing with ΔT, so linear proportionality is a conservative assumption. A 2% increase in temperature has been found to reduce demand by 37% [18].).

We now formulate a general case utility model of the principal. Let $\xi_i \in \{0, 1\}$ be a variable where $\xi_i = 1$ if the principal offers an incentive $i \in \mathbb{I}$ and $\xi_i = 0$ otherwise. We additionally assume that when the principal offers some incentive $i \in \mathbb{I}$, it incurs a cost χ_i. An *incentive policy* is specified by a vector $P^{\text{in}} = (\xi_1, \ldots, \xi_{|\mathbb{I}|})$ indicating which incentives are offered by the principal. The cost associated with a particular incentive policy is $\text{cost}(P^{\text{in}}) = \sum_{i \in \mathbb{I}} \chi_i \xi_i$. A *policy*, P, is an incentive policy coupled with an expected temperature increase, ΔT, the principal wishes to implement. That is, $P = \langle P^{\text{in}}, \Delta T \rangle$.

We assume that the principal implements a policy only if a minimum threshold participation of λ^* is met. The threshold λ^* signifies a lower bound on the acceptable electricity savings determined according to the peak demand reduction requirement. Given a policy, P, let $\lambda(P)$ be the number of agents willing to participate in the program given the policy. Mathematically, $\lambda(P)$ is given by:

$$\lambda(P) = |\{j | \Pi_j(P) \geq 0\}|.$$

where $\Pi_j(P)$ represents agent j's utility under policy P. For simplicity, we assume that the agent's utility from different incentives is additive, so that for calculating $\Pi_j(P)$ we

[1] A particular policy set is identified and implemented at the beginning, i.e., the principal gets only one chance at selecting and implementing a policy in the game.

add the utility from only those incentives that are included in the policy P. We discuss this in further detail in 3.3.2.

Given these conditions, it is possible to formulate the principal's optimization problem as follows:

$$\begin{array}{ll} \max\limits_{P} & cost(\lambda(P)\Delta T) - \sum_{i\in\mathbb{I}} \chi_i \xi_i \\ \text{s.t.} & \qquad \lambda(P) \geq \lambda^* \end{array} \tag{3.1}$$

where $\text{cost}(\lambda\Delta T)$ denotes the monetary value of the electricity savings. Here, a uniform ΔT temperature increase is used for agents whose utility stays non-negative after accounting for discomfort due to the increase, and a zero temperature increase for all agents whose utility becomes negative upon imposing a ΔT increase in thermostat temperature.

3.3.2 Agent Model

Let there be n agents in the system. Each agent i has a private type θ_i that determines its utility for different aspects of the policy, and let the type space be given by Θ. For agent i, let S_i, C_i, N_i and G_i denote the utility due to status quo bias, commitment devices, social norms, and gamification respectively; let $d_i(\Delta T)$ represent the absolute utility value of its thermal discomfort under a ΔT increase signal from the principal; and let $\Pi_i(P)$ be the net utility derived by the agent from the biases and the discomfort under policy P.

Let $\mathcal{A} = \{0, 1\}$ denote the action set, where 1 and 0 denote if the agent chooses to participate or not participate, respectively. Let $a_i \in \mathcal{A}$ denote the action taken by agent i. We assume that agents are individually rational and thus will participate only when their utility from participating is greater than or equal to zero.

Let $\mathcal{S} = (S_1, \ldots, S_n), \mathcal{C} = (C_1, \ldots, C_n), \mathcal{N} = (N_1, \ldots, N_n), \mathcal{G} = (G_1, \ldots, G_n), \mathcal{D}(\Delta T) = (d_1(\Delta T), \ldots, d_n(\Delta T))$, and $\mathcal{Z} = (a_1, \ldots, a_n), \quad \forall a_i \in \mathcal{A}, i \in \{1, 2, \ldots, n\}$ where $\mathcal{Z}$ represents the action profile.

Then we define the following function mappings

$$\mathbb{A} : \mathcal{S} \times \mathcal{C} \times \mathcal{N} \times \mathcal{G} \times \mathcal{D} \rightarrow \mathcal{Z}$$

where $\mathbb{A}$ maps from the agents' utility distributions to the action profile. The action profile is also the outcome of the game, as it represents the agents' participation distribution.

We now consider modeling Π_i, the net utility derived by an agent i. In [37], Keeney demonstrates how assessing a multiattribute utility function is more of an art than science; and coincidentally, deals with formulating a utility function for an energy policy. Many prominent psychologists and economists like Meehl, Kahneman, Dawes et al. in their respective works [25, 34, 46], have argued, firstly, in favor of using precise mathematical algorithms instead of intuition arising from expertise, and secondly, how even a simple and naive model works efficiently as a composite predictor. Therefore, for evaluating the net utility obtained from combining utility from different biases, we use an additive linear utility model [2]. At this point, we still need to model the weights that appropriately scale the utilities so that they can be combined linearly.

Mathematically, we calculate the net utility for agent i as follows:

$$\Pi_i(P) = \xi_S w_i^S S_i + \xi_C w_i^C C_i + \xi_N w_i^N N_i + \xi_N w_i^N N_i - d_i(\Delta T)$$

where w_i^j represents agent i's weight for bias $j \in \{S, C, N, G\}$, and $(\xi_S, \xi_C, \xi_N, \xi_G)$ is P^{in}.

For succinctness in presentation, we represent $\Pi_i(P)$ as follows:

$$\Pi_i(P) = (P^{\text{in}} \circ (w_i^S, w_i^C, w_i^G, w_i^N))(S_i, C_i, N_i, G_i)^\intercal - d_i(\Delta T)$$

[2]We assume that the net utility Π depends only on the marginal distributions of the utility $\mathcal{S}$, $\mathcal{C}$, $\mathcal{N}$, and $\mathcal{G}$,derived from the biases, and not on their joint distributions [28]. Again, due to the mutual independence of utilities from different biases, without loss of generality, our utility model can also be assumed to be linear.

where $\circ$ denotes the entry-wise product or Hadamard product operation.

It is noteworthy that even though participants of the *peaksaver* PLUS program receive monetary benefits due to reduced consumption, it is inappropriate to include it in the policy and the utility functions, as it may counter-act the utility received from behavioral policies. This is because framing is important part of behavioral policy design, and there is evidence that bringing money into the picture may change people's mindset [43].

3.3.3 Agent Participation Criteria

For an agent to participate in the programs of our policy, the utility of the agent under participation should be greater than or equal to the utility of the agent under non-participation. Consider an agent who is willing to participate if the maximum temperature by which the thermostat is allowed to increase is ΔT_{max}. Clearly, the agent's utility is 0 at ΔT_{max}. If the temperature increase is less than ΔT_{max}, the utility for the agent under participation is always non-negative. Hence, it is axiomatically a dominant strategy for the agent to participate. We have three core assumptions that allow us to make this argument. These are, for all agents:

- If the utility derived is zero, the agent's action is to conform with the policy. That is, agents prefer weakly dominant strategies that lead to a desirable outcome, to those that lead to undesirable outcomes (from the point of view of the principal).
- Utility derived from non-cash incentives is non-negative
- Utility derived from non-cash incentives under non-participation is zero.

3.4 Measuring Utility of the Agent

Much of the agent's utility model that we presented in the previous section depends on the type of the agent. We shall discuss more about agent types and how we determine them in the following chapters. However, for calculating discomfort, we use a modification of Fanger's Predicted Mean Vote (PMV) model [27]. Being able to evaluate the value of $\mathcal{D}(\Delta T)$, would enable us to calculate the net negative utility that the principal needs to offset using incentives, to be able to meet its target energy saving goal.

3.4.1 Utility from Thermal Discomfort

We use Fanger's equations for computing the PMV of a human occupant in the house. According to Fanger's equations, the PMV depends on several factors like air temperature, mean radiant temperature, air speed, humidity, metabolic rate of the person, and their clothing level. Since we need to measure the comfort levels of humans in closed houses that are equipped with thermostats, we make certain reasonable assumptions that would be very likely to hold in practice. We assume the humidity level to be at 50%, the mean radiant temperature to be equal to the air temperature, the airspeed to be zero, a metabolic rate of $1.2met$, and clothing level to be $0.5clo$ that represents typical summer indoor clothing. Using these values, we find the varying levels of discomfort experienced by a person.

The Fanger's equations take six different parameters as input and render the PMV as output. This makes analysis of the equation difficult. Since in our case, we deal with only a particular set of values of these parameters, we find the PMV values obtained for a range of temperatures with these parameters, and fit a simple linear equation to the PMV values using regression. For the purpose of regression, we obtain the data points from [4]. We find that the PMV values can be accurately represented by the following equation, with a Mean Square error of 0.02%.

$$PMV(t) = (0.306)|t - 24.6| - 0.48$$

where t represents temperature in Celsius. Now, according to ASHRAE's Standard 55, humans feel comfortable in the PMV range of $[-0.5, 0.5]$. Therefore, we may calculate the absolute value of the utility derived from discomfort by the following equation.

$$d(|t-24.6|) = \begin{cases} 0 & \text{if } |t-24.6| \leq 1.6 \\ PMV(t) & \text{if } |t-24.6| > 1.6 \end{cases}$$

In the above definitions of discomfort and PMV, it is assumed that the average human is comfortable at 24.6 degrees Celsius. In our model, we assume that the agent is comfortable at the temperature t^* of the household (set by the agent), whatever that temperature is. So, for calculation of both discomfort and PMV, we replace 24.6 by t^*. With this modification, we define discomfort of agent i as follows:

$$d_i(|T_{set}-t^*|) = \begin{cases} 0 & \text{if } |T_{set}-t^*| \leq 1.6 \\ PMV(T_{set}) & \text{if } |T_{set}-t^*| > 1.6 \end{cases}$$

where T_{set} represents the higher temperature set by the mechanism administrator during peak load hours. Now, $|T_{set}-t^*|$ is equal to the ΔT increase in the thermostat, as signaled by the principal. Therefore, substituting ΔT for $|T_{set}-t^*|$, we obtain the following equation for discomfort:

$$d_i(\Delta T) = \begin{cases} 0 & \text{if } \Delta T \leq 1.6 \\ (0.306)(\Delta T) - 0.48 & \text{if } \Delta T > 1.6 \end{cases}$$

3.4.2 Utility from Biases

For measuring utility from biases, we first determine the underlying agent types. To that end, we conduct a survey that elicits behavioral preferences. We will discuss more about the design, execution, and interpretation of the survey in the coming chapters.

Chapter 4

Survey for Agent Types

As discussed earlier, the general area of behavioral economics is making its presence felt in several disciplines. However, the majority of the research contributions to this field are from psychologists and behavioral economists. Contributions are mostly derived from meticulously crafted experimental routines that can be rigid in their design. The conditions for these experiments are tightly controlled. This makes it difficult to adapt these techniques for other purposes. For example, the Self-control scale [60] measures a person's ability to resist temptations, but it is hard to generalize scores on this scale to derive information about how much the person values commitment devices in the context of thermal comfort. Given the potential gain from behavioral economics in areas of research, it would be useful to derive guidelines that enable researchers from other areas to conduct similar behavioral studies.

One way to conduct a behavioral study is by using a psychometric survey. The primary aim of the survey questionnaire is to objectively quantify aspects of human behavior. We specifically choose surveys to be our preferred means of collecting data, as it is cheaper than conducting elaborate experiments. Surveys are also easy to conduct using online platforms. In this chapter, we present guidelines and procedures we followed in conducting a survey for eliciting preferences towards cognitive biases in energy policies, which may prove beneficial for researchers. We also provide details about the design elements and structure of our survey.

4.1 Designing and Conducting a Behavioral Survey

Compared to other means of data collection, surveys are inexpensive, highly scalable, and help obtain responses from a potentially diverse set of audiences. These advantages are further enhanced by online platforms for creating and conducting surveys. However, designing a psychometric survey has some subtle issues. For example, to prevent bias in survey responses, individuals from the target population should interpret the questions objectively. Such subtleties may demand several iterations before the questionnaire is ready to be deployed. In this section, we will discuss our survey design and execution processes in detail.

4.1.1 Survey Structure

A well-thought-out survey structure is essential for ensuring data quality. For example, all questions pertaining to a particular behavioral phenomena should be grouped together, thereby forming clearly demarcated sections. This prevents confusion, makes the survey more readable, and also helps in organization and analysis of responses.

4.1.2 Framing Questions

Recall that the primary aim of the survey questionnaire is to objectively quantify aspects of human behavior. As human behavior is intrinsically qualitative in nature, several precautions should be taken while framing questions. We recommend the following rules of thumb and best practices [12, 13, 32]:

1. Questions should articulate precise and objective situations or provide concrete examples. This reduces errors arising out of subjective interpretation by the respondents.

An example of a question to avoid is: *How often do you fulfill your commitments?* An example of a better question is: *Suppose you and your friend agree to go to the gym together, a few times a week. How likely are you to do so?*

2. It is a good practice to ask a number of specific questions that all aim to gauge the same behavioral aspect. For example, in place of a single general question like, *How often do you fulfill your commitments?*, it is better to ask a set of specific questions like the following:

   - *Suppose you and your friend agree to go to the gym together, a few times a week. How likely are you to?*
   - *Suppose you promised yourself that you would read more books. How likely are you to follow through?*
   - *Suppose you are on a diet. How likely are you to successfully resist that extra piece of cake?*

3. Questions should not cite multiple examples. Doing so may introduce biases and potential anomalies in the responses due to possible miscommunication or misinterpretation. For example, the question, *How often do you fulfill your commitments like going to the gym with a friend, or sticking to a diet?*, is undesirable because it cites multiple examples. Instead, it is better to use a question with a single example, such as, *How often do you fulfill your commitment to go the gym?*

4. Questions should not be complicated. An example of a poorly constructed question is *Your recently inherited money is in a moderate risk mutual fund. Assuming that you want to keep the money invested somewhere, and you have other real-world options like directly investing in stocks, purchasing treasury bills or bonds, investing in another mutual fund, investing as a venture capitalist, etc. How likely are you to cash out the money and invest it somewhere else.* An example of a better question is, *Your recently inherited money is in a moderate risk investment. Assuming that you wish to keep the money invested, how likely is it that you would change how the money is invested*

5. Clarify wording for obtaining unbiased or incorrect responses. For example, consider the following questions:

    - *How much higher are you willing to adjust your thermostat than your usual, if you are provided incentive A?*
    - *How much higher are you willing to adjust your thermostat than your usual, if we are provided incentive B?*
    - *How much higher are you willing to adjust your thermostat than your usual, if we are provided both incentive A and B?*

    In this case, suppose it is implicitly assumed that incentives A and B together have a greater impact than either of them alone. Then, to avoid incorrect responses to the third question, it is better to explicitly clarify by adding a statement – *The answer to the third question should either be the maximum value of the answers to the previous two questions, or more than that.*

6. Group related questions together under separate pages for good structure, and randomize question order within each page to remove unforeseen biases.

7. Having a dual sample, i.e. collecting data from different samples of the population, helps in comparison and further validation of results. Alternatively, this can be enforced by making *splits* in the data set, and *cross-validating* the results.

8. Consistently use a single scale throughout the survey. Using different scales within the same survey might be confusing for participants, and may introduce anomalies in their responses. The five-point Likert Scale [44] is a good choice for psychometric surveys.

9. Avoid *leading* questions, i.e. questions that are suggestive in nature and may influence the participant's response. A common sanity-check is to take a step back and observe if the question could be interpreted in a different way by people from different backgrounds. This can also be caught in a pilot survey. An example of a leading question – *Do you think your electricity rates are high?* This question suggests that

electricity prices are high. A more neutral question would be – *Do you think your electricity rates are appropriate?*

10. Multiple-choice questions eliciting factual information should present all possible choices. In such cases, adding a *Cannot answer* option or providing an additional text field for explanation are both good practices.

    For example, consider the question *On an average, how much is your monthly electricity bill?* Let us assume this question has two answer choices – $25 or less, and more than $25. However, it should be noted that the options are not exhaustive. What if the participant does not have this information? Not having the right option may drive the participant to choose an answer that may not be accurate. In such cases, it is a good practice to include an additional *Cannot answer* option.

11. To ensure good quality of responses, it is often wise to include trap questions and attention checks like, *To what extent do you agree with the statement Five dollars plus fifteen dollars is equal to three dollars.* Similarly, recording response time to judge quality can also help, i.e., if a question is answered too quickly, it may be incorrectly answered. We also recommend IP-filtering to remove duplicate answers, since two answers from the same IP address are very likely to be from the same person.

12. If certain questions elicit sensitive information, it is often a good idea to *load* the question. Loading refers to creating a comfortable circumstance and building a certain level of trust, so that the respondent feels comfortable to provide the *real* answer and not just a *socially desirable* answer. For example, a participant may feel uncomfortable answering to – *Have you ever made an inappropriate comment in a social situation that you found was embarrassing?* A better way to frame such a question would be by making the participant feel comfortable about disclosing the information – *It has been observed that most people have had several experiences of embarrassing themselves in public. Have you ever made an inappropriate comment in a social situation that you found was embarrassing?*

13. Add a question towards the end of the survey that asks the respondent if they feel they had appropriately answered all the questions. Here, it is also imperative to

state that they will not be denied payment if they chose that they did not answer questions adequately.

14. Add a seriousness check by stating that good research findings depend on good quality data. For example, one could add a disclaimer at the beginning of the survey that states, *Only datasets from participants with a motivation for serious participation will be analyzed.* Using University or research organization credibility also urges respondents to answer questions seriously.

15. Research findings suggest that survey respondents' focus level is higher in the beginning and lower towards the end of the survey [20]. Therefore, placing more cognitively demanding questions early on is a good idea. That said, placing harder questions first, may decrease survey participation since most dropouts take place during initial parts of the survey.

16. Place sensitive questions towards the end of the survey. The respondent can be assumed to have developed a certain level of comfort with the questions towards the end of the survey, and is, therefore, more likely to answer them without inhibitions.

17. It may be beneficial to mention how much time the survey would take to complete. In fact, inflating this figure by say 3-4 minutes is also a good idea. This sets the expectations correctly, builds trust, and weeds out non-serious respondents. Also, conduct an objective assessment of the estimated time (using a pilot survey, for example), instead of quoting an approximate length of time that's purely based on judgment.

18. While designing the questions, it is always a good idea to have your peers or colleagues review them. Seeking opinions about confusing questions, ambiguous wording, etc., helps in improving the quality of responses.

19. When the survey questionnaire is ready, we recommend conducting a pilot survey before the actual survey. One way of obtaining feedback on a pilot survey is to have an additional option that respondents can click for each question, if they feel the question is not clear to them.

20. Before collecting responses, it is vital to sketch an outline of the steps involved in the interpretation and analysis of the survey. This helps in avoiding potential unforeseen circumstances and ensures that the survey questionnaire meets the needs of the study. A mock analysis of responses from the pilot survey often helps. For example, one may miss out on including a question on a particular topic, and realize that they need the information only after all the survey responses are in. Conducting a preliminary analysis helps in such cases.

21. As a part of the survey's post processing, it is better to drop *flat-liners* – responses with very low variance in their answers.

22. Opt-in panels, i.e. panels where the respondents willingly choose to participate in the survey, should not report sampling error. There is no theoretical basis for calculating sampling error since participation is not random.

23. Demographic information, if required, should be sought at the end of the survey. Some examples of basic demographic information that can be sought are age, gender, ethnicity, and educational qualification.

4.1.3 Research Ethics

It is the responsibility of the designer of the survey to abide by ethics regulations set by a government body or the concerned organization that the designer represents. For example, when collecting data, appropriate means must be adopted to maintain data anonymity, confidentiality, and privacy. Some organizations may impose approval prerequisites from an Ethics body, before conducting the study. Additionally, participants should receive a fair remuneration for their time.

Our survey had been reviewed by, and received ethics clearance through, a University of Waterloo Research Ethics Committee, before we conducted the survey. More information about University of Waterloo's research ethics policies can be found at [8, 9].

4.1.4 Conducting the Survey

There are several different ways to execute a survey and collect responses. The avenues depend on available funds and goals.

Completely outsourcing the survey to third-party panel firms is an option, but may be quite expensive. In this case, the third party would be responsible for recruiting participants, seeking responses, drafting the results, and even analyzing the results. The recruitment process could be carried out either in person, via telephone calls, or via emails and other electronic media.

Another approach would be to procure a dataset containing the concerned localities' consumer information such as name, address, telephone number, etc. Several websites provide business to business (B2B) and business to consumer (B2C) datasets for a price [5]. These datasets may only have the name, address and telephone numbers of the consumers. Small businesses pay and acquire such datasets for targeted marketing. With such datasets, one would need to call people from the list, and invite them to participate in the survey. This makes the process very cumbersome to execute. A natural alternative is to procure email-ids and request participation via emails. However, it is worth noting that Canadian anti-spam laws in effect since July 1, 2014, prevent the sale of datasets that contain email-ids [1, 2]. In fact, this law applies to any commercial electronic communication, including email, SMS, social media or instant messaging. More specifically, the law also prevents electronic communication for persuasion to participate in surveys. For these reasons, direct communication may not be the best way to recruit participants.

Online surveys can be conducted using websites like SurveyMonkey or Qualtrics. These websites provide simple and streamlined platforms for hosting surveys, collecting responses, and organizing the collected data. Such websites also provide additional facilities like online survey panels, but these are expensive. An economical alternative is to use crowdsourcing platforms like Amazon Mechanical Turk or CrowdFlower to recruit participants. One major criticism about the use of crowdsourcing platforms is that the sample of participants, having gone through an opt-in process, cannot be regarded as a statistically randomized sample of the population. Nevertheless, they are an economical and attractive option.

In our case, we hosted our survey using SurveyMonkey and obtained responses through CrowdFlower. We opted for these platforms because of their ease of use, scalability, and cost-effectiveness. Our target audience consisted of people living in the United States and Canada. A map showing locations of survey respondents can be seen in Figure 4.1.

4.1.5 Executing Jobs on CrowdFlower

Creating and executing jobs on CrowdFlower is user-friendly for the most part. One can design and launch jobs fairly easily, and monitor the progress of jobs on its dashboard. The only part of executing a job on CrowdFlower that may be challenging, is to make sure that the remuneration is paid fairly, and without any security loopholes, as provisioning of remuneration codes is not handled by CrowdFlower. When contributors accepted our job on CrowdFlower, they were linked to our survey on SurveyMonkey. Upon completion of the survey, we asked the contributor to access a webpage hosted our server (blizzard.cs.uwaterloo.ca), where they obtained a remuneration code for CrowdFlower. CrowdFlower provides a client-side interface in the form of a text-box portal, for writing custom JavaScript. We used this interface to configure remuneration codes. To generate unique valid codes for every contributor, we maintained a database of codes on our server. For every new contributor trying to use a remuneration code, the custom JavaScript queried our server and verified the validity of the code. In the database on our server, we also maintained the contributor's IP address, and a flag variable to check if the code had been used before. These steps ensured that the code worked for every contributor, and that the code worked exactly once. The JavaScript code can be found in Appendix C.

4.1.6 Statistical Know-How

For appropriate design and interpretation of the survey, it is important to have knowledge of basic statistical concepts like experiment design, statistical sampling, hypothesis testing

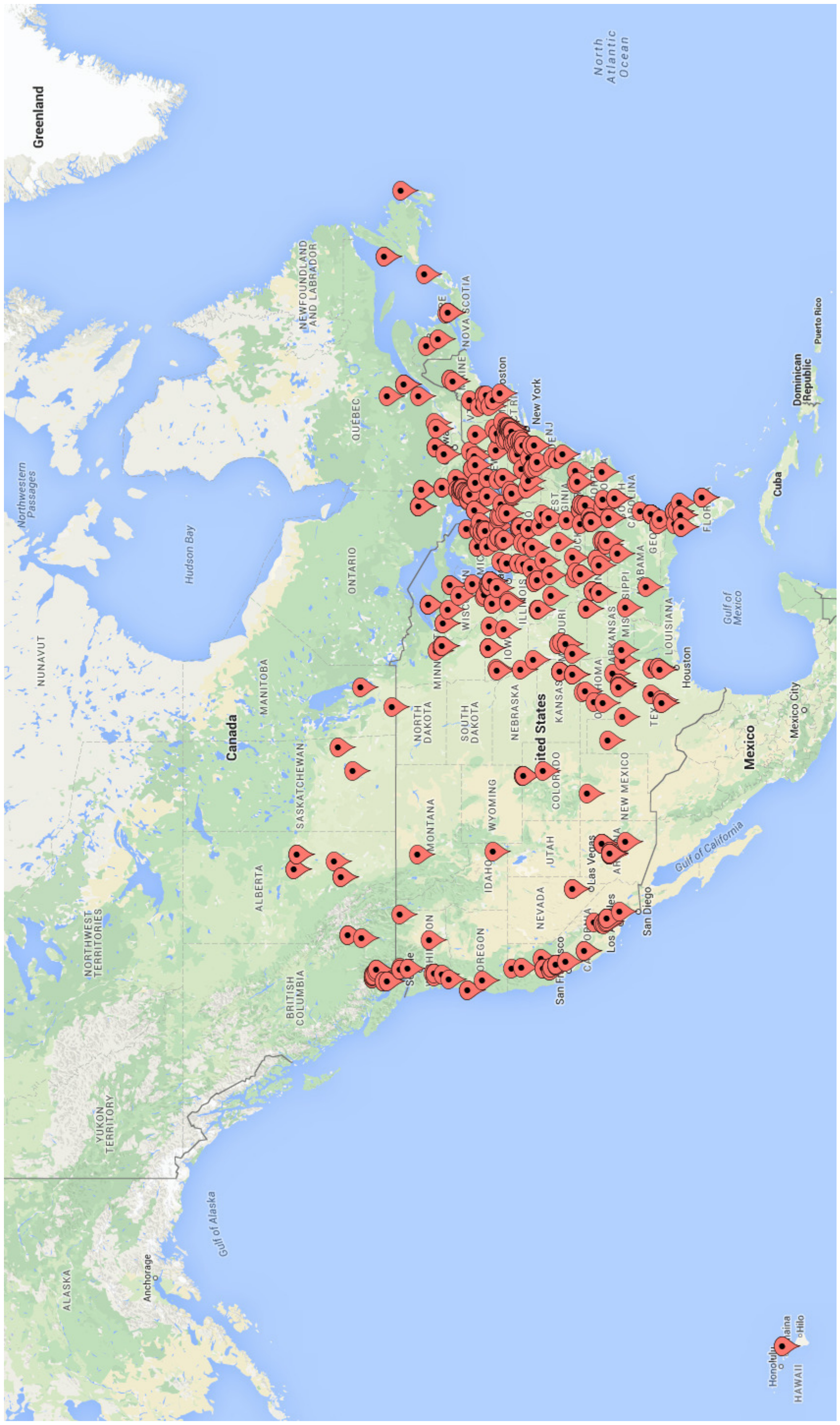


Figure 4.1: Map showing locations of survey participants

and confidence intervals, etc. If lacking an adequate knowledge of statistics, we strongly recommend consulting a statistician before conducting the survey.

For our survey, we used the services of Statistical Consulting Service at the University of Waterloo, wherein they reviewed and validated that our survey.

4.1.7 Appropriate Sample Size for Survey

The population of the United States and Canada is roughly 332.5 million [3,10]. Given this population size, if we had a truely random sample, we would have aimed to have a sample size of 385, to have the worst case confidence interval (error bars) of 5% for any observed distribution of responses to a question, at 95% level of significance [11]. However, since our sample is not truly statistically randomized, the numbers we provide here should only be considered as indicative of the inferences we may be able to derive from a well-randomized sample.

We obtained a total of 990 responses to our survey, and after data-cleaning were left with 425 responses. Our inferences are based on these 425 responses. We will discuss more about survey responses in the next chapter.

4.1.8 Major Components of Our Survey

Our entire survey can be found in Appendix A. The purpose of our survey is to extract preferences about the four biases (status quo bias, social norms, commitment bias, and gamification), temperature sensitivity and thermostat behavior, and the impact of the biases on the thermostat settings. We have eight major sections within the survey that measure these attributes. Within the survey, each of these sections start on a new page.

The first section has qualifier questions. Based on responses to these questions, we decide if the response belongs to our target population.

The next four sections of the survey elicit preferences regarding commitment devices, status quo bias, social norms, and gamification respectively. These questions aim at finding the degree to which respondents might be predisposed to the biases in general, rather than specific circumstances.

The sixth section inquires about house occupancy during peak load hours in the summer. We broadly target the two scenarios where people's houses may systematically and predictably be unoccupied during the peak load hours. The two scenarios being: either the occupants are out on a vacation for a long stretch of time, or the occupants are consistently away from home during the said hours due to external commitments such as work. We hypothesize that this information may be suggestive of the 'unreliability' of a particular household during the summer season. By unreliability, we mean that the household may be unoccupied and hence we may or may not be able to change the temperature for these households depending on whether the thermostat is left working or switched off [1].

The seventh section asks the respondents about their usual thermostat settings during peak load times in summer, and to what extent the occupants are sensitive to temperature [2].

The eighth and final section inquires about the extent (in terms of temperature in degree Celsius) by which the respondents would be willing to set their temperature higher. In particular, we seek this information under two cases. First, while considering the incentive from each bias individually, and second, while considering the net incentive when all the biases are implemented. These responses enable us to find the proportion of contribution of each of the biases towards their net utility. We will discuss about utility from the biases in further detail in Chapter 5.

[1] To avoid uncertainty, we later decided to exclude information from our model

[2] As a part of modeling choice, we later decided not to use this information. Instead, we made a modeling assumption, according to which we only needed the change in thermostat temperature and not the absolute value of the temperature; as discussed in section 5.4.3.

4.1.9 Scoring

For all the questions in the survey that measure utility from the biases or the discomfort, we use a five-point Likert scale [44]. For evaluating scores obtained from the survey responses, we use an adjusted scale with unequal distances between the Likert items, as opposed to a uniformly linear scale [39]. In particular, we use unequal distances with the scoring set $\{0, 0.2, 2, 3.8, 4\}$, to overcome the *middle-of-scale* effect, which states that using verbal anchors at all points on the Likert scale creates "a larger perceived distance between points in the middle of the scale" [39]. [3]

4.2 Survey Takers' Happiness

To the best of our ability, we ensured that the respondents were happy with the survey taking experience. To this end, we offered competitive remuneration for the task, and prompt email support for any respondents having trouble with receiving their remuneration.

4.2.1 Deciding Participant Remuneration

We decided to award a remuneration of one United States Dollar (USD) per participant, based on data samples collected from Amazon Mturk.

- For a sample of 11 "research" surveys, a plot of Frequency v/s Price in USD is shown in Figure 4.2.
- For a sample of 52 surveys of all kinds, a plot of Frequency v/s Price in USD is shown in Figure 4.3.

For both considered samples, we observed the following additional criteria:

[3] It should be noted that during our initial exploratory analysis of the survey data, we use a linear Likert scale, to get a clear understanding of the nature of the data. Later, during our final analysis, we use the curved Likert scale for scoring.

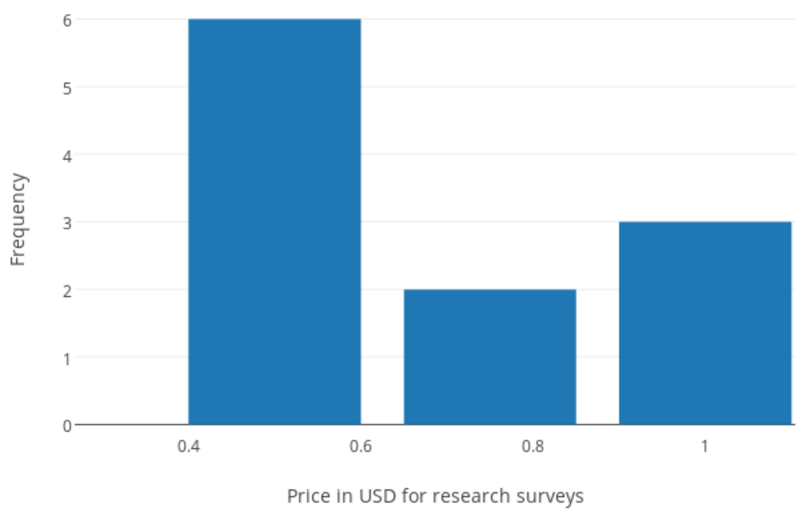


Figure 4.2: Frequency distribution of how much research surveys pay on Amazon MTurk

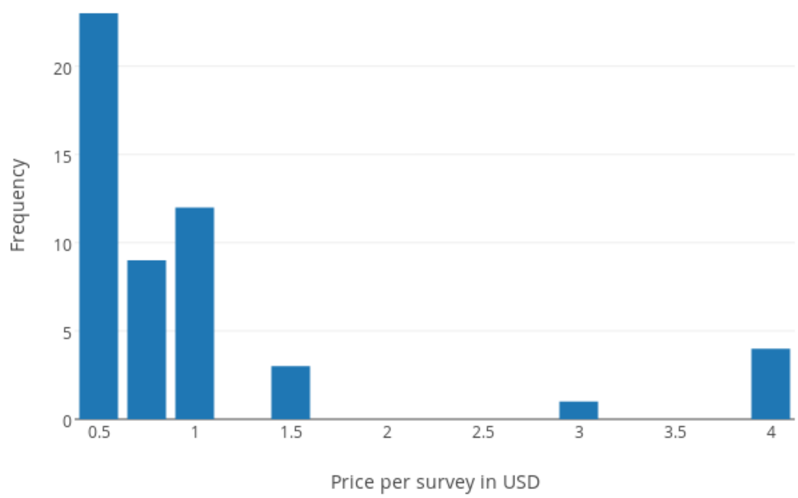


Figure 4.3: Frequency distribution of how much surveys, in general, pay on Amazon MTurk

Figure 4.4: Contributors' feedback rating on CrowdFlower

- Minimum cutoff payment was \$0.5
- Only respondents who had completed at least 1000 tasks were considered. This is a standard qualifier for contributors on Amazon MTurk.
- All tasks considered were less than 30 minutes long.

4.2.2 Respondents' Feedback Rating

From among the 990 participants of the survey task, 74 participants rated our survey task on clarity of instructions, fairness of test questions, ease of job, and remuneration. A screenshot of the average ratings from these 74 participants can be seen in Figure 4.4.

Chapter 5

Survey Interpretation, Data Analysis and Results

In this chapter, we analyze and interpret the data from our survey. In particular, we seek to identify agent types based on their preferences. We also use data from the survey to calculate our model's parameters, and to calibrate weights for combining utility from the biases and discomfort. Ultimately, we solve the principal's optimization problem of maximizing energy savings under uniform and non-uniform signals sent to the agents.

5.1 Data Cleaning

Conducting a survey on a crowd sourced platform like CrowdFlower has its trade-offs. To our advantage, it allows us to conduct our survey in a large geographical region and quickly obtain numerous responses. However, to our disadvantage, several factors lead to anomalous *dirty* responses in our survey data. We discuss these next.

- Remuneration provided on crowd sourcing platforms is generally low. We pay one US dollar for completing our survey. While this remuneration is competitive, and in fact, better than the average, workers are still underpaid compared to a real world job. Therefore, it is expected that some workers may not answer the survey seriously.

- Moreover, due to restrictions enforced by Research Ethics, we allow the respondent to not answer questions and skip to the end of the survey.

Since we cannot rely on the quality of data we obtain, we employ a set of data cleaning rules to eliminate untrustworthy responses, before data analysis. We carry out the following steps for data cleaning. The number of responses removed in each step, from a total of 990 responses, is shown in parenthesis.

1. Removed responses from surveys that were quit midway (9 responses removed).

2. Removed responses that were completed but had at least one question blank (101 responses removed).

3. Removed any duplicate responses from the same IP address (65 responses removed).

4. Removed responses from surveys that asserted that they do not have thermostats (44 responses removed).

5. Removed responses from surveys that asserted that they do not have air conditioners (89 responses removed).

6. Removed responses that failed to answer the trap questions correctly. Responses that answered the trap questions with the middle option on the Likert scale were considered to have failed. For example, a response of 'neither agree nor disagree' to the question "Is five dollars plus fifteen dollars equal to three dollars?" would be disqualified.

   One of the questions asked if the hottest day of summer is expected to be colder than the coldest day of winter. The other question asserted the converse of this question. Because of the confusing nature of these questions, we deleted only those responses that had failed to answer both these questions correctly (149 responses removed).

7. Removed responses that did not answer questions 35 and 36 as requested (98 responses removed).

8. Removed *flat–liners*, that is, responses with very low variance (< 1.0) across all questions seeking responses on the linear Likert scale (6 responses removed).

9. Those responses that said in Q38 that they did not answer attentively and would not want us to use their data (4 responses removed).

After the data cleaning process, we were left with 425 responses – we believe a sufficient sample for drawing inferences.

5.2 Pilot Survey Statistics

We ran a pilot survey on 114 respondents before we conducting the full version of the survey. The primary aims of the pilot survey were to check for anomalies in interpretation of the questionnaire, and to obtain preliminary insight into the results of the survey. We applied the data cleaning process, as explained in the previous section, and looked for anomalies in the data. We observed the following from the pilot survey:

- About 30% of respondents had answered Q36 incorrectly. The question inquires about the net temperature by which the respondent is willing to set their thermostat higher, given all four incentives from Q35 are implemented. Since we assume that the utility from the biases are non-negative and additive, the answer to Q36 should be at least the maximum value reported in Q35. This essentially rendered these 30% of responses unusable according to our data cleaning standards. This lead us to change the language of Q36 and explicitly state our requirements in the question.

  In particular, we replaced the statements "To what extent would you be willing to volunteer? (Hint: Should not be less than the maximum value you chose in Q35, since all programs are implemented at once)," with "For this question, your answer should either be the maximum value of the four responses you gave in the last question, or more than that. Keeping that in mind, in total, how much higher are you willing to adjust your thermostat than your usual (during summer afternoons)?"

With this modification, we did not observe this data anomaly in our final survey version.

- After applying the data cleaning process, we obtained about 64% of usable responses.

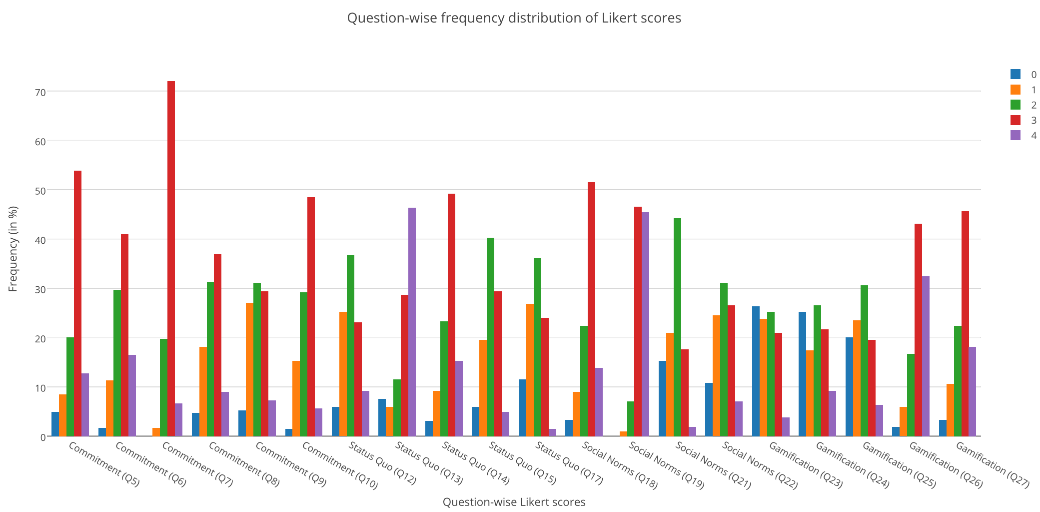
5.3 Exploring the Data

In this section, we conduct detailed exploratory analysis on the data that we obtained from the full survey, after cleaning.

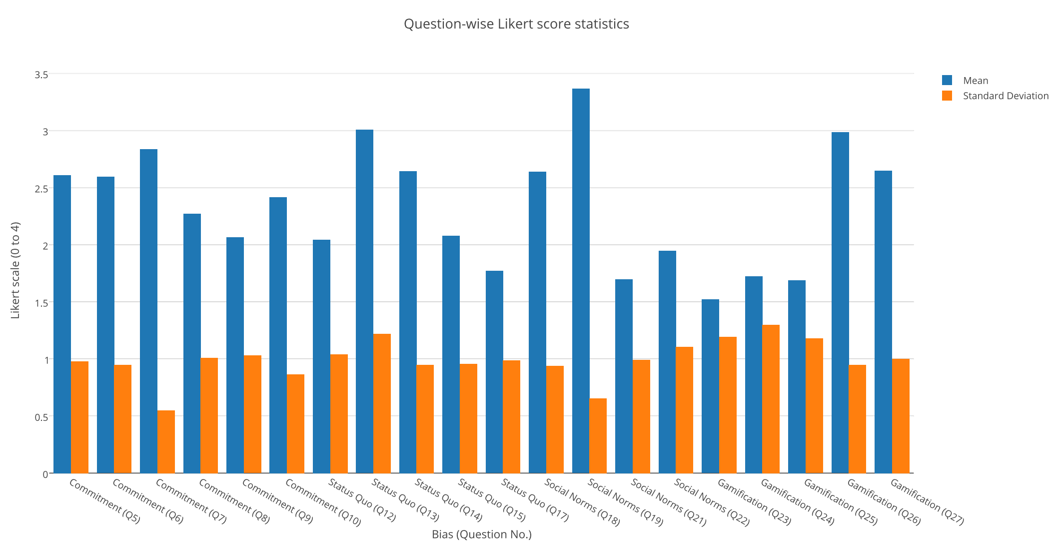
5.3.1 Full Survey Data Statistics - a Preview

We first explore the nature of the data through some basic plots. Here, we use values on the linear Likert scale $\{0, 1, 2, 3, 4\}$, where 0 corresponds to the least bias, and 4 represents the highest bias on the scale. Figure 5.1a shows the distribution of responses (on Likert scale) across the population for every question that elicits generic preferences towards biases. Note that this plot may not reveal much information about agent types, though. Agent types would depend on collective analysis across individual responses and not individual questions. It merely gives us some insight into how effective a bias is in general.

Figure 5.1b shows the mean and standard deviations of responses of each question across the population. Computed values are based on the linear Likert scale of $\{0, 1, 2, 3, 4\}$. Since the response distributions are unimodal and show central tendency, a high mean and low standard deviation indicates that the bias may have strong influence on people, in general. A low mean and high standard deviation, as seen for some questions pertaining to gamification and social norms, shows that the bias in question may not have a strong influence on the general population.



(a)



(b)

Figure 5.1: Statistics of responses for every question eliciting non-cash incentive preferences

5.3.2 Preliminary Screening for Clusters

The purpose of clustering is to identify a small number of agent types. Identifying agent types from the data is inherently hard because of its dimensionality. If we had only one question for each bias, it would be easy to identify types. Based on the Likert scores of the response, we could classify them into a discrete number of types. However, we cannot draw inferences based on one question per bias, as this would introduce unpredictability into our psychometric scale. Now, given the fact that we have multiple questions eliciting information about the same bias, a particular response could have inconsistent Likert scores on these questions. Indeed, we have such observations in our data. This poses challenges for identification of agent types. However, the presence of clusters in the data would be a sufficient condition for asserting that those clusters correspond to the underlying agent types. This motivates us to perform cluster analysis on the data.

We performed a centroid-based k-means cluster analysis on the *raw data* from the survey that included unprocessed data from questions eliciting bias information. However, the raw data did not yield any clusters. So, we conducted Exploratory Factor Analysis and Principal Component Analysis (PCA) on the raw data, and reran the cluster analysis on the factors and principal components respectively [1]. We still did not observe clusters.

5.3.3 Cluster Analysis Revisited

As a means of reducing the dimensionality, we use the aggregate score (section 5.4.2) for each bias, thereby reducing the data representing the four biases to a four dimensional space. We analyze this data for centroid-based k-means clusters. We observe standard analysis practices like calculating the within-group-sum-of-squares (WGSS) [38] and Calinski Harabasz index (ch-index) [22], to determine the ideal number of clusters in the dataset. Our results still do not indicate the presence of any prominently distinguishable clusters. Therefore, we do not have a reasonably small number of discrete agent types. Figure 5.2 and Figure 5.3 show the plots of WGSS and ch-index for commitment bias, respectively.

[1] Although, as a side note, we did not particularly prefer conducting factor analysis or PCA, as this would result in loss of original variables during the analysis.

The WGSS plot does not have a "knee" and the ch-index plot does not have clear "peak". Results for other biases are similar.

5.4 Data Interpretation and Analysis Procedures

In this section we discuss how we utilized data from the survey to directly calibrate our game theoretic model. With the calibrated model in place, we deduce inferences about optimality of policy sets, possible actions for the principal, and their corresponding outcomes.

5.4.1 Post hoc Survey Questionnaire Cleaning

Due to the lack of clusters in the data set, we decided to post hoc remove some questions from the survey. These are questions whose answers are inconsistent with other answers relating to the same bias. Arguably, these are due to confusion on the part of the survey respondents. For selecting the questions from the questionnaire that best represent the underlying psychological bias, we carry out an optimization process based on the correlation matrix obtained for questions relating to each bias. This selection process is based on the assumption that all questions that elicit a general measure of predisposition of a person towards a particular bias, should have well-correlated responses. If the data from a particular question is not well-correlated, we argue that this question is not representative of the concerned bias and should not be part of the survey. Since we could not have predicted which questions are best suited for eliciting preferences of this nature before conducting the survey, we employ this post hoc selection process.

The correlation matrix provides the pairwise correlation of responses to questions for each bias (see Figures 5.4 5.5 5.6 5.7). To find the *best correlated* subset of questions, we enumerate all possible question combinations and use a normalized measure of correlation for each subset – we call this Within-Group Average Correlation (WGAC). We calculate WGAC on a set of random variables as the sum of their pairwise correlations, divided by

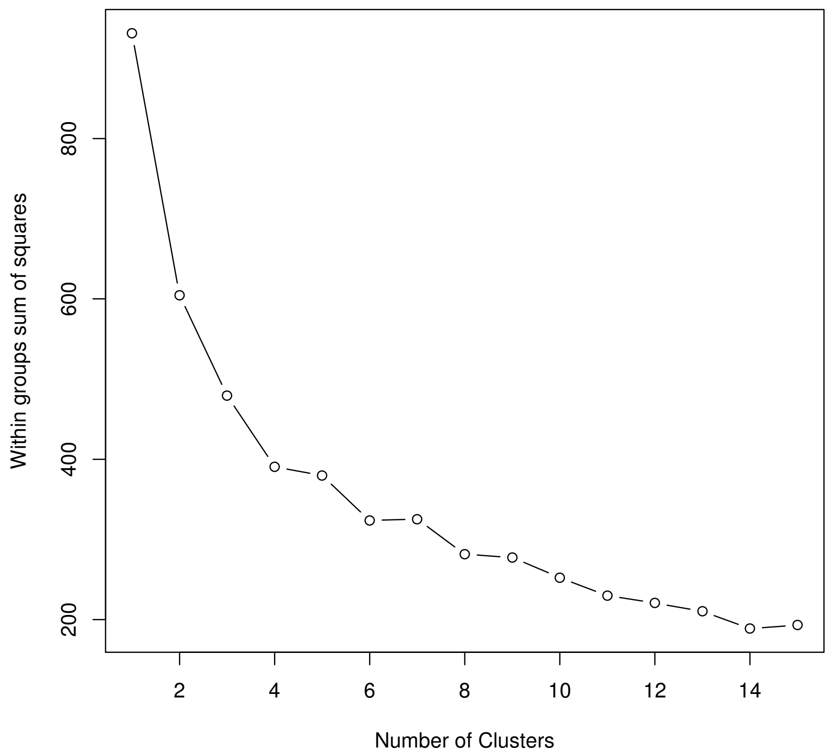


Figure 5.2: Within Group Sum of Squares vs. Number of Clusters for commitment bias

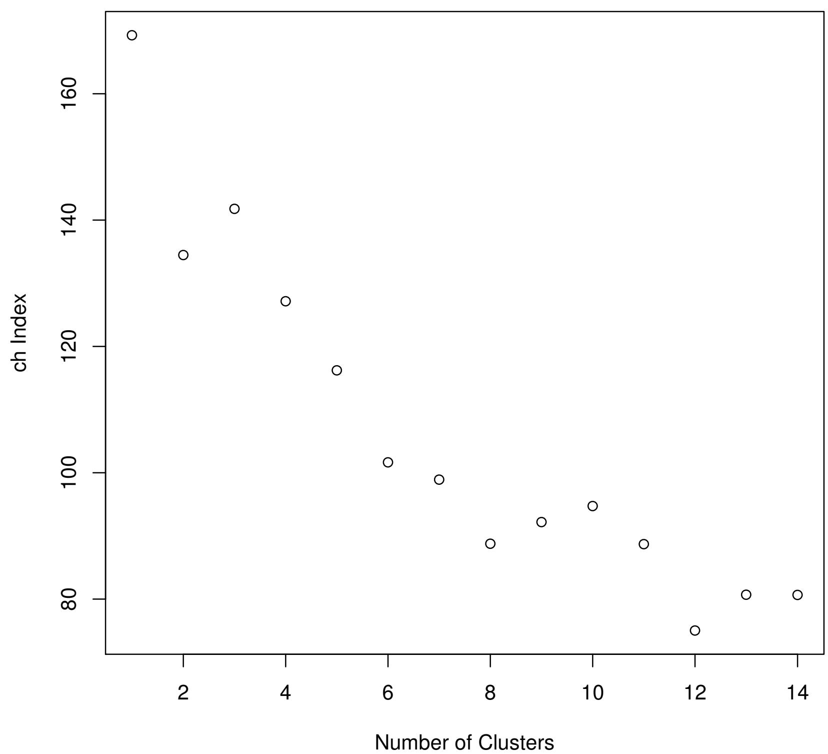


Figure 5.3: ch-Index vs. Number of Clusters for commitment bias

the total number of pairs. Mathematically, WGAC is calculated as follows:

$$\text{WGAC}(\mathbf{V}) = \frac{\sum_{X,Y\in\mathbf{V},X\neq Y} \text{corr}(X,Y)}{\binom{|\mathbf{V}|}{2}}$$

where $\mathbf{V}$ is the set of random variables for which we are computing WGAC, and $\text{corr}(X, Y)$ is the correlation between random variables X and Y. This measure is good for comparison across different subsets of questions, as it provides (averaged) comparable values of dependence among questions in that subset. Further, we multiply a reward factor with the normalized correlation measure, and maximize this value across all possible subsets. We define the reward factor as the number of questions present in the subset being considered. This reward factor penalizes smaller subsets and rewards bigger subsets, thus helping minimize the number of questions that we eliminate.

The (arbitrary) choice of the reward function has an impact on the number of questions in our optimal subset. Considering this flexibility in the reward function, we face the choice of choosing responses from three or four questions per bias. Choosing three questions per bias gives a stronger pairwise correlation, while choosing four question per bias retains more information from the survey. Also, it is worth noting that for all four biases, we find that the best *4-member* subset contains the optimal *3-member* subset. These sets are questions $\{Q8, Q9, Q10\}$ (0.41 WGAC) and $\{Q7, Q8, Q9, Q10\}$ (0.36 WGAC) respectively for commitment bias, questions $\{Q13, Q14, Q15\}$ (0.19 WGAC) and $\{Q13, Q14, Q15, Q17\}$ (0.14 WGAC) respectively for status quo bias, questions $\{Q18, Q19, Q21\}$ (0.22 WGAC) and $\{Q18, Q19, Q21, Q22\}$ (0.17 WGAC) respectively for social norms, and questions $\{Q23, Q24, Q25\}$ (0.43 WGAC) and $\{Q23, Q24, Q25, Q27\}$ (0.34 WGAC) respectively for gamification. In the end, we decided to keep as many questions as possible, and we opt for *4-member* subsets of questions for each bias. For further details on the correlation matrices, please refer to Figures 5.4 5.5 5.6 5.7.

Having selected the questions from the survey that we would want to consider, we now focus on identifying agent types.

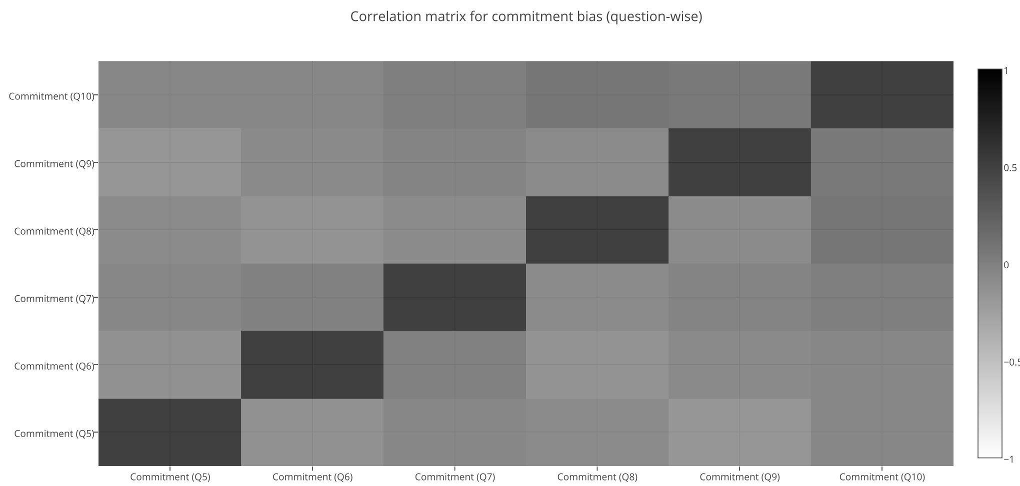


Figure 5.4: Correlation matrix showing pairwise correlation of responses for Commitment bias

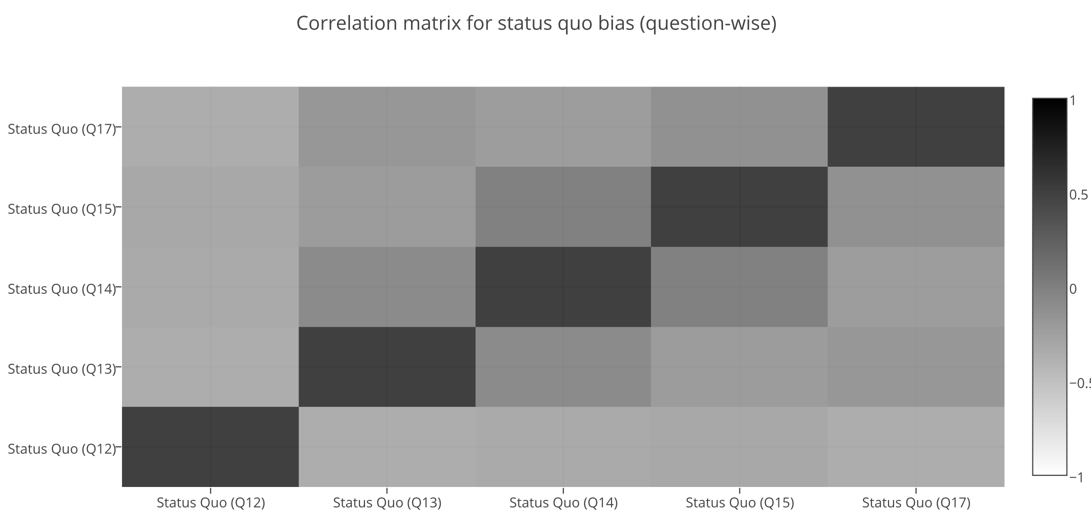


Figure 5.5: Correlation matrix showing pairwise correlation of responses for Status quo bias

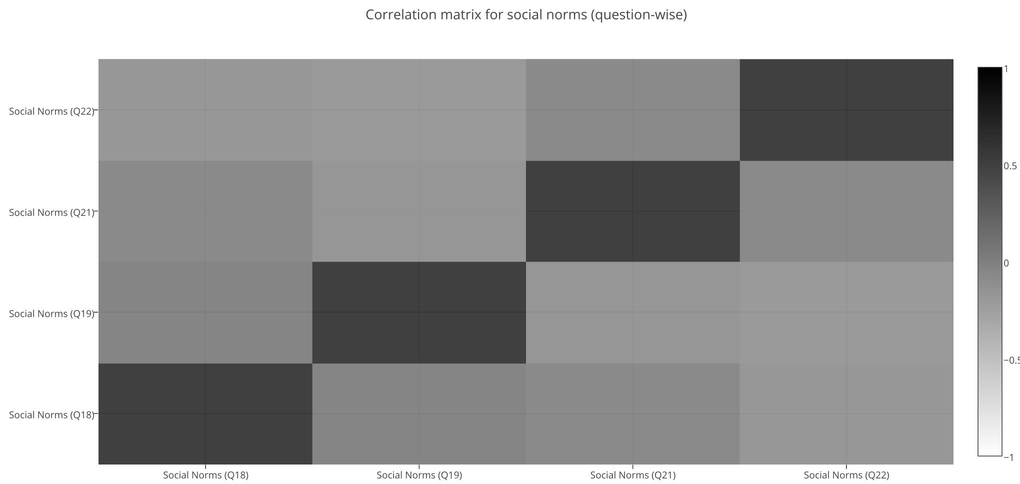


Figure 5.6: Correlation matrices showing pairwise correlation of responses for Social Norms

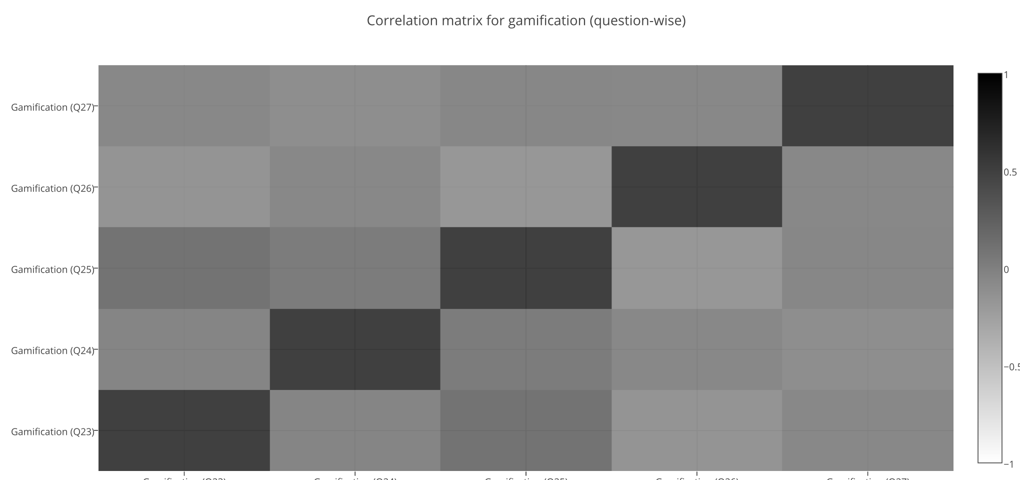


Figure 5.7: Correlation matrices showing pairwise correlation of responses for Gamification

5.4.2 Aggregate Score and Data Agreement

For better reliability of our psychometric scale, we have multiple questions for each bias. However, a survey participant may not have answered all the questions for a bias with similar and coherent Likert scores. For example, consider two responses on a linear Likert scale {0,1,2,3,4}: one that has all '4's on the Likert scale, and another that has two '4's, a '2', and a '1' on the Likert scale. We need to be able to assign a score to each response, for each bias, so that we have a metric for comparing across different survey responses.

We have four questions for measuring each bias. Specifically, for each survey response, we calculate an aggregate score for each bias, as the sum of Likert scores for that bias. We set the values of the utility parameters S, C, N, and G to their respective aggregate scores. The notion of aggregate score achieves two purposes. Firstly, it reduces dimensionality of observations from the four-dimensional space to the one-dimensional space. Secondly, it incorporates a concept of *agreement*, that refers to how close or distant a particular response is to the *ideal response*, thereby giving us a comparison metric across responses (separately for each bias). Since we measure the degree of inclination towards various biases, our *ideal response*, in this case, is a response that has '4's on the Likert scale for all questions, thus representing the highest bias score. We measure the distance (L1-norm) of a particular response from this ideal response. Additionally, we switch to using curved Likert interpretation, i.e. $\{0, 0.2, 2, 3.8, 4\}$ instead of $\{0, 1, 2, 3, 4\}$, which captures a notion of *psychological distance* between Likert items by overcoming the middle-of-scale effect [39]. Thus, for each bias, lower the L1-norm, better is the inclination towards that particular bias. It is worth noting that this method of comparing responses is identical to simply comparing the aggregate scores, where a higher aggregate score means better inclination towards the bias. However, we feel the former description provides a more intuitive understanding of why we choose to use this particular measure for agreement between responses.

The aggregate score that we computed in the previous paragraph, provides an absolute measurement of each bias. That is, it helps compare and rank across responses for each bias, but is inadequate for comparing responses between biases, as the effect of some biases may be more pronounced than others. Therefore, we have the need to rescale these aggregate scores for different biases by multiplying weights, so that they can be rendered comparable to each other. This is important since our ultimate aim is to calculate the agents' utilities derived from each of these biases, and hence the overall utility from all biases combined.

We obtain the scaling weights from question 35 of the survey, that inquires about the respondent's relative preferences across all four biases. Let us denote the values reported against the four biases in question 35, for agent i, as t_i^S, t_i^N, t_i^C, and t_i^G (for status quo, social norm, commitment and gamification respectively). Then, the weight for bias $b \in \{S, C, N, G\}$ is given by:

$$w_i^b = \frac{t_i^b}{\sum\limits_{\forall b \in \{S,C,N,G\}} t_i^b}$$

5.4.3 Calculating and Discounting Discomfort from Total Utility

Questions 35 and 36 from the survey elicit very specific information from respondents. In question 36, we ask the respondents of the maximum amount by which they would be willing to increase their thermostat temperatures, assuming that policies incorporating all the considered biases are implemented together. This information registers the maximum temperature ΔT_{max} that the respondent is willing to have their thermostat increased by. Let us call this respondent agent i. Based on the statement of the question, we can assume that the net utility obtained by agent i from all the biases exactly nullifies the discomfort experienced at ΔT_{max}. We can then calculate $\Pi_i(P^{\text{in}}, \Delta T_{max})$ from the data, using the following equation:

$$\Pi_i(P^{\text{in}}, \Delta T_{max}) = \alpha_i(P^{\text{in}} \circ (w_i^S, w_i^C, w_i^G, w_i^N))(S_i, C_i, N_i, G_i)^\intercal - d_i(\Delta T_{max})$$

where α_i is a scaling constant.

Recall from chapter 3, that we can calculate $d(\Delta T_{max})$ as follows:

$$d_i(\Delta T_{max}) = \begin{cases} 0 & \text{if } \Delta T \leq 1.6 \\ (0.306)(\Delta T) - 0.48 & \text{if } \Delta T > 1.6 \end{cases}$$

Since the utility from all biases cancels out the utility from discomfort at ΔT_{max}, we set $\Pi_i(P^{\text{in}} = (1,1,1,1), \Delta T_{max})$ to zero, and obtain the value of α_i as follows:

$$\alpha_i(w_i^S, w_i^C, w_i^G, w_i^N)(S_i, C_i, N_i, G_i)^\intercal - d(\Delta T_{max}) = 0 \implies \alpha_i = \frac{d_i(\Delta T_{max})}{(w_i^S, w_i^C, w_i^G, w_i^N)(S_i, C_i, N_i, G_i)^\intercal}$$

Finally, we can compute the utility of agent i at any given temperature change ΔT and incentives P^{in} using the following equation:

$$\Pi_i(P^{\text{in}}, \Delta T) = \alpha_i(P^{\text{in}} \circ (w_i^S, w_i^C, w_i^G, w_i^N))(S_i, C_i, N_i, G_i)^\intercal - d_i(\Delta T)$$

5.4.4 Principal's Action and Optimal Policy Model

A part of the principal's (utility company's) action is the temperature increase signals it sends to the agents' thermostat controller. Assuming that the principal sends the same temperature increase signal to all the agents, we now investigate the impact of inclusion or exclusion of each non-cash incentive in our policy. This lets us find the policy for which the principal maximizes its *energy saving objective*. We define the energy saving objective as the product of the number of participants and the signaled temperature increase.

We present our results based on data from 425 participants. The absolute values presented in our results would scale accordingly, if the target population is larger. Further, we base our results on the assumption that the cost of implementation of incentives is zero, since we do not have data about what these costs might be. In practice, the gains from each policy will need to be balanced by the implementation cost (per user) for that policy.

We consider all the 16 possible combinations of policy incentives that can be implemented. For each of these policy sets, we calculate the temperature increase signal for

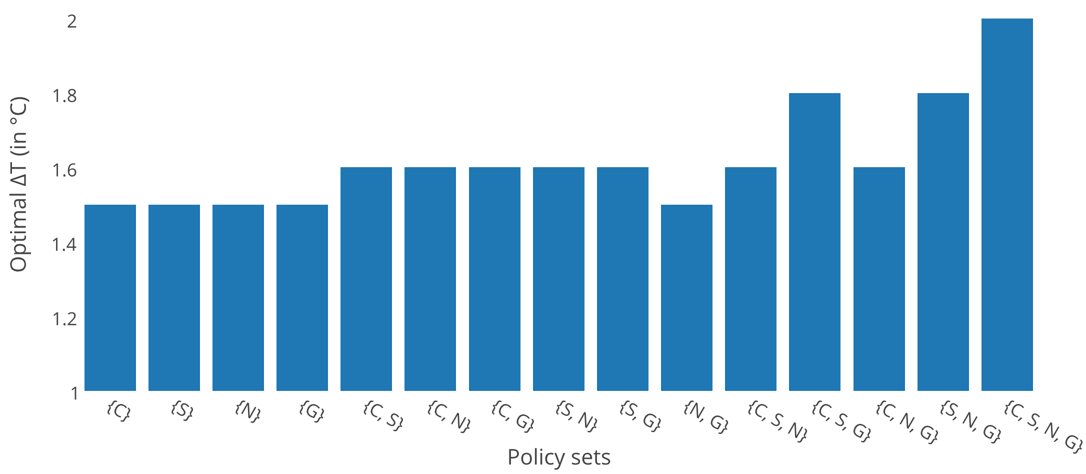


Figure 5.8: Optimal ΔT for each incentive policy. We use the notation C for commitment device, S for status quo bias, G for gamification and N for social norm.

which that policy set attains its maximum participation and maximum energy saving objective. We find that for the policy sets incorporating all four incentives, the energy saving objective is maximized at a temperature increase of $2°C$. For all other non-empty policy sets, the principal's energy saving objective is maximized at a lower value of ΔT. Figure 5.8 shows a plot of the optimal ΔT across policy sets. From among other policy sets, {C,S,G} and {S,N,G} record a better optimal ΔT of $1.8°C$.

Figure 5.9 shows plots of optimal energy saving objective (or principal's objective) and participation across policy sets, and optimal energy saving objective vs. policy set respectively, at these optimal ΔT signals. From the plots, it can be observed that implementing all the incentives in the policy results in the maximum energy saving objective of $542°C$, summed over all participants. It should be noted that, in this case, the participation is the lowest – showing that it is important to identify sections of the population that react positively to these policies, allowing for higher ΔT.

Figure 5.10 shows a plots of how the principal's objective and the participation vary with

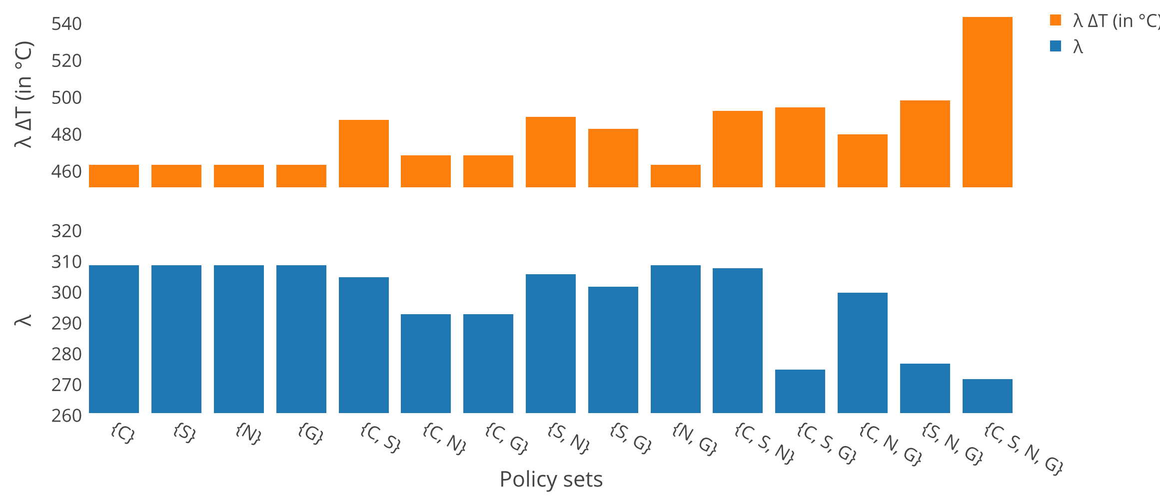


Figure 5.9: Participation rate (bottom graph) and principal's objective as a function of the optimal ΔT for each incentive policy (top graph). Where C denotes commitment device, S denotes status quo bias, G denotes gamification and N denotes social norm.

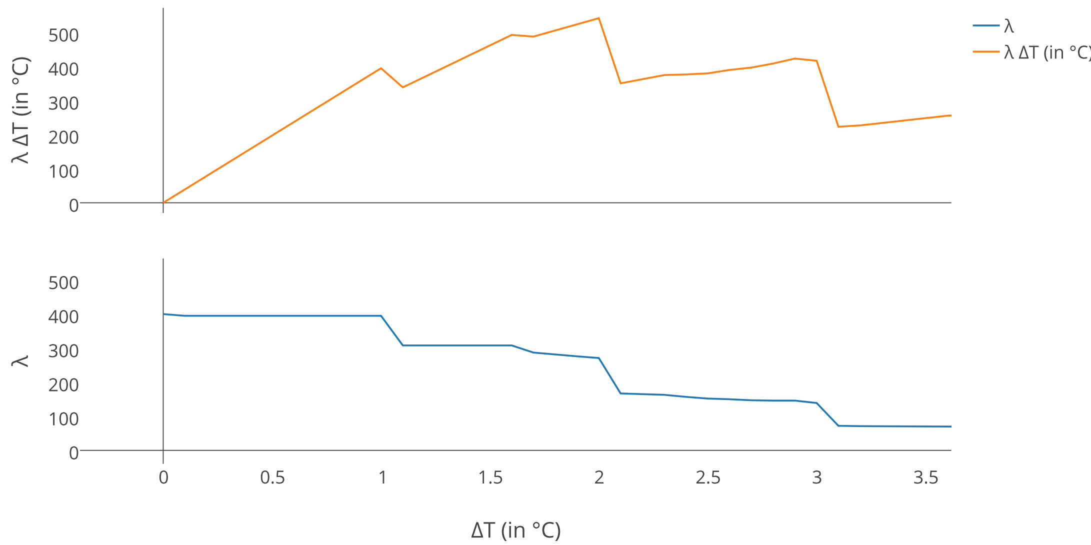


Figure 5.10: Change in participation rate (bottom graph) and principal's utility (top graph) as a function of ΔT when all incentives are used.

change in ΔT, for the policy set implementing all policies. As is expected, participation drops as ΔT increases. The principal's objective is maximized at $2°C$. Sharp drops in the plot of $\lambda\Delta T$ are observed due to granularity of survey data.

To study the marginal utility of each incentive in each policy set, we plot the marginals using a Hasse diagram, showed in Figure 5.11. For a policy maker, this diagram is designed to provide key insights, as it depicts the marginal improvements in utility for all possible combinations of incentives and policy sets – particularly, since our results are based on the assumption that the cost of implementation of incentives is zero.

In the Hasse diagram, the nodes represent policy sets and the edges represent marginal utilities (in terms of the principal's objective i.e. $\lambda\Delta T$). Nodes appear in four levels, based on the number of incentives they contain. Incentives are color-coded. The colors red, green, blue, and yellow, respectively represent incentives from commitment, status quo, social norms, and gamification. The Hasse diagram may be read starting from the bottom-most node (the empty policy set). As we move up, incentives are added into the nodes and the edges leading to those nodes (from below) reflect the marginal utility of the respective color-coded incentive. The marginal utilities of all incentives are equal, if the policy sets contain exactly one incentive. This shows that when only one incentive is offered, the discomfort factor dominates, i.e., agents make their decisions based solely on their discomfort preferences. If we add more incentives, we start observing different marginal utilities for different incentives, based on the preferences of the agents. When exactly two incentives are provided, the policy sets $\{Commitment, Status\ quo\}$, $\{Status\ quo, Social\ Norms\}$, and $\{Status\ quo, Gamification\}$ outperform others. This clearly shows that marginal utility from adding *Status quo* bias is greater than other biases. In the case the policy sets have exactly three incentives, if we inspect edges joining the nodes (with three incentives) from below, the green edges representing *Status quo* bias have significantly higher values than other edges – showing that the marginal contribution of the *Status quo* bias was the highest. This reflects in the values of the net utility of these policy sets as well, as $\{Commitment, Social\ norms, Gamification\}$ performs the worst – with the other three policy sets having approximately similar total utility. To summarize, this shows that {Status quo} bias has the greatest impact among the four incentives that we studied.

If the principal sends non-uniform signals, its objective can be further maximized. Since the principal already has the values of ΔT_{max} for each participant from the survey (Q36), it may directly send these values as signals to the respective agents. In that case, according to our data from 425 participants, the gains for the principal would be $967°C$, which is roughly 78% higher than the case where the principal sends identical temperature change signals. However, for practically implementing non-uniform signals, survey responses must be obtained for each participant from the target population so that their individual preferences can be clearly estimated. Although, conducting such a survey for the entire population could be expensive.

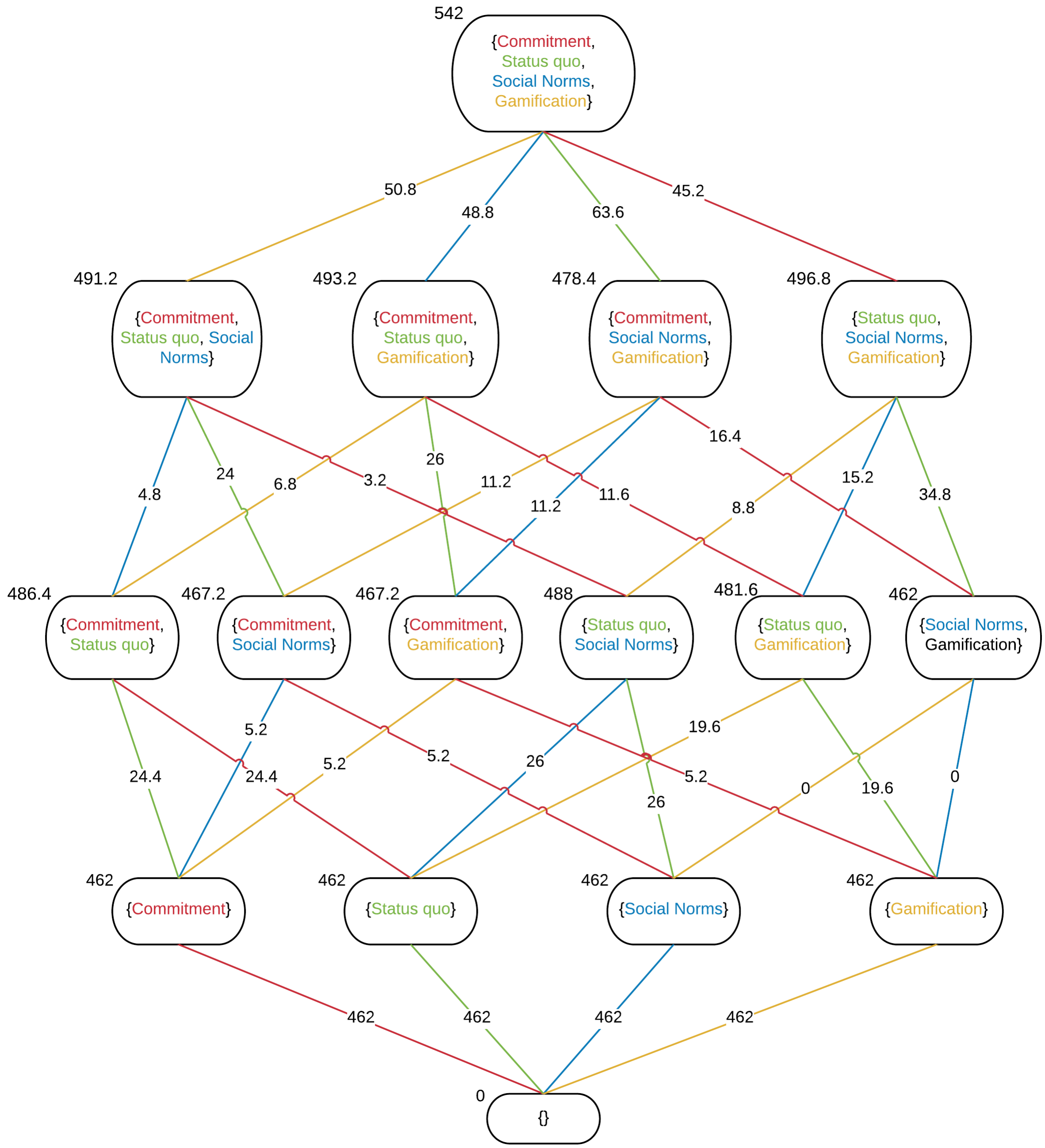


Figure 5.11: Hasse diagram showing marginal utilities of incentives over the power set of policies.

Chapter 6

Conclusion and Future Work

In this thesis, we showed that energy policies can significantly benefit from behavioral research. To our knowledge, this is the first game-theoretic and data-driven analysis that quantifies and models the role of behavioral incentives in designing energy policies. Our most important contributions have been laying the framework for comparing and contrasting the efficacy of policies, identifying how target population sections respond to individual policies, and hence proposing an evidence-based approach for decision-making in policy design. Our proposed methodologies provide policy makers with the necessary tools to make wise and informed choices.

With respect to our specific problem instantiation of *peaksaver* PLUS penetration, we showed that the utility companies can make a homogeneous increase in thermostat temperatures by up to $2°C$ for targeted sections of the population. Further, financial viability admitting, the utility companies may elicit preferences from the entire population enabling them to send heterogeneous signals. Our computations suggest that heterogeneous signals can achieve about 78% more energy savings.

To obtain ground truth data, we designed and conducted a survey. We also recommended a subset of questions from our questionnaire that can be used for gauging the participants' stimuli to commitment devices, status quo bias, social norms, and gamification; and how these biases relate to thermal discomfort. We proposed a post hoc survey cleaning methodology that ensures higher reliability of data. This is particularly important for surveys that aim to elicit psychological or behavioral preferences, as the inherent concepts that we try to quantify are abstract in nature – thus providing more structure to the art of designing psychometric surveys.

It is worthy of note that our survey results are based on responses from all across Canada and the United States. Within the scope of our research study, we conducted the survey on a sample population of 425, obtained from an original data set of 990 responses. However, both in principle and practice, our propositions are highly scalable. Which is why we coin the term "Big Data Mechanism Design" for our work. With adequate support from government authorities and administrations, we envisage our model and policy to be implementable on a large scale with millions of people participating. Further, a similar baseline approach can be used to verify, compare, and work with other newer policies – possibly in domains other than energy. Thus, our work also carries a broader appeal to policy makers in general.

Due to various theoretical and practical constraints, we made some assumptions to tractably study the problem. These assumptions pose some limitations on our proposed models and methodologies. We assumed that the agents' utility depends only on the biases and thermal discomfort. In reality, the agents' utility could depend on other external factors too. Also, in calculating the net utility, we assume that the utility from biases and discomfort are additive and linear. This assumption was critical in quantifying the utility from biases, but it was also the most naive way to formulate the utility model. A better model for formulating the utility functions could indeed exist. As for the behavior of the agents, we made the standard game-theoretic assumption that they are individually rational and that they participate if and only if their utility is non-negative (a weakly dominant strategy). From the principal's point of view, we assume that it would implement a policy based on its expected energy savings. Although this is a reasonable assumption, due to lack of data we were unable to factor in the cost of policy implementation – an important factor in the principal's decision making. Instead, we provide the policy makers with information regarding the marginal benefit of adding or subtracting incentives to the policy. Ultimately, the credibility of our results are greatly dependent on the correctness of our survey elicitation procedure and our instantiation of the PMV model.

Prospects for future work based on our current study are numerous. We tried clustering agent types based on their utility from different biases, albeit with limited success. Further investigation in the direction of using machine learning approaches could yield interesting results. For example, one could take a completely computational approach and train a classifier to determine agent participation, instead of relying on a utility model. This method, however, would require collecting data on an even larger scale to avoid overfitting concerns. We elicited a small set of features through our survey, and likely missed some important features like potentially useful demographic information, unaccounted psychological and economic variables, etc. A larger and richer dataset could present more valuable insights. Other directions for future research could include extensions in domains of policy making outside of the energy sector, and mechanisms for studying treatment of a multiple customized policies for targeted sections of the population (neither homogeneous nor completely heterogeneous). Also, surveys, as a method for preference elicitation, could be replaced by more elaborate field experiments. Data from field experiments could help calibrate the degree of response bias that might be implicitly prevalent in survey data.

References

[1] Archived - harper government delivers on commitment to protect canadian consumers from spam and online threats - canada news centre. http://news.gc.ca/web/article-en.do?nid=864329. (Visited on 10/19/2015).

[2] Canada's law on spam and other electronic threats - home - canada's anti-spam legislation. http://fightspam.gc.ca/eic/site/030.nsf/eng/home. (Visited on 08/13/2015).

[3] Cansim - 051-0005 - estimates of population, canada, provinces and territories. http://www5.statcan.gc.ca/cansim/a26?lang=eng&retrLang=eng&id=0510005&paSer=&pattern=&stByVal=1&p1=1&p2=31&tabMode=dataTable&csid=. (Visited on 08/13/2015).

[4] Cbe thermal comfort tool for ashrae-55. http://comfort.cbe.berkeley.edu/. (Visited on 09/28/2015).

[5] E-markets (b2b, b2c, b2g)- canadian company capabilities - industry canada. http://www.ic.gc.ca/app/ccc/sld/cmpny.do?lang=eng&profileId=2059&tag=025031008002. (Visited on 10/19/2015).

[6] Enersource. http://www.enersource.com/energy-savings-tips/Pages/peak_saver.aspx. Accessed: 2015-08-13.

[7] Hydro one. http://www.hydroone.com/MyHome/SaveEnergy/Pages/Peaksaver.aspx. Accessed: 2015-08-13.

[8] Office of research ethics, university of waterloo, application process. https://uwaterloo.ca/research/office-research-ethics/research-human-participants/application-process. (Visited on 11/09/2015).

[9] Office of research ethics, university of waterloo, homepage. https://uwaterloo.ca/research/office-research-ethics. (Visited on 11/09/2015).

[10] Population clock. http://www.census.gov/popclock/. (Visited on 08/13/2015).

[11] Sample size calculator by raosoft, inc. http://www.raosoft.com/samplesize.html. (Visited on 08/13/2015).

[12] Survey methods and practices. http://www.statcan.gc.ca/pub/12-587-x/12-587-x2003001-eng.pdf. (Visited on 11/11/2015).

[13] http://www.qualtrics.com/blog/writing-survey-questions/.

[14] Hunt Allcott and Sendhil Mullainathan. Behavioral science and energy policy. *Science*, 327(5970):1204–1205, 2010.

[15] Hunt Allcott and Todd Rogers. The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. Technical report, National Bureau of Economic Research, 2012.

[16] Dan Ariely and Simon Jones. *Predictably irrational*. HarperCollins New York, 2008.

[17] Dan Ariely and Simon Jones. *The upside of irrationality: The unexpected benefits of defying logic at work and at home*, volume 159. Harper New York, 2010.

[18] Ontario Power Authority. peaksaver residential air conditioner measurement and verification study. http://www.powerauthority.on.ca/sites/default/files/new_files/2009/2009%20peaksaver%20Residential%20Air%20Conditioner%20Measurement%20and%20Verification%20Study.pdf, 2009. (Visited on 11/30/2015).

[19] VSKM Balijepalli, Vedanta Pradhan, SA Khaparde, and RM Shereef. Review of demand response under smart grid paradigm. In *Innovative Smart Grid Technologies-India (ISGT India), 2011 IEEE PES*, pages 236–243. IEEE, 2011.

[20] P Ben-Nun. Respondent fatigue. *Encyclopedia of survey research methods*, pages 743–744, 2008.

[21] Gharad Bryan, Dean Karlan, and Scott Nelson. Commitment devices. *Annu. Rev. Econ.*, 2(1):671–698, 2010.

[22] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[23] Robert B Cialdini and Melanie R Trost. Social influence: Social norms, conformity and compliance. 1998.

[24] Jon Chandler Creyts. *Reducing US greenhouse gas emissions: how much at what cost?: US Greenhouse Gas Abatement Mapping Initiative*. McKinsey & Co., 2007.

[25] Robyn M Dawes. The robust beauty of improper linear models in decision making. *American psychologist*, 34(7):571, 1979.

[26] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, pages 9–15. ACM, 2011.

[27] Poul O Fanger et al. Thermal comfort. analysis and applications in environmental engineering. *Thermal comfort. Analysis and applications in environmental engineering.*, 1970.

[28] Peter C Fishburn. Independence in utility theory with whole product sets. *Operations Research*, 13(1):28–45, 1965.

[29] Frederick X Gibbons and Bram P Buunk. Individual differences in social comparison: development of a scale of social comparison orientation. *Journal of personality and social psychology*, 76(1):129, 1999.

[30] Kenneth Gillingham, Richard G Newell, and Karen Palmer. Energy efficiency economics and policy. Technical report, National Bureau of Economic Research, 2009.

[31] Hannah Choi Granade, Jon Creyts, Anton Derkach, Philip Farese, Scott Nyquist, and Ken Ostrowski. Unlocking energy efficiency in the us economy. 2009.

[32] Kelly Garrett Jason Peifer. http://www.comm.ohio-state.edu/Opt-in_panel_best_practices.pdf.

[33] Oliver P John and Sanjay Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138, 1999.

[34] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[35] Daniel Kahneman, Jack L Knetsch, and Richard H Thaler. Anomalies: The endowment effect, loss aversion, and status quo bias. *The journal of economic perspectives*, pages 193–206, 1991.

[36] Daniel Kahneman and Amos Tversky. Choices, values, and frames. *American psychologist*, 39(4):341, 1984.

[37] Ralph L Keeney. The art of assessing multiattribute utility functions. *Organizational Behavior and Human Performance*, 19(2):267–310, 1977.

[38] Wojtek J Krzanowski and YT Lai. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, pages 23–34, 1988.

[39] Bjorn Lantz. Equidistance of likert-type scales and validation of inferential methods using experiments and simulations. *Electronic Journal of Business Research Methods*, 11(1), 2013.

[40] Alex Laskey and Ogi Kavazovic. Opower. *XRDS: Crossroads, The ACM Magazine for Students*, 17(4):47–51, 2011.

[41] Kibeom Lee and Michael C Ashton. Psychometric properties of the hexaco personality inventory. *Multivariate Behavioral Research*, 39(2):329–358, 2004.

[42] Richard D Lennox and Raymond N Wolfe. Revision of the self-monitoring scale. 1984.

[43] Thomas C Leonard. Richard h. thaler, cass r. sunstein, nudge: Improving decisions about health, wealth, and happiness. *Constitutional Political Economy*, 19(4):356–360, 2008.

[44] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[45] Jessie Ma, Danilo Yu, Craig Brown, and Bala Venkatesh. Conservation and demand management: Where we are today. 2015.

[46] Paul E Meehl. When shall we use our heads instead of the formula? *Journal of Counseling Psychology*, 4(4):268, 1957.

[47] Tomas Nauclér and Per-Anders Enkvist. Pathways to a low-carbon economy: Version 2 of the global greenhouse gas abatement cost curve. *McKinsey & Company*, 192, 2009.

[48] Guy R. Newsham and Brent G. Bowker. The effect of utility time-varying pricing and load control strategies on residential summer peak electricity use: A review. *Energy Policy*, 38(7):3289 – 3296, 2010. ¡ce:title¿Large-scale wind power in electricity markets with Regular Papers¡/ce:title¿.

[49] US Department of Energy. Benefits of Demand Response in Electricity Markets and Recommendations for Achieving them. *Report to the United States Congress*, 2006.

[50] Julia Reinaud and Amelie Goldberg. The boardroom perspective: How does energy efficiency policy influence decision making in industry? 2012.

[51] Lee Ross and Richard E Nisbett. *The person and the situation: Perspectives of social psychology*. Pinter & Martin Publishers, 2011.

[52] Ian H Rowlands. Demand response in ontario: exploring the issues. *A report for the Independent Electricity System Operator (IESO) of Ontario*, 2008.

[53] William Samuelson and Richard Zeckhauser. Status quo bias in decision making. *Journal of risk and uncertainty*, 1(1):7–59, 1988.

[54] Simone Schneider and Jürgen Schupp. The social comparison scale: Testing the validity, reliability, and applicability of the iowa-netherlands comparison orientation measure (incom) on the german population. 2011.

[55] Ralf Schwarzer and Matthias Jerusalem. Generalized self-efficacy scale. *Measures in health psychology: A users portfolio. Causal and control beliefs*, 1:35–37, 1995.

[56] Swati Singla and Srinivasan Keshav. Demand response through a temperature setpoint market in ontario. In *Smart Grid Communications (SmartGridComm), 2012 IEEE Third International Conference on*, pages 103–108. IEEE, 2012.

[57] Marcin Sklad and Rene Diekstra. The development of the heuristics and biases scale (hbs). *Procedia-Social and Behavioral Sciences*, 112:710–718, 2014.

[58] ASHRAE Standard. Standard 55-2013. thermal environmental conditions for human occupancy. ashrae. *Atlanta, USA*, 2013.

[59] Valerie Sugarman and Edward Lank. Designing persuasive technology to manage peak electricity demand in ontario homes. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1975–1984. ACM, 2015.

[60] June P Tangney, Roy F Baumeister, and Angie Luzio Boone. High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of personality*, 72(2):271–324, 2004.

[61] Yevgeniy Vorobeychik. Mechanism design and analysis using empirical game models. Technical report, Technical report, University of Michigan, 2006.

[62] Yevgeniy Vorobeychik. *Mechanism design and analysis using simulation-based game models*. ProQuest, 2008.

[63] Yevgeniy Vorobeychik, Christopher Kiekintveld, and Michael P Wellman. Empirical mechanism design: Methods, with application to a supply-chain scenario. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 306–315. ACM, 2006.

[64] Michael P Wellman. Methods for empirical game-theoretic analysis. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1552. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

APPENDICES

Appendix A

Survey Questionnaire

Preferences regarding everyday scenarios and home thermostat settings

Information about the survey

You are invited to participate in a research study conducted by Ankit Pat, under the supervision of Prof. S. Keshav and Prof. Kate Larson of the Department of Computer Science, University of Waterloo, Canada. The objectives of the research study are to gauge preferences of occupants of residential buildings with respect to temperature settings of their thermostat. Our key focus is to design a flexible architecture that lets us experiment with a variety of schemes for non-cash incentives. To this end, we ask questions about how you make decisions in different everyday scenarios. The study is for a master's thesis.

If you decide to volunteer, you will be asked to complete a 5-minute online survey that is completed anonymously. Participation in this study is voluntary. We appreciate your answering all the questions; however, you are not obliged to answer any questions that you deem objectionable or are uncomfortable answering. You may decline to answer any questions that you do not wish to answer by leaving it blank, and you can withdraw your participation at any time by not submitting your responses and closing the survey's browser window, without penalty or loss of remuneration. To receive remuneration please proceed to the end of the questionnaire, obtain the unique code for this HIT, and submit it. There will be no adverse consequence of choosing not to participate in the survey. If you withdraw prior to completing and submitting the survey all data entered will be permanently removed. There are no known or anticipated risks from participating in this study.

Any information that you provide will be confidential. All data will be summarized and no individual could be identified from the summarized data. Also, the website collects responses alone and will not collect any information that could potentially identify you. As this is an anonymous survey, the researchers have no way of identifying you or getting in touch with you should you choose to tell us something about yourself or your life experiences. The data, with no personal identifiers will be maintained on a password-protected computer database in a restricted access area of the university. The data will be electronically archived after completion of the study and maintained for two years and then erased. Moreover, no individual data will be shared with the public in any form.

This survey uses Survey MonkeyTM which is a United States of America company. Consequently, USA authorities under provisions of the Patriot Act may access this survey data. If you prefer not to submit your data through Survey MonkeyTM, please do not accept this HIT.

Any remuneration amount that you receive by completing this HIT may be taxable. It is your responsibility to report this amount for income tax purposes.

The results of this survey may be published in professional journals or presented at scientific conferences, but such presentations will report only aggregated findings, which in some instances may be illustrated by short, anonymous quotes carefully selected so as not to breach individual confidentiality. If you have any questions about the survey, or are interested in receiving information about the study's findings, you can contact Prof. Keshav at keshav@uwaterloo.ca or (519) 888-4567x34456. Any questions about survey may also be directed to Prof. S. Keshav.

This survey has been reviewed by and received ethics clearance through a University of Waterloo Research Ethics Committee. However, the final decision about participation is yours. If you have any comments/concerns resulting from your participation in this study, please feel free to contact Dr. Maureen Nummelin in the Office of Research Ethics at 1-519-888-4567, Ext. 36005 or maureen.nummelin@uwaterloo.ca.

This research is funded by a grant from the Canadian National Science and Engineering Research Council with additional support from Cisco Systems.

Thank you for considering participation in this study.

Preferences regarding everyday scenarios and home thermostat settings

1. By selecting “yes” as an answer to survey question 1 below, you agree:

(i) to have read and understood the information presented in this letter.
(ii) that you are aware that you may withdraw from the study without penalty at any time by simply not completing the online survey.
(iii) have had the opportunity to ask any questions related to this study and receive satisfactory answers to your questions, and any additional details you may have wanted.
(iv) to being at least 18 years old.

Please confirm that you understood the purpose and conditions of the study and with full knowledge of all foregoing, agree of your own free will, to participate in this study. Otherwise, you may simply close this browser window.

○ Yes, I agree to participate in the study.

Preferences regarding everyday scenarios and home thermostat settings

2. Do you have a thermostat in your house?

Yes	No	Can not answer / Prefer not to answer
○	○	○

3. How is your electricity bill determined?

I pay a variable cost based on consumption	I pay a fixed monthly cost (e.g., included in rent)	Can not answer / Prefer not to answer
○	○	○

Other (please specify)

4. Do you have air conditioning in your house?

Yes	No	Can not answer / Prefer not to answer
○	○	○

Preferences regarding everyday scenarios and home thermostat settings

5. Suppose you and your friend agree to go to the gym together, a few times a week. How likely are you to?

Not likely	Less likely	As likely as not	Very likely	Definitely
○	○	○	○	○

6. Suppose you promised yourself that you would read more books. How likely are you to follow through?

Not likely	Less likely	As likely as not	Very likely	Definitely
○	○	○	○	○

7. When you commit to do something, in general, how often do you carry it out?

Never	Rarely	Sometimes	Most of the Time	Always
○	○	○	○	○

8. Suppose you are on a diet. How likely are you to successfully resist that extra piece of cake?

Not likely	Less likely	As likely as not	Very likely	Definitely
○	○	○	○	○

9. How likely are you to successfully abstain from browsing online, and instead concentrate on work?

Not likely	Less likely	As likely as not	Very likely	Definitely
○	○	○	○	○

10. How good are you at resisting temptations, in general?

Very bad	Bad	Neither good nor bad	Good	Very good
○	○	○	○	○

11. To what extent do you agree or disagree with the statement that the sun is bigger than the earth.

Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly agree
○	○	○	○	○

Preferences regarding everyday scenarios and home thermostat settings

12. Imagine that you inherited some money, and this money was already in a moderate-risk investment. Assuming that you wish to keep the money invested, how likely would you change how the money is invested.

Not likely	Less likely	As likely as not	Very likely	Definitely
○	○	○	○	○

13. Suppose you got a new mobile phone. How likely are you to change the default ringtone?

Not likely	Less likely	As likely as not	Very likely	Definitely
○	○	○	○	○

14. While grocery shopping, how likely are you to stick to the same brand of milk every time?

Not likely	Less likely	As likely as not	Very likely	Definitely
○	○	○	○	○

15. To what extent would you agree or disagree with the statement, "I usually use the same travel website to make my travel bookings."

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
○	○	○	○	○

16. All of us commute to work or school everyday. Many avail public transport like buses. The mode of transport may have an effect on how people feel when they reach work/school. To ensure good quality responses, we sometimes include questions that tell us if a participant was serious about the answers. To show that you understand this, please ignore everything else in this question and select only the Forgiving option below.

- [ ] Alert
- [ ] Ashamed
- [ ] Determined
- [ ] Disappointed
- [ ] Excited
- [ ] Forgiving
- [ ] Guilty
- [ ] Happy
- [ ] Irritated
- [ ] Nervous
- [ ] Proud
- [ ] Sad
- [ ] Scared
- [ ] Upset

17. Another bank is offering a slightly higher interest rate on savings, as compared to your bank. How likely are you to move your savings account to this other bank?

Not likely	Less likely	As likely as not	Very likely	Definitely
○	○	○	○	○

Preferences regarding everyday scenarios and home thermostat settings

18. Suppose an acquaintance of yours invited you to a dinner party. How likely are you to bring a gift, for e.g., a bottle of wine or flowers?

Not likely	Less likely	As likely as not	Very likely	Definitely
○	○	○	○	○

19. If others around you are tipping the waiter, how likely are you to leave a tip as well?

Not likely	Less likely	As likely as not	Very likely	Definitely
○	○	○	○	○

20. To what extent do you agree with the statement "Five dollars plus fifteen dollars is equal to three dollars."

Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly agree
○	○	○	○	○

21. How likely are you to participate in an activity if others around you are also participating? For example, the Ice Bucket Challenge.

Not likely	Less likely	As likely as not	Very likely	Definitely
○	○	○	○	○

22. To what extent do you care what people around you think of you?

Very little	A little	Neither little nor much	Much	Very much
○	○	○	○	○

Preferences regarding everyday scenarios and home thermostat settings

23. To what extent do you find public endorsements such as Facebook "Likes" gratifying?

Very little	A little	Neither little nor much	Much	Very much
○	○	○	○	○

24. How much do you like earning stars with Starbucks coffee, or a chance at a lottery with Tim Horton's coffee?

Very little	A little	Neither little nor much	Much	Very much
○	○	○	○	○

25. To what extent do you appreciate reputation points on online communities?

Very little	A little	Neither little nor much	Much	Very much
○	○	○	○	○

26. How likely are you to use Shopper's cards with reward points?

Not likely	Less likely	As likely as not	Very likely	Definitely
○	○	○	○	○

27. To what extent do you think feedback mechanisms like grades, employee ratings, etc. make (have made) you push harder to perform better?

Very little	A little	Neither little nor much	Much	Very much
○	○	○	○	○

Preferences regarding everyday scenarios and home thermostat settings

28. During the summer (Jun to Aug), on an average, how many weeks per month is nobody home? (for e.g. out on vacation etc.)

Less than 1 week per month	Between 1 and 2 weeks, per month	Between 2 and 3 weeks, per month	More than 3 weeks, per month	Prefer not to answer
○	○	○	○	○

29. In summer (Jun to Aug), on an average, how many weeks per month is nobody home during the afternoon (from 3pm to 6pm)?

Less than 1 week per month	Between 1 and 2 weeks, per month	Between 2 and 3 weeks, per month	More than 3 weeks, per month	Prefer not to answer
○	○	○	○	○

30. How likely do you think the hottest day of summer will be hotter than the coldest day of winter?

Not likely	Less likely	As likely as not	Very likely	Definitely
○	○	○	○	○

Preferences regarding everyday scenarios and home thermostat settings

31. What is your usual thermostat (or air conditioner) setting during summer afternoons (between 3pm and 6pm)? You may choose more than one option.

- ☐ 16°C / 61°F or less
- ☐ 17°C / 62.5°F
- ☐ 18°C / 64.5°F
- ☐ 19°C / 66°F
- ☐ 20°C / 68°F
- ☐ 21°C / 70°F
- ☐ 22°C / 71.5°F
- ☐ 23°C / 73.5°F
- ☐ 24°C / 75°F
- ☐ 25°C / 77°F
- ☐ 26°C / 79°F
- ☐ 27°C / 80.5°F
- ☐ 28°C / 82.5°F or more
- ☐ Can not answer

32. How often during a day does your thermostat temperature setting change, either manually or due to a programmed setting?

Does not change	Changes 1 or 2 times	Changes 3 or 4 times	Changes 5 or 6 times	Changes more than 6 times
○	○	○	○	○

Other (please specify)

33. How likely do you think the hottest day of summer will be colder than the coldest day of winter?

Not likely	Less likely	As likely as not	Very likely	Definitely
○	○	○	○	○

34. How easy/hard is it for occupants of your residence to agree on a single temperature setting?

Very easy	Easy	Neither easy nor hard	Hard	Very hard
○	○	○	○	○

Preferences regarding everyday scenarios and home thermostat settings

35. To the best of your ability, please consider each of the following scenarios as independent situations while answering.

	0°C / 0° F	1°C / 2° F	2°C / 3.5°F	3°C / 5.5°F	4°C / 7° F or more
Would you be willing to make a commitment to set your thermostat higher than you currently do during summer afternoons (3PM to 6PM)? If so, how much higher are you willing to commit to adjust your thermostat?	○	○	○	○	○
Suppose many of your neighbors are participating in an Energy Saving Program that sets your thermostat slightly higher during summer afternoons (3PM to 6PM). By how much would you volunteer to set your thermostat higher?	○	○	○	○	○
How many degrees warmer could your thermostat be adjusted during summer afternoons (3PM to 6PM) before you would get up and change it?	○	○	○	○	○
Consider "Green Score", a numerical rating in 0 to 5 range, which is awarded basing on the level of contribution to the Energy Saving Program. Imagine your residence is awarded a Green Score badge, showing your dedication to take actions to save energy and help the environment. You may also be given additional perks for participating in the Energy Saving Program, like an annual dinner with the mayor, to which only people who have also earned a Green Score badge are invited. To what extent would the introduction of Green Score badge encourage you to set your thermostat higher during summer afternoons (3PM to 6PM)?	○	○	○	○	○

36. Now, please consider the combined effect of the four scenarios presented in the previous question.

	0°C / 0° F	1°C / 2° F	2°C / 3.5°F	3°C / 5.5°F	4°C / 7° F or more
For this question, your answer should either be the maximum value of the four responses you gave in the last question, or more than that. Keeping that in mind, in total, how much higher are you willing to adjust your thermostat than your usual (during summer afternoons)?	○	○	○	○	○

37. Please enter your CrowdFlower Contributor ID.

38. Lastly, it is important to our study that we only include responses from people who devoted their full attention to this study.

Your answer to this question will not affect your payment in any manner whatsoever.

In your honest opinion, should we use your data?

- No, I was not paying attention while answering.
- Yes, go ahead. I answered to the best of my knowledge and capability.

Appendix B

Psychometric Scales

B.1 Social Comparison Scale

The social comparison scale [54] is a short version of the INCOM scale [29]. It deals with two distinctive underlying dimensions of social comparisons: (a) comparisons of abilities referring to the question "How am I doing as compared to others?", and (b) comparisons of opinions referring to the question "How do my feelings/thoughts compare to those of others?" In particular, the scale consists of the following questions seeking responses on the Likert scale [44] ranging from "Strongly disagree" to "Strongly agree":

1. I always pay a lot of attention to how I do things compared with how others do things.

2. I often compare how I am doing socially (e.g., social skills, popularity) with other people.

3. I am not the type of person who often compares with others.

4. I am not the type of person who often compares with others.I often try to find out what others think who face similar problems as I face.

5. I always like to know what others in a similar situation would do.

6. If I want to learn more about something, I try to find out what others think about it.

B.2 HEXACO & the Big Five Inventory Scales

The HEXACO [41] and the Big Five Inventory [33] scales claim to measure fundamental personality traits along different dimensions. The Big Five Inventory measures personality traits along the following five dimensions:

1. Extraversion vs. introversion: Correlated behavioral qualities along this dimension include gregariousness, assertiveness, activity (energetic), excitement-seeking, enthusiastic, outgoing nature, etc.

2. Agreeableness vs. antagonism: Correlated behavioral qualities include trust, straightforwardness, altruism, compliance, modesty, Tender-mindedness, etc.

3. Conscientiousness vs. lack of direction: Correlated behavioral qualities include competence, order (organized), dutifulness, achievement striving, self-discipline, deliberation (not impulsive), etc.

4. Neuroticism vs. emotional stability: Correlated behavioral qualities include anxiety, angry hostility, depression, self-consciousness, impulsiveness, vulnerability (not self-confident), etc.

5. Openness vs. exclusiveness to experience: Correlated behavioral qualities include curious, fantasy (imaginative), aesthetics (artistic), wide interests, feelings (excitable), unconventional, etc.

The set of personality traits considered by the HEXACO scale is a superset of the Big Five Inventory, with the additional dimension of Honesty-Humility. This additional dimension represents traits like manipulative behavior for personal gain, temptation to break rules, lack of interest in lavish wealth and luxuries, feeling no special entitlement to elevated social status, etc.

B.3 Self-Monitoring Scale

Self-monitoring scale [42] measures the respondent's sensitivity to the expressive behavior of others, and their ability to modify self-presentation.

A representative set of questions from the self-monitoring scale follow:

1. In different situations and with different people, I often act like very different persons.
2. Although I know myself, I find that others do not know me.
3. Different people tend to have different impressions about the type of person I am.
4. I'm not always the person I appear to be.
5. I'm pretty good at entertaining people with jokes, anecdotes, and stories.
6. I guess I put on a show to impress or entertain people.
7. I can make impromptu speeches even on topics about which I have almost no information.
8. I have a quick wit.
9. I have never been good at games like charades or improvisational acting.
10. I can look anyone in the eye and tell a lie with a straight face (if for the right end).
11. I find it hard to imitate the behavior of other people. (C)'b
12. Even if I am not enjoying myself, I often pretend to be having a good time.
13. It's important to me to fit into the group I'm with.

B.4 Self-Efficacy Scale

Jerusalem and Schwarzer devised the General Self-efficacy scale [55]. The scale consists of ten items and, as its name indicates, was created to assess a general sense of perceived self-efficacy. Specifically, the scale seeks answers to the extent to which the respondent agrees to the following statements:

1. I can always manage to solve difficult problems if I try hard enough.
2. If someone opposes me, I can find the means and ways to get what I want.
3. It is easy for me to stick to my aims and accomplish my goals.
4. I am confident that I could deal efficiently with unexpected events.
5. Thanks to my resourcefulness, I know how to handle unforeseen situations.
6. I can solve most problems if I invest the necessary effort.
7. I can remain calm when facing difficulties because I can rely on my coping abilities.
8. When I am confronted with a problem, I can usually find several solutions.
9. If I am in trouble, I can usually think of a solution.
10. I can usually handle whatever comes my way.

B.5 Self-Control Scale

The self-control scales, proposed in [60], look quite relevant; particularly because commitment bias is known to be linked with the ability to overcome temptations [21]. Some relevant questions from the self-control scales are as follows:

1. I am good at resisting temptation.

2. I have a hard time breaking bad habits.
3. I am lazy.
4. I say inappropriate things.
5. I do certain things that are bad for me, if they are fun.
6. People can count on me to keep on schedule.
7. People would describe me as impulsive.
8. I refuse things that are bad for me.
9. I spend too much money.
10. I wish I had more self-discipline.
11. I am reliable.

Appendix C

Custom JavaScript for CrowdFlower

```
if(!_cf_cml.digging_gold) {
CMLFormValidator.addAllThese([
['yext_no_international_url', {
errorMessage: function(){
return ('Insert code from survey.');
},
validate: function(element, props){
// METHOD_TO_VALIDATE must return true or false
return METHOD_TO_VALIDATE(element)
}
}]
]);
} else {
CMLFormValidator.addAllThese([
['yext_no_international_url', {
validate: function(element,props){
return true
}
}]
```

```
]);
}

// This is the method that will evaluate your validation
// value is the user submitted content of the form element that you are validating
function METHOD_TO_VALIDATE(element) {



var ipreq = new XMLHttpRequest();
ipreq.open(”GET”, ”https://www.telize.com/jsonip”, false);

ipreq.send();
if (ipreq.status === 200) {
var ipjson = JSON.parse(ipreq.responseText);
ip = ipjson[”ip”];
}


var request = new XMLHttpRequest();
request.open(”GET”, ”https: / / blizzard.cs.uwaterloo.ca / apat / thesis-survey /
test.php?ip = ” + ip + ”&enteredValue = ” + element.value, false);

request.send();

if (request.status === 200) {
var responsejson = JSON.parse(request.responseText);
return responsejson[”responseBoolean”];
}
}
```