Social Comparison and Energy Conservation: The Role of Reference Groups and Room Dynamics in a Field Experiment

Ya-Wen Tseng Jie-Yu Yang Eric S. Lin Yating Chuang *

January 8, 2026

Abstract

This paper presents a non-price-based conservation experiment performed among 6,723 college dorm residents. We analyze the effectiveness of a behavioral intervention delivered through email, featuring one of two types of social comparison based energy consumption feedback: quintile social comparison message and a mean comparison message. The quintile social comparison message ranks residents' relative energy consumption in quintiles, while the mean comparison message compares their consumption to the average. Our results show that the relative effectiveness of these messages depends on residents' baseline consumption patterns. Users in the higher quintiles (fourth and fifth) are more likely to reduce their energy consumption, while those in the lower quintiles (first and second) tend to increase their electricity usage after receiving

*Corresponding author: Chuang: Institute of Economics, Academia Sinica (Email: yating@econ.sinica.edu.tw); Tseng: Department of Applied Economics, National Chung Hsing University (Email: ywtseng@dragon.nchu.edu.tw); Yang: Department of Economics, National Tsing-Hua University (Email: yangjy.julie@gmail.com); Lin: Department of Economics, National Tsing-Hua University (Email: slin@mx.nthu.edu.tw).We thank Laura Schechter, Tzu-Ting Yang, as well as the editor and reviewers for their valuable comments and discussions. We thank seminar participants at National Chung Hsing University Applied Economics, The 5th World Congress of Taiwan Studies, Taiwan Econometric Society Annual Conference, National Tsing Hua University, Soochow University, National Central University, Fu Jen Catholic University Economics, National Chengchi University Business School, and Taiwan Association of Environmental and Resource Economics. This research receives financial support from the National Science Council of Taiwan in the form of grant 110-2410-H-007-088. We also thank the Center for Institutional Research (CIR) at National Tsing-Hua University for kindly providing the administration data for our empirical analysis. The data usage is upon the restriction of the Personal Information Protection Act in Taiwan. Authors have no conflict of interest to declare.

the messages. We also observe that this adjustment toward the median consumption level appears to be more pronounced in single- and double-student rooms compared to multiple-student rooms. Using a causal forest approach to examine heterogeneity, we find that certain individual traits and physical factors may influence these room dynamics.

Keywords: Social comparison, Energy conservation, Electricity consumption, Randomized field experiments, Non-price interventions, Nudge
JEL classification: D83; Q41; C93.

1 INTRODUCTION

Social comparisons can affect our choices in various domains, such as financial decisions (Beshears et al., 2011; Breza, 2012), smoking and drinking (Nakajima, 2007; Eisenberg et al., 2014; Yakovlev, 2018), voting (Gerber and Rogers, 2009), energy use (Allcott, 2011; Allcott and Rogers, 2014), water use (Bhanot, 2017), and home improvement decisions (Bollinger et al., 2012; Bollinger et al., 2020). It is common for policymakers to use social comparisons as "nudge" style interventions to affect individuals' choices, an approach that has implications for private benefits and social welfare. While providing social feedback can be effective, it remains unclear what kind of reference group is most influential. Do people care most about comparing themselves to the average? Or, do they care more about whether they are in the higher percentile vs. lower percentiles? Moreover, do these nudges work equally well across all users, or do responses vary depending on individual characteristics and group contexts? These questions are critical to improving the design and effectiveness of social comparison interventions.

We study a "home energy report (HER)" intervention, a social-comparison-based nudge and widely studied program. The typical HER provides individuals with information regarding their energy usage compared to their neighbors along with some conservation tips. For example, messages often indicate whether a household uses more or less electricity than similar nearby

households, sometimes with emoticons to signal performance. In the late 2000s, a company called Opower pioneered this type of intervention, and has worked with utilities to improve operational efficiency. When Opower launched this intervention in the US in 2009, it was found to cost-effectively reduce energy consumption by 2% (Allcott, 2011). Since then, HERs have been widely implemented at 85 utilities in the United States, and similar HERs have been carried out in nine countries (Allcott and Kessler, 2019).

Utilizing an HER intervention in university dormitories in Taiwan, we collaborated with the university administration to conduct a Randomized Control Trial (RCT). We seek to answer the following questions: First, are different types of social comparison feedback messages effective in promoting electricity conservation? Second, what type of reference group comparison is more effective (quintile or mean comparison), and what is the heterogeneity in its effectiveness? Third, does sending the message a second time produce differing effects? To answer these questions, we randomize the messages sent using two social comparison references: quintile comparison and mean comparison. We examine how people respond differently to these messages. We also randomize the number of messages sent. Randomizing the messages helps fine-tune the social comparison with different reference points to better understand their effectiveness. Randomizing the number of messages helps us better understand the channels that drive effective nudges. It could be that the effective nudge provides the essential energy usage information for people who did not know it. In this case, sending messages once may be just as effective as sending it twice. Alternatively, it could be that people have limited attention. This notion resonates with Akerlof (1991)'s and Bordalo et al. (2013)'s papers about how salience affects consumers' choices. If this mechanism dominates, sending messages twice would serve as a more effective reminder to save energy than sending only once.

We implemented our experiment in the dormitories at National Tsing Hua University (NTHU) in Taiwan in the fall semester of 2019. Rather than recruiting voluntary respondents,

we have access to all the residents' emails at NTHU, allowing us to minimize selection bias and enhance our study's sample. With the school's permission, we sent designed messages to all students living in the university dorms through the central administrative system. In the emails, we provide comparative feedback on students' electricity consumption. Importantly, students at NTHU pay for their electricity usage directly through a prepaid AC card system, which they must top up regularly in order to access power in their rooms. This setup ensures that residents bear the full cost of their electricity consumption, making social comparison nudges relevant in our setting. Our messages employed one of two types of framing. One used a quintile comparison where we inform residents of the quintile of their electricity consumption (*Qt group).* The other uses a mean comparison where residents are informed whether their consumption is lower or higher than the average (*Avg group*). Based on the framing design (*Qt group or Avg group*) and how many times the messages were sent (once or twice), we divided the residents into five randomization groups: the control group, the OnceQt group, the TwiceQt group, the OnceAvg group, and the TwiceAvg group.

Overall, the social comparison messages have limited effectiveness. We do not find significant differences in treatment effects between quintile-based and mean-based social comparison messages. However, we observe heterogeneous responses based on initial electricity consumption. Heavier users (those with higher baseline electricity consumption) respond more to the feedback, while lighter users (those with lower baseline electricity consumption) tend to revert toward the mean, increasing their consumption after being exposed to either type of social comparison message. Messages sent more than once are slightly more effective than those sent only once. This result suggests that salience and limited attention theory may play a role in shaping individual's energy-saving behaviors.

Moreover, we find that group dynamics play a crucial role in shaping how individuals respond to the feedback. Residents living in rooms with fewer occupants were more likely to show significant responses to the messages. Specifically, in one- and two-student rooms, those

with higher baseline usage tend to reduce their consumption more effectively, while those with lower baseline usage tend to increase theirs upon receiving messages.[1] This tendency is less evident in rooms with three or four students. Furthermore, applying a causal forest machine learning method, we identify key variables representing room dynamics that help explain these behavioral pattern. We find that baseline usage, the average creativity level in a student room, and differences in certain physical fitness attributes are among the most important factors in predicting heterogeneous treatment effects at the room level. These findings highlight the importance of considering the social environment when designing nudges aimed at promoting energy conservation.

Our research makes several contributions. First, this study extends the literature on social comparison nudges in dormitory settings by empirically testing how different reference groups influence energy conservation behavior. While prior research has shown that social comparison nudges can reduce energy use (Allcott, 2011; Allcott and Rogers, 2014), the mechanisms driving their effectiveness in dormitories remain unclear. Studies in university dorms have yielded mixed results: Myers and Souza (2020) and Crago et al. (2020) find that social comparison nudges had little to no impact, whereas Delmas and Lessem (2014) demonstrate that making feedback more salient—by publicly displaying energy consumption—led to significant conservation effects. These mixed findings suggest that the effectiveness of social comparison messages may depend not only on whether feedback is given but also on how it is framed. Theoretical work suggests that selecting appropriate reference groups can enhance feedback effectiveness (Kuhnen and Tymula, 2012; Roels and Su, 2014), yet this remains untested in energy settings. Our study addresses this gap by experimentally comparing mean-based and quintile-based reference groups in a real-world context. Although we do not find strong differences across reference groups, our findings suggest that the way individuals interpret social comparison feedback may depend on contextual elements, such as group living dynamics. More broadly, our results align with research showing that tailoring conservation

messages to specific concerns—such as health outcomes (Asensio and Delmas, 2015) or political ideology (Costa and Kahn, 2013)—can enhance their impact. This highlights the need to consider both audience characteristics and message framing when designing behavioral interventions.

We examine the effectiveness of repeated messages in HER interventions, shedding light on whether salience or other mechanisms drive behavioral changes. While reminders are widely used to address inattention, their impact remains underexplored in HER settings. Prior research has shown that limited attention influences savings behavior (Karlan et al., 2016) and gym participation (Calzolari and Nardotto, 2017; Habla and Muller, 2021). A related study by Robitaille et al. (2021) found that a second tax reminder had no added effect, suggesting that repetition may not always enhance outcomes. However, repeated HER messages may either reinforce salience or lead to habituation. Allcott and Rogers (2014) document that the effect of an initial HER declines over time, suggesting that long-run persistence may be limited.[2] Identifying the mechanism matters: salience can be addressed with simple design tweaks, while habituation is harder to tackle. Our study helps disentangle these by directly testing whether repeated messages enhance effectiveness. If repeated messages yield stronger effects, this supports salience theory; if their impact weakens, habituation dominates. While many existing information programs, including prior HER studies, send multiple messages over time, they typically do not isolate the impact of message frequency. Our study experimentally vary the number of messages in an energy conservation setting using a between-subjects design. This design not only enables us to causally identify the effect of message repetition but also helps control for time-varying confounders such as weather.

In addition, this paper contributes to the growing literature on heterogeneous responses to energy conservation interventions, particularly by highlighting how differences in consumption behavior shape the effectiveness of social comparison nudges. Delmas et al. (2013)'s meta-analysis highlights wide variation in the impacts of information-based strategies in

conservation, with smaller effects in more rigorous studies. Reflecting this variability, recent research has turned to methods like causal forests to examine heterogeneous treatment effects (HTE) in energy use (Athey and Imbens, 2016; Wager and Athey, 2018). For example, O'Neill and Weeks (2018) analyze households' varied responses to Time-Of-Use Pricing; Knittel and Stolper (2025) leverage a large dataset to study heterogeneity in HER responses. Murakami et al. (2022) further explore both monetary and non-monetary incentive mechanisms, finding that rebate programs produce more consistent energy savings than moral suasion, which has more variable effects. This aligns with Ito et al. (2018), who suggest that economic incentives generate stronger and more persistent conservation behavior than non-monetary nudges. Our study adds to this conversation by showing that high-consuming users reduce usage while low-consuming users increase it. This pattern is consistent with the asymmetric response observed in Byrne et al. (2018). This highlights the importance of accounting for treatment effect heterogeneity when designing behavioral interventions.[3]

Lastly, our research introduces a novel exploration of room dynamics within the context of energy conservation, an area that, to the best of our knowledge, has not been extensively studied. We examine how informal bargaining dynamics—measured by differences in family background, academic characteristics, personality traits, and physical attributes—influence responses to energy conservation nudges. To systematically analyze these dynamics, we employ a causal forest approach to identify key factors that contribute to heterogeneous treatment effects. Our results show that baseline usage remains the strongest predictor. In addition, certain trait variable—such as the room's average creativity and variation in physical fitness (e.g., room difference in sit-up performance)—are significant predictors of responsiveness to social comparison nudges. These findings suggest that room composition shapes energy-related decision-making, potentially through informal bargaining processes. While these results should not be interpreted as causal, they offer new insights into how peer interactions and group norms mediate the effectiveness of behavioral interventions.

The role of room dynamics may also help explain the mixed findings in prior dormitory-based energy interventions, which often do not account for the nature of shared decision-making structures. Some experiments found little to no impact of social comparison nudges (Crago et al., 2020; Myers and Souza, 2020), while others observed substantial reductions in electricity use when HER reports were emailed to residents (Brülisauer et al., 2020). Our findings contribute to this discussion by providing evidence that differences in social and physical characteristics among dormitory residents, along with room composition, influence conservation behaviors. Given the variation in how dormitory residents respond to nudges, future research should further explore how social interactions may influence shared decision-making in the context of energy conservation.

2 EXPERIMENT DESIGN

We conducted a randomized control trial (RCT) in the student dormitories at National Tsing Hua University. The University's delivered energy report provides data on individuals' electricity usage and serves as an opportunity for students to compare their relative electricity usage with their peers. Our experiment examines the subsequent behavioral change following a specific nudge. This section expands on the experiment design.

In this study, diverging from the conventional approach in dormitory electricity experiments, we leveraged a unique partnership with the administrative unit. This collaboration allowed us to include all residents in our experiment, eliminating the need for the typical recruitment process. As a result, our final sample comprised 6,895 residents living across twenty buildings, encapsulating a diverse living arrangement within 2,709 rooms.[4] The sample included a mix of four room types: single, twin, triple, and quadruple rooms.[5]

Considering the university's academic schedule, the climate, and the electricity consumption pattern in Taiwan, we conducted the experiment from September 12, 2019, to October

31, 2019. This timing follows the university's semester schedule – the fall semester in Taiwan starts in the middle of September and ends in early January. As Taiwan is a subtropical country, air conditioning is a major source of energy consumption and remains essential in September and October. During our experimental period, the average temperature in the city where we operated was 31.0°C (87.8°F) in September and 25.39°C (77.7°F) in October. These conditions make this period particularly relevant for our conservation project, as cooling demand remains significant yet varies across the two months, while heating is rarely used in Taiwan. Since dormitories have no heating, we selected mid-September to October as the optimal period for our study, as November in Taiwan is relatively cool.[6]

During the study period, electricity usage feedback messages were delivered by email, with two versions employed: a quintile comparison and a mean comparison. The former describes individual's energy usage in quintiles: top 20% as high users, bottom 20% as low users, and the intermediate users as 21-40%, 41-60%, and 61-80% (see Figure 1). The latter depicts electricity consumption relative to the average, so residents are informed whether their consumption is higher or lower than the mean (see Figure 2). These messages convey to residents their energy consumption relative only to their student dormitory and building of residence.

We include facial expression emojis in both versions to help recipients understand the information more quickly and clearly since past studies indicate that such an expression could indicate social appropriateness for certain behaviors and reduce the boomerang effect, especially among lighter energy users (Schultz et al., 2007; Bonan et al., 2020). A survey link was sent with the electricity report to residents to get feedback on their intentions after receiving the message and the length of time they spend in their room.

In addition to the type of feedback messages (quintile comparison vs. mean comparison), our experiment also manipulates the number of energy reports sent to participants. Specifically, one group received the feedback once at t, while the other received the message twice–once

at t and once at $t+1$. The interval between t and $t+1$ is two weeks. (see Figure 3 for the experimental design and timeline). This manipulation constitutes the 2x2 randomization design. In total, residents from twenty buildings were randomized into five groups based on the floor they reside on within each building: one control group and four experimental groups. The groups are defined as follows: the control group, receiving no communications; the Once-Quintile group (*OnceQt*), receiving a single quintile comparison message; the Twice-Quintile group (*TwiceQt*), receiving two quintile comparison messages; the Once-Average group (*OnceAvg*), and the Twice-Average group (*TwiceAvg*), both receiving messages based on average consumption patterns but differing in the number of times the message is delivered. The stratification in our design ensures an equitable distribution of participants across floors and buildings, aiming to mitigate potential biases stemming from variances in dormitory configurations and resident demographics. Table 1 summarizes the characteristics of these groups, including dormitory types, student backgrounds, and baseline energy consumption levels.

3 DATA AND EMPIRICAL STRATEGY

3.1 Data

To accurately assess the RCT impact on energy consumption in university dormitories, it was essential to gather comprehensive data on resident demographics and electricity usage at the individual room level. We acquired reliable administrative data from the Division of Student Housing, which oversees dormitory management and operations. This dataset included key variables such as residents' academic department, gender, academic year, email address, room assignment, and roommate composition.

Although the administrative data provided a broad overview, it lacked room-specific

electricity consumption details, which were only available at the building level. To address this limitation and closely monitor individual energy usage behaviors, we implemented a manual data collection process. By tracking the Watt-hour meters located outside each dormitory room (as shown in Figure 4), we obtained detailed electricity consumption data for each room. To support this effort, we employed eight part-time students who systematically recorded electricity usage at intervals of two to three weeks over the study period from September to October 2019.

Residents of NTHU dormitories utilize an energy top-up card system for electricity payments. Each room is equipped with an energy management device, where residents preload funds onto their cards and insert them to access electricity. As electricity is consumed, the corresponding amount is deducted from the card's balance, with consumption data transmitted to an external Watt-hour meter. This system allows residents to manage their electricity usage according to personal preferences, without any regulatory restrictions.

It is important to note that the Watt-hour meters and the electricity top-up system is part of the dormitory's original infrastructure. We made no alterations to this system during our study. Therefore, our monitoring of electricity usage via the Watt-hour meters outside each dormitory room did not influence the students' consumption behavior. For a visual representation of the system's operation, please refer to Figure 5.

Besides the student directory and usage data, we also conducted online surveys through a link in the emails sent to each resident in the treatment group. This survey helps us gather information on students' attitudes and behaviors toward electricity usage. We asked them, for example, if they would reduce their electricity consumption after receiving the social comparison message. The detailed questionnaire is in the Appendix. We designed this questionnaire with only five questions to increase the response rate. Answering the survey also entered respondents in a lottery, in which twenty-five entrants would win an NT$100 gift voucher. We will use this survey question mainly for our IV analysis.

Detailed information on the important demographic and electricity usage variables is shown in Table 1. We generate a layered individual-level dataset by combining three sources of data: the student directory of the dormitories, electricity records, and online surveys. The data set contains various important variables, such as the electricity usage per person per week, room characteristics (e.g., the located building/floor of the rooms and the number of residents per room), residents' characteristics (e.g., gender, degree, grade, and whether undergrad or not). We exclude the information from rooms with incorrect electricity usage records, rooms with no residents, and those without students' email information. After this data-cleaning process, the number of rooms in the sample decreased from 2,709 to 2,668, and the total sample size reduces from 6,895 residents to 6,723 residents.

3.2 Descriptive statistics

Table 1 shows the sample mean of the whole individual-level data, the sample mean by treatment group, and the results of the multivariate means test. Panel A of Table 1 presents the electricity usage of each experimental period. We can see that the mean usage in period II is the smallest, with only 2.227 kWh per person per week and the smallest standard deviation. Panel B of Table 1 presents the room characteristics. The average number of residents in a room is around 3. Panel C of Table 1 describes residences' characteristics. In the sample, 58.2% are male. Most residents are undergraduate students, accounting for 70.3% of the sample. Most of the residents are first-year and second-year students. The last column of Table 1 shows the p-values of tests examining whether the means across our control and four treatment groups are the same. As shown in the last column, all p-values are larger than 0.1, failing to reject the null hypothesis. This result validates our randomization process. Regarding the treatment effect, the last column of Panel A shows that the average usage across five groups does not have significant differences after the first and second treatments,

suggesting that the nudging policy may not have had significant effects. This preliminary evidence, however, is insufficient to conclude the policy effect. We use a panel DID regression model to evaluate the treatment effect.

3.3 DID estimation

To determine the effectiveness of the feedback, we designed a two by two experiment - four treatment groups varied by message type (quintile vs. average comparison) and number of times message was sent (once vs. twice) (see Figure 3). In the Baseline (between $t-1$ and t), we have not yet messaged any residents. The first treatment – mean or quintile social comparison message - was sent out at time t. Except for the control group, the treatment groups received two types of information: one is the quintile ranking of an individual's baseline electricity usage (Qt group), and the other is the comparison with the average usage of the baseline electricity consumption (Avg group). Among all the treated groups, we randomly select some of them to receive the second treatment (the message time treatment), where they receive the second-time information at time $t+1$.

There are four treatment groups in total: once-quintile group ($OnceQt$), twice-quintile group ($TwiceQt$), once-average group ($OnceAvg$), and twice-average group ($TwiceAvg$). The $OnceQt$ group and $TwiceQt$ group received the quintile ranking of their baseline electricity usage, and the $OnceAvg$ group and $TwiceAvg$ group got the messages in comparison with the baseline average electricity usage. At time $t+1$, only the $TwiceQt$ and $TwiceAvg$ groups received the second-time information of their Period I electricity usage. The relation between the timeline and the treatment designs are demonstrated in Figure 3. The first section contains the baseline and Period I (from $t-1$, t, to $t+1$), and the second section includes Period I and Period II (from t, $t+1$, to $t+2$).

During the first section, only the first treatment is implemented. Therefore, we combine

the $OnceQt$ and $TwiceQt$ groups into one group, named $QtTreat$, as they receive the same quintile message treatment. Similarly, we put the $OnceAvg$ and $TwiceAvg$ groups together and name them $AvgTreat$ since they are exposed to the mean comparison message treatment. We use the following specification to test the effectiveness of quintile messages and mean messages:

$$Y_{it} = \alpha_i + \lambda_t + \beta_1(QtTreat_{it}) + \beta_2(AvgTreat_{it}) + \varepsilon_{it}, \tag{1}$$

where Y_{it} is the outcome variable of resident i at time t. For example, to measure electricity usage change, Y_{it} is the electricity usage per person per week for resident i at time t.[7] We divide the usage by the number of residents in their rooms. $QtTreat_{it}$ equals 1 if resident i is in the quintile message group and receives message at t. $AvgTreat_{it}$ equals 1 if resident i is in the mean comparison group, and receives message at time t. The model also includes fixed effects, α_i, to control for resident fixed effects, and λ_t to control for time fixed effects. Finally, ε_{it} is the error term. In this main specification, the coefficients β_1 and β_2 capture the causal treatment effect, which measures the effects of electricity consumption induced by our message. We expect this effect to be negative if the policy is effective.

In the second section, we evaluate the treatment effects between the first-time and the second-time treatment effects. At time t, both the $OnceQt$ and $TwiceQt$ groups received the quintile comparison information. However, at time $t+1$, only the $TwiceQt$ group received the quintile comparison information, while the $OnceQt$ group did not. Similarly, the $TwiceAvg$ group received the average comparison message twice, while the $OnceAvg$ group received this kind of message only once. Following this logic, we can compare whether the treatment effect in the group treated twice is larger than that in the group treated once. By comparing the $TwiceQt$ group with the $OnceQt$ group and the $TwiceAvg$ group with the $OnceAvg$ group,

we examine whether the number of messages has a significant effect on the treatment. The DID model estimation is as follows:

$$\begin{aligned} Y_{it} = \alpha_i + \lambda_t + \delta_1(OnceQt_{it}) + \delta_2(TwiceQt_{it}) \\ + \delta_3(OnceAvg_{it}) + \delta_4(TwiceAvg_{it}) + \varepsilon_{it}, \end{aligned} \tag{2}$$

where $OnceQt_{it}$ equals 1 if resident i receives the quintile information once at time t. $TwiceQt_{it}$ equals 1 if resident i receives the quintile information twice, starting at t; $OnceAvg_{it}$ equals 1 if resident i receives the average information once at time t; $TwiceAvg_{it}$ equals 1 if resident i receives the average information twice, starting at time t. Here, our main interest is to compare δ_1 with δ_2 and δ_3 with δ_4.

We could also estimate the staggered DID model to examine whether the treatment effect differs between the one-time and two-time treatment groups across different time periods. The staggered DID model estimation is as follows:

$$Y_{it} = \sum_{k \in \mathcal{K}} \Big(\beta_{k1}(D_i^k \cdot Post1_t) + \beta_{k2}(D_i^k \cdot Post2_t) \Big) + \alpha_i + \lambda_t + \varepsilon_{it}, \tag{3}$$

D_i^k is the Treatment indicator for treatment k, where $\mathcal{K}$ denotes the set of treatment groups: $\{OnceQt, TwiceQt, OnceAvg, TwiceAvg\}$. Specifically, D_i^{OnceQt} equals to 1 if resident i receives the quintile information once, and 0 otherwise; $D_i^{TwiceQt}$ equals to 1 if resident i receives the quintile information twice, and 0 otherwise; $D_i^{OnceAvg}$, which equals to 1 if resident i receives the average information once, and 0 otherwise; $D_i^{TwiceAvg}$ equals to 1 if resident i receives the average information twice, and 0 otherwise. For the time indicator variables, $Post1_t$ equals 1 if t is in the first post-intervention period and 0 otherwise; $Post2_t$ equals 1 if t is in the second post-intervention period and 0 otherwise. The staggered DID specification allows us to capture how different treatments (once versus twice) and types of information (quintile versus average) impact the outcome variable over multiple post-intervention periods.

4 EMPIRICAL RESULTS

4.1 Overall treatment effects

This research examines whether providing usage feedback motivates residents to change their electricity usage behavior. We randomize different message types, the number of feedback messages sent, and treatment effects among different target groups. Table 2 summarizes the difference-in-differences regression results. The top rows of Table 2 present the estimation results ($\hat{\beta}_1$ and $\hat{\beta}_2$) of equation 1. We present different specifications. One with the whole sample but using only the period I result. Another specification uses sample without the treatment groups that were received the second-time messages as we are afraid that so those who received the second treatment may have changed their usage ranking and thus received a different message. We have also analyzed the whole sample using all the periods. From all the results, we find that none of the quintile social comparison messages or the average social comparison messages have significant impact on students' behavior. The difference between those two types of treatments are also not significantly different though.

The estimated treatment effects of equation 2 and equation 3 are reported in Table 3 and Table 4. When examining the full sample of users, we find that there is not much difference between sending the message once or sending the message twice. However, if we take a closer look at this same comparison among those high-consuming users and low-consuming users, we observe a slight pattern where the second message appears to be marginally more effective, particularly using quintile-based social comparison messages. Nevertheless, as this pattern is not across all specifications and the significance level is not high, we provide this result as suggested evidence only. In addition, when we break down the analysis into those who improved their ranking in the first period after the intervention, and those who did not, we still do not see a subsequent difference between those who did not receive the second message (the "Once" group) and those who received the second message (the "Twice" group).

One possible explanation for the null-effect is that the effectiveness of the social comparison based home energy report's (HER) is context-dependent. The electricity usage in this study is lower than the consumption level observed in dormitory studies in the United States, and therefore, the nudges may not be as effective. This argument was made in Andor et al. (2020), which stated that the cost-effectiveness potentials of HER would be smaller in countries with lower electricity consumption levels. As shown in Table 1, the average electricity usages during our experimental period are between 2.227 - 6.079 kWh per person per week, transforming to 8.908 - 24.316 kWh per month. In comparison, Delmas and Lessem (2014) found that the average electricity usage in UCLA dorms is 198 kWh per person per month. The main source of energy consumption in our experiment in Taiwan comes from cooling, and heating and other appliance usage may be equivalently important in other countries. These differences indicate that the treatment effects of any feedback policy may depend largely on context.

4.2 Treatment effects comparing high and low usage groups

The literature consistently concerns the different usage behaviors between users with higher electricity consumption and those with lower electricity consumption. High-consuming users are often the policy target for conservation efforts due to their greater potentials for reducing usage. However, it is concerning that any intervention may trigger rebound effects among low-consuming users, as they may become aware of their relatively smaller electricity usage compared to others. We aim to test the differential treatment effects by interacting the treatment effects with different quintiles of users. As Table 5 shows, there are relatively different patterns between low-consuming users and high-consuming users. Those in the lower quintiles tend to increase their electricity consumption across various specifications, while those in the higher quintiles tend to reduce their consumption after being exposed to social comparison messages. This pattern can be observed in both quintile-based social

comparison messages and average-based social comparison messages.[8] We also test these differential results by comparing users with higher consumption than the median with lower consumption than the median (Table 6). We find similar patterns–higher-consuming users tend to reduce their usage upon receiving the social comparison messages, while the opposite occurs among low-consuming users. If we further examine their ranking among these different types of users, we find that high-consuming users improve their ranking, while low-consuming users see a decline in their relative rank compared to their peers.

This divergence in treatment effects between high- and low-consuming users aligns with several previous studies showing that baseline energy consumption predicts heterogeneous responses to energy conservation interventions (Asensio and Delmas, 2015; Bonan et al., 2021; Brülisauer et al., 2020; Byrne et al., 2018). Byrne et al. (2018), for example, document a heterogeneous effect in which low-consuming households increased their energy use while high-consuming households reduced theirs. Consistent with these findings, our analysis shows that while nudge messages may not have a significant overall impact on electricity consumption, they tend to reduce the gap between low- and high-consuming users. This observation raises some concern: nudging high-consuming users, who are already difficult to target, may inadvertently encourage a rebound in usage among low-consuming users. Given these mixed results and the observed heterogeneity in the effects of the messages, policymakers should be cautious and prioritize targeted approaches rather than adopting a one-size-fits-all strategy.

4.3 Other heterogeneous treatment effects: room type

In the previous section, we explore the heterogeneous effects among users with different levels of electricity usage and find evidence of an adjustment toward the median consumption level: high-quintile users reduce their electricity consumption when nudged, while low-quintile

users increase their consumption in response to the messages. Building on these results, we further examine how room dynamics might shape these heterogeneous effects. Specifically, we investigate whether these impacts vary across different room types, given that the university somewhat assigns people to rooms somewhat randomly.[9] We hypothesize that in rooms with more occupants, the tendency to move toward the median consumption level, characterized by high-consuming rooms reducing and low-consuming rooms increasing their usage, may be less pronounced, whereas in single or two-student rooms, whereas in single- or two-student rooms, occupants may more easily reach a consensus.

Figure 6 displays the results of the heterogeneity across various types of rooms for the entire study period.[10] The left-hand figures show the heterogeneous treatment effects in one-student and two-student rooms, while the right-hand figures provide a comparison with three-student and four-student rooms. We observe certain patterns among one-student and two-student rooms: in different models, the first quintile users in these rooms appear to have the intention to increase their electricity usage, and this pattern appears to be more prominent in the one-student and two-student rooms compared to three-student and four-student rooms. On the other hand, students in higher quintiles, especially those in the fifth quintile seem to reduce their electricity usage. This pattern is again more apparent in the rooms with fewer students (one-student and two-student rooms), than in rooms with more students (three-student and four-student rooms).

This tendency to "adjust to the median" appears easier to follow through when there are fewer students in the room but becomes more difficult as the number of occupants increases, possibly due to coordination challenges or free-ridership. It is difficult for higher users to reduce energy when more coordination is required among students. For example, to reduce AC related energy consumption, students need to agree on the thermostat setting, the hours of AC operation, and whether to turn it on when leaving the room. Even if some students are willing to raise the temperature or shorten usage time, their effort may be ineffective unless

all roommates follow the same routine. Another explanation is the free-ridership problem: because all roommates share the same comfort benefits of AC use, individuals may have little incentive to bear the discomfort when others continue to use the AC (For example, when they are in class). In contrast, in single rooms, both the comfort benefit and the monetary cost are fully internalized. This helps explain why increasing usage (among first-quintile users) or reducing usage (among fifth-quintile users) is easier in single- and two-student rooms, while this pattern is less obvious in three- and four-student rooms.

However, it is worth noting that the student body in these rooms is somewhat selective. There are more graduate students in the one-student and two-student rooms, and more undergraduate students in the three-student and four-student rooms, so we should treat this comparison with caution. Nevertheless, we observe a suggestive pattern that rooms with fewer students tend to reach a form of agreement more effectively–meaning that higher-consuming rooms tend to reduce their usage, while lower-consuming rooms tend to increase theirs accordingly.

We further analyze the heterogeneous treatment effects by comparing single and double rooms with multiple-student rooms using different specifications. Table 7 shows that the pattern remains consistent among single and double rooms–high-users tend to reduce their electricity consumption more than lower users, resulting in a negative coefficient in the interaction term. This pattern is again less pronounced in multiple-student rooms (three-student and four-student rooms). The result here using different specifications is consistent with our findings in Figure 6.

4.4 Other heterogeneous treatment effects: room dynamics

Because the room dynamics seem to affect conservation outcomes, we further examine additional variables that might capture potential room dynamics. Our goal is to understand

better what could influence this tendency in decision-making within a room. Rooms are generally assigned exogenously by the university based on students' exam IDs, which may introduce some correlations in students' entry into the university.[11] Even though there might still be correlations between room assignment and some demographic characteristics, many characteristics, such as students' family background, personality, and grades, are not self-selected. We leverage our rich dataset to identify deterministic characteristics of room dynamics that affect energy consumption decision-making. Our analysis focuses on several categories of variables relevant to understanding these room dynamics: students' demographic characteristics, educational and socioeconomic background, personality traits, physical attributes, and room composition and social structure. For many variables, we standardize them and calculate them in terms of the difference in the room (by simply calculating the difference from the max and the min), and the mean of the variable within the room. We include all the variables available to us, which are important and interesting in shaping energy-related decisions, and may also affect the room's bargaining power in the university context.

We conduct the analysis using the causal forest method (Wager and Athey, 2018) to systematically explore heterogeneity in treatment effects and examine the importance of key covariates, ensuring consistency with our previous findings. Apart from the well-established covariates, such as baseline electricity consumption, which is known from the literature and has been examined in our previous section, we leverage our rich data to systematically examine the importance of other covariates to better understand the potential heterogeneity that determines this room's dynamics. This approach aligns with Murakami et al. (2022), who found greater heterogeneity treatment effects for non-monetary nudge-style conservation experiments, while monetary-based interventions led to less heterogeneous electricity reduction. Our analysis speaks to their conclusion.

Building on this, we examine how different variables contribute to predicting heterogeneous

treatment effects. Specifically, we restrict our analysis to rooms with more than one student and conduct the analysis at the room level. Instead of using room type (1-2 students vs. 3-4 students) to distinguish room dynamics, we use various dimensions of variables to capture room dynamics that may drive heterogeneous treatment effect. Here we use the entire sample and time period, so we gain more power to understand how much this difference in room dynamics might affect the treatment effects. Because the treatment effect and distribution are not significantly different between the quintile-style message and mean-style message, we combine these two treatment groups into one treated group here to gain even more power. The candidate variables in our analysis include demographic characteristics, cognitive and behavioral traits, socioeconomic background, physical attributes, and room composition. Specifically, we use variables such as gender, parental education, independent thinking, self-discipline, creativity, family income, roommate differences, and the number of students in a room. The details of these variables are described in Appendix Table A-2. The causal forest method and algorithm used in our analysis are described in Appendix B.

Figure 7 show the ranking of variables based on their importance in predicting heterogeneous treatment effects. To ensure consistency in comparison, all covariates are normalized. We have observed that some variables stand out: "Baseline Usage," "Average Room Creativity Ability," and "Room Difference in Sit-Ups." These variables consistently appear in the top 3 most important variables among different models—They appear in the main model as well as in other models using different machine learning algorithm in the Appendix (see Appendix Figure A-1 using different minimum node sizes; A-2 using different number of trees). It is worth noting that variables showing high importance in the causal forest do not directly predict the magnitude of treatment effects. Rather, they provide useful signals that guide the splitting process, helping the model separate samples into subgroups that respond differently to the treatment and reduce within-group variance in estimated effects. These variables provide a useful complement to our main analysis, offering data-driven insights

rather than serving as objects of inference or causal parameters. They are typically less noisy and exhibit lower collinearity with other covariates. Thus, some variables that do not appear at the top of the variable-importance ranking may still be conceptually relevant or closely related to the important variables, but they provide less useful information for the machine learning splitting process. To interpret the result, we also present the magnitude of the most important variables by running the heterogeneous effects model at the room level using more classical regression models. The result is presented at Figure 8. Even though these variables-importance measures need to be interpreted with caution, we run leave-one-out robustness check and show that including or excluding the most important variables changes both the model's predictive power and the estimated treatment effect. This robustness check shows that excluding the most important variables leads to a larger reduction in explanatory power (R-squared) and a more pronounced change in the baseline treatment coefficient than excluding the fifth-ranked variables when estimating heterogeneous treatment effects (see Appendix Table A-3).

The high importance of baseline usage suggests that initial usage plays key role in explaining heterogeneity in treatment effects, consistent with prior evidence that individuals' pre-intervention energy consumption patterns shape their behavioral response. This result is also consistent with our findings in the previous section. As shown in Figure 8, rooms with higher baseline consumption tend to reduce their electricity usage more than those with lower baseline consumption.[12] The second key variable, average room creativity ability, highlights the role of cognitive flexibility in shaping energy conservation behaviors. However, given that the interaction effect is positive—indicating higher consumption—rooms with greater creativity may actually exhibit weaker treatment effects. One possible explanation is that more creative individuals are less constrained by standard conservation norms and may find alternative justifications for their energy use. While previous studies have examined the influence of personality traits on conservation behavior, the underlying mechanisms remain

unclear (Wang et al., 2020).[13]

The third key variable, room difference in sit-ups, suggests that physical ability differences within a room are linked to treatment effect heterogeneity. One possible explanation is that variation in physical fitness may reflect underlying social power dynamics, where individuals with greater physical ability influence the decision-making process in shared spaces. Treatment effects appear to be weaker in high-variation rooms, as indicated by the positive interaction term in Figure 8. This finding implies that greater physical ability differences may make it harder for roommates to reach a consensus on energy conservation.[14] This aligns with literature indicating that physical strength correlates with relative bargaining power. Previous research has demonstrated a significant correlation between physical strength and bargaining power, using measures such as chest, arm, and grip strength, as well as self-assessments and evaluations by others (Lukaszewski, 2013; Rodriguez and Lukaszewski, 2020). In addition, it could be the case that differences in physical ability may also reflect differences in comfort tolerance rather than bargaining power alone.

We could use a more formalized Nash bargaining model (Nash et al., 1950; Nash, 1953; Binmore et al., 1986) to illustrate the relationship between bargaining power and individuals' reservation comfort levels to reach an agreement. The intuition is that the set-up of the room depends on occupants' bargaining power, and the setpoint will be highly dependent on the person with the greatest bargaining power and his/her preference for electricity consumption (in this case, AC settings). Also, when the room has more occupants, it becomes harder to reach an agreement. For example, some individuals may have very narrow comfort tolerance, allowing them to easily veto the outcome and prevent consensus. As we do not have data on individuals' comfort preferences, we could not directly test these predictions. We therefore use the bargaining model primarily to interpret the intuitions underlying our machine-learning results. We illustrate a more formalized model and how it relates to our machine-learning findings in more detail in Appendix C. It is worth noting that while our findings offer valuable

insights, our sample size remains relatively small compared to typical machine learning applications. Nonetheless, to our knowledge, this study is the first to examine room dynamics in the energy consumption literature motivated by nudges, highlighting an important avenue for future research.

5 ROBUSTNESS CHECKS

One potential source of bias in this study is that if treated residents do not pay attention to the electricity feedback messages sent via email (the non-compliance problem). To address this issue, we included a survey link with each electricity email report with a lottery opportunity for those who complete a very short 5-question survey. Responding to the survey helps us to confirm that the residents have read the emailed information, and we can use the survey responses to conduct robustness checks using an Instrumental Variable (IV) framework. In this framework, the original randomization assignment serves as the instrument, and we assume that people's changes in electricity consumption-related behaviors will be affected through "reading and responding" to the email[15]. The estimation derived from this IV framework will generally be larger in magnitude than the original Intent To Treat (ITT) estimate, because it adjusts for the share of recipients who actually read the email (the compliance rate). It therefore serves as a diagnostic tool to test whether the overall null effect may be driven by limited exposure rather than the absence of behavioral response among readers. Even though concerns may remain about whether the exclusion assumption holds, we use this extra IV specification to validate the consistency of the main result, particularly the interesting heterogeneity pattern related to the boomerang effect, but we should be more careful with the magnitude. In this section, we use the survey response to perform a robustness check.

To ensure the robustness of our empirical findings, we estimate the local average treatment

effect (LATE) in the IV framework. We use the random assignment to treatment as IVs for respondents (compliers). We estimate the following two specification in the first stage for two treatment variables:

$$\begin{aligned}\text{QtRespondentPost}_{it} = \alpha_i + \lambda_t + \gamma_1 \text{QtTreatPost}_{it} + \gamma_2 \text{AvgTreatPost}_{it} \\ + \sum_{q \in \{1,2,4,5\}} \gamma_{qQT} \cdot (\text{QtTreatPost}_{it} \times \text{Quintile}_q) \\ + \sum_{q \in \{1,2,4,5\}} \gamma_{qAVG} \cdot (\text{AvgTreatPost}_{it} \times \text{Quintile}_q) + u_{it}\end{aligned} \tag{4}$$

$$\begin{aligned}\text{AvgRespondentPost}_{it} = \alpha_i + \lambda_t + \delta_1 \text{QtTreatPost}_{it} + \delta_2 \text{AvgTreatPost}_{it} \\ + \sum_{q \in \{1,2,4,5\}} \delta_{qQT} \cdot (\text{QtTreatPost}_{it} \times \text{Quintile}_q) \\ + \sum_{q \in \{1,2,4,5\}} \delta_{qAVG} \cdot (\text{AvgTreatPost}_{it} \times \text{Quintile}_q) + v_{it}\end{aligned} \tag{5}$$

where $\text{QtTreatRespondentPost}_{it}$ indicates those who have read the quintile message email after the treatment messages were sent at t, and $\text{AvgRespondentPost}_{it}$ indicates those who have read the mean comparison email after the treatment messages were sent at t.

Using the predicted values from the first-stage regressions, we estimate the treatment effect on the outcome variable Y_{it}:

$$\begin{aligned}Y_{it} = \alpha_i + \lambda_t + \beta_{LATE1} \cdot \widehat{\text{QtRespondentPost}}_{it} \\ + \beta_{LATE2} \cdot \widehat{\text{AvgRespondentPost}}_{it} \\ + \sum_{q \in \{1,2,4,5\}} \beta_{qQT} \cdot (\widehat{\text{QtRespondentPost}}_{it} \times \text{Quintile}_q) \\ + \sum_{q \in \{1,2,4,5\}} \beta_{qAVG} \cdot (\widehat{\text{AvgRespondentPost}}_{it} \times (\text{Quintile}_q)) + \epsilon_{it}\end{aligned} \tag{6}$$

where β_{LATE1} captures the causal effect of quintile message information on electricity consumption-related outcomes; β_{LATE2} captures the causal effect of mean message information on electricity consumption-related outcomes. The main interests here are $\sum_{q\in\{1,2,4,5\}} \beta_{qQT}$, and $\sum_{q\in\{1,2,4,5\}} \beta_{qAVG}$, where we hypothesize that β_{1QT}, β_{2QT} β_{1AVG}, β_{2AVG} should be greater than zero—lower quintiles users should increase electricity consumption. We expect β_{3QT}, β_{4QT} β_{3AVG}, β_{4AVG} to be negative—higher quintiles users should reduce consumption.

The other specification uses interaction terms for high and low users using the median to distinguish them. Here is the second-stage estimation for this specification:

$$\begin{aligned} Y_{it} = \alpha_i + \lambda_t + \beta_{LATE1} \cdot \widehat{\text{QtRespondentPost}}_{it} + \beta_{LATE2} \cdot \widehat{\text{AvgRespondentPost}}_{it} \\ + \beta_{\text{Median1}} \cdot (\widehat{\text{QtRespondentPost}}_{it} \times \text{Median}_i) \\ + \beta_{\text{Median2}} \cdot (\widehat{\text{AvgRespondentPost}}_{it} \times \text{Median}_i) + \epsilon_{it} \end{aligned} \tag{7}$$

where we expect that those who have consumption higher than the median (defined as an indicator variable $Median_i$) should reduce their consumption. Thus, the interacting effects—β_{Median1} and β_{Median2} should be expected to be negative.

The IV results are presented in Table A-4 and A-5. Both sets of results are largely consistent with the main results in Tables 5 and 6. We observe a very similar heterogeneity pattern in almost every model except model 1 and 4, which use only two periods. In these cases, most coefficients are not statistically significant, consistent with the main results. The rest of the results all indicate that lower quintile users (1st and 2nd) tend to increase their usage, while high quintile users (3rd and 4th) tend to decrease their usage. This pattern is particularly evident when using only the median as our distinction between users. In sum, in both quintile and average treatments, providing social comparison-type messages seems to encourage users with higher consumption to reduce their usage relative to those with lower consumption, and they also appear motivated to improve their ranking in terms of their

consumption.

6 CONCLUSION

In this study, we conducted an experiment to assess the effectiveness of providing non-price-based conservation feedback to students at NTHU in Taiwan through email messages. The experiment involved sending two types of messages to participants: a quintile comparison (Qt group) and a mean comparison (Avg group). We also varied the frequency of message delivery, sending messages either once or twice. The purpose of this experiment was to examine the effects of the feedback on energy conservation behavior by comparing the electricity consumption of the treatment groups to a control group.

Our results indicate that providing feedback on electricity consumption through email has no significant effect on changing residents' average electricity behavior. However, when examining the treatment effects across different groups of residents, we find that the feedback has heterogeneous effects on conservation behavior: high-consuming users tend to reduce their electricity usage more than low-consuming users, making it difficult to observe an overall treatment effect.

Interestingly, we observed some suggested patterns indicating that sending the message twice resulted in slightly more positive effects than sending it once, suggesting a small salience effect. This finding implies that repeated messaging might help maintain the feedback's presence. Furthermore, the heterogeneous effects identified through machine learning methods highlights the importance of designing targeted conservation messages. Specifically, emphasizing the targeting of high versus low electricity users or those with different family backgrounds or personal traits may be more effective in future interventions. A good example is Gerarden and Yang (2023), who demonstrate that targeting based on historical data can significantly improve program efficiency.

In addition, we identify the role of group dynamics in shaping response patterns. The group setting, such as the number of roommates and their coordination, seems to drive the effectiveness of the messages. For example, in rooms with fewer occupants, where consensus is easier to form, high-consuming users were more likely to reduce their usage, while low-consuming users tended to increase their usage. Our ML analysis provides some evidence that baseline usage, average room creativity, and room difference in physical fitness may play a role in shaping room dynamics. These factors could affect how individuals coordinate decisions, potentially shaping behavioral patterns. This highlights the importance of considering room dynamics in fostering certain behaviors and suggests that policymakers could benefit from future research that better account for these dynamics.

This study contributes to the growing body of literature on conservation efforts in university settings and offers insights into how social comparison-based messages can be leveraged to affect behaviors. The results suggest that future interventions could benefit from being more tailored to specific target groups and considering group dynamics, to optimize the impact of conservation messages.

It is important to note that this study is limited by its short time frame, constrained by the academic semester and seasonal variations in electricity usage. As such, we were unable to observe long-term behavior change. Future research could examine the dynamics of behavior change over a longer period, incorporating a broader range of group interactions. Future research could also explore innovative design features that helps disentangle coordination and free-rider mechanisms, for example, by introducing coordination tools such as shared temperature-setting plans or group chat reminders, and individualized incentives that make personal contributions more salient. Additionally, collecting richer measures of room dynamics or bargaining power could further refine strategies for effective energy conservation in shared living environments.

ACKNOWLEDGMENTS

We thank Laura Schechter, Tzu-Ting Yang, as well as the editor and reviewers for their valuable comments and discussions. We thank seminar participants at National Chung Hsing University Applied Economics, The 5th World Congress of Taiwan Studies, Taiwan Econometric Society Annual Conference, National Tsing Hua University, Soochow University, National Central University, Fu Jen Catholic University Economics, National Chengchi University Business School, and Taiwan Association of Environmental and Resource Economics. This research receives financial support from the National Science Council of Taiwan in the form of grant 110-2410-H-007-088. We also thank the Center for Institutional Research (CIR) at National Tsing-Hua University for kindly providing the administration data for our empirical analysis. The data usage is upon the restriction of the Personal Information Protection Act in Taiwan. Authors have no conflict of interest to declare.

ENDNOTES

1 Room size does not vary based on the number of occupants. Within the same dormitory building, all rooms are the same size regardless of occupancy, ranging from approximately 116 to 253 square feet. Additionally, all estimations are clustered at the room level and control for individual fixed effects.

2 Ito et al. (2018) further differentiate habituation effects between moral suasion and economic incentives, finding that the former declines faster.

3 See Chen et al. (2023)'s paper for the only other RCT study in Taiwan that uses performance-based monetary incentives to encourage energy conservation among newly recruited participants through competitive framing. Their context differs from ours, as they focus on group contest, but their findings highlight an alternative pathway for influencing energy use behavior in Taiwan.

4 Here is how the buildings are classified according to student status and gender: Thirteen of the dormitories are for undergraduates, three are for graduate dormitories, and the remaining are for mixed-use. Regarding gender, there are six female-only buildings, nine male-only buildings, and five mixed-gender buildings. The overall building complex goes from the second to tenth floors, while most buildings have three or four stories.

5 In our study, the composition of room types reflects the diverse living arrangements typically found in university dormitories. Specifically, the sample included 248 single rooms, 1,522 twin rooms, 153 triple rooms, and 786 quadruple rooms. Multi-occupancy rooms are a common feature in university settings, as adopted by similar studies such as those conducted by Delmas and Lessem (2014), Anderson et al. (2017), Crago et al. (2020), and Myers and Souza (2020). By examining a wide range of room types, our study aims to provide comprehensive and relevant insights into the typical residential experience in university dormitories.

6 Brülisauer et al. (2020) experiment demonstrates that providing feedback of high-energy-use appliances

could promote graduate students' energy conservation more effectively than providing general energy usage feedback. This finding supports taking high-energy-consuming appliances into account.

[7] Because the frequency of collecting electricity records changes from every two to three weeks during the experiment, we transform the usage unit into a one-week unit.

[8] The test results can be provided upon request but they are not shown here.

[9] The university rooms at NTHU are assigned almost randomly but may be influenced by observable characteristics, such as geographic origin and academic department, due to logistical convenience for the administrative staff. These observables were controlled by the individual fixed effects in our model.

[10] The detailed numbers are presented in Appendix Table A-1.

[11] Students have different channels through which they are admitted to the university. For example, some admissions are based more on exams, while others are based on oral interviews. Their ID organization, which determines the dorm assignment, may randomly reflect this administrative process, and different admission channels may attract different profiles of students.

[12] Please note that the coefficient in this boxplot is presented for more intuitive illustration. This presentation may not fully capture the heterogeneity and complex interactions identified by the Causal Forest.

[13] Wang et al. (2020) also identify a nuanced relationship between the Openness trait and energy conservation. They find that Openness is positively correlated with energy-saving behaviors in household environments but not significant in office settings. Additionally, our measure is based on local expert assessments, which may not fully align with the personality trait measures used in their study.

[14] Average high school ranking and average father's education appear to be fourth and fifth important predictors. Higher academic ranking may indicate greater self-discipline, while parental education could influence environmental awareness. However, neither is significant in the heterogeneity regression analysis.

[15] Exclusion restriction assumption may violated if some students act on the email without completing the survey; however, because the email and survey present very similar information, any such violation would bias the IV estimate away from zero, especially with a weak first stage. Our near-zero LATE estimate therefore remains information as diagnostic remains informative.

REFERENCES

Akerlof, G. A. (1991). Procrastination and obedience. *American Economic Review*, 81(2):1–19.

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, 95(9-10):1082–1095.

Allcott, H. and Kessler, J. B. (2019). The welfare effects of nudges: A case study of energy use social comparisons. *American Economic Journal: Applied Economics*, 11(1):236–276.

Allcott, H. and Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review*, 104(10):3003–3037.

Anderson, K., Song, K., Lee, S., Krupka, E., Lee, H., and Park, M. (2017). Longitudinal analysis of normative energy use feedback on dormitory occupants. *Applied Energy*, 189:623–639.

Andor, M. A., Gerster, A., Peters, J., and Schmidt, C. M. (2020). Social norms and energy conservation beyond the us. *Journal of Environmental Economics and Management*, 103:102351.

Asensio, O. I. and Delmas, M. A. (2015). Nonprice incentives and energy conservation. *Proceedings of the National Academy of Sciences*, 112(6):E510–E515.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Beshears, J., Choi, J. J., Laibson, D., and Madrian, B. C. (2011). The availability and utilization of 401 (k) loans. Technical report, National Bureau of Economic Research.

Bhanot, S. P. (2017). Rank and response: A field experiment on peer information and water use behavior. *Journal of Economic Psychology*, 62:155–172.

Binmore, K., Rubinstein, A., and Wolinsky, A. (1986). The nash bargaining solution in economic modelling. *The RAND Journal of Economics*, pages 176–188.

Bonan, J., Cattaneo, C., d'Adda, G., and Tavoni, M. (2020). The interaction of descriptive and injunctive social norms in promoting energy conservation. *Nature Energy*, 5(11):900–909.

Bonan, J., Cattaneo, C., d'Adda, G., and Tavoni, M. (2021). Can social information programs be more effective? the role of environmental identity for energy conservation. *Journal of Environmental Economics and Management*, 108:102467.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2013). Salience and consumer choice. *Journal of Political Economy*, 121(5):803–843.

Breza, E. (2012). Peer effects and loan repayment: Evidence from the krishna default crisis. *Job market paper MIT*, 1.

Brülisauer, M., Goette, L., Jiang, Z., Schmitz, J., and Schubert, R. (2020). Appliance-specific feedback and social comparisons: Evidence from a field experiment on energy conservation. *Energy Policy*, 145:111742.

Byrne, D. P., Nauze, A. L., and Martin, L. A. (2018). Tell me something i don't already know: Informedness and the impact of information programs. *Review of Economics and Statistics*, 100(3):510–527.

Calzolari, G. and Nardotto, M. (2017). Effective reminders. *Management Science*, 63(9):2915–2932.

Chen, C.-W., Chen, J. I., and Lin, M.-J. (2023). Performance feedback in a group contest: A field experiment on electricity conservation. *Journal of Economics & Management Strategy*, 32(3):558–582.

Costa, D. L. and Kahn, M. E. (2013). Energy conservation "nudges" and environmentalist ideology: Evidence from a randomized residential electricity field experiment. *Journal of the European Economic Association*, 11(3):680–702.

Crago, C. L., Spraggon, J. M., and Hunter, E. (2020). Motivating non-ratepaying households with feedback and social nudges: A cautionary tale. *Energy Policy*, 145:111764.

Delmas, M. A., Fischlein, M., and Asensio, O. I. (2013). Information strategies and energy conservation behavior: A meta-analysis of experimental studies from 1975 to 2012. *Energy Policy*, 61:729–739.

Delmas, M. A. and Lessem, N. (2014). Saving power to conserve your reputation? the effectiveness of private versus public information. *Journal of Environmental Economics and Management*, 67(3):353–370.

Eisenberg, D., Golberstein, E., and Whitlock, J. L. (2014). Peer effects on risky behaviors: New evidence from college roommate assignments. *Journal of Health Economics*, 33:126–138.

Gerarden, T. D. and Yang, M. (2023). Using targeting to optimize program design: evidence from an energy conservation experiment. *Journal of the Association of Environmental and Resource Economists*, 10(3):687–716.

Gerber, A. S. and Rogers, T. (2009). Descriptive social norms and motivation to vote: Everybody's voting and so should you. *The Journal of Politics*, 71(1):178–191.

Habla, W. and Muller, P. (2021). Experimental evidence of limited attention at the gym. *Experimental Economics*, 24(4):1156–1184.

Ito, K., Ida, T., and Tanaka, M. (2018). Moral suasion and economic incentives: Field experimental evidence from energy demand. *American Economic Journal: Economic Policy*, 10(1):240–267.

Karlan, D., McConnell, M., Mullainathan, S., and Zinman, J. (2016). Getting to the top of mind: How reminders increase saving. *Management Science*, 62(12):3393–3411.

Knittel, C. R. and Stolper, S. (2025). Using machine learning to target treatment: The case of household energy use. *The Economic Journal*, 135(672):2377–2401.

Kuhnen, C. M. and Tymula, A. (2012). Feedback, self-esteem, and performance in organizations. *Management Science*, 58(1):94–113.

Lukaszewski, A. W. (2013). Testing an adaptationist theory of trait covariation: Relative bargaining power as a common calibrator of an interpersonal syndrome. *European Journal of Personality*, 27(4):328–345.

Murakami, K., Shimada, H., Ushifusa, Y., and Ida, T. (2022). Heterogeneous treatment effects of nudge and rebate: Causal machine learning in a field experiment on electricity conservation. *International Economic Review*, 63(4):1779–1803.

Myers, E. and Souza, M. (2020). Social comparison nudges without monetary incentives: Evidence from home energy reports. *Journal of Environmental Economics and Management*, 101:102315.

Nakajima, R. (2007). Measuring peer effects on youth smoking behaviour. *The Review of Economic Studies*, 74(3):897–935.

Nash, J. (1953). Two-person cooperative games. *Econometrica: Journal of the Econometric Society*, pages 128–140.

Nash, J. F. et al. (1950). The bargaining problem. *Econometrica*, 18(2):155–162.

O'Neill, E. and Weeks, M. (2018). Causal tree estimation of heterogeneous household response to time-of-use electricity pricing schemes. *arXiv preprint arXiv:1810.09179*.

Robitaille, N., House, J., and Mazar, N. (2021). Effectiveness of planning prompts on organizations' likelihood to file their overdue taxes: A multi-wave field experiment. *Management Science*, 67(7):4327–4340.

Rodriguez, N. N. and Lukaszewski, A. W. (2020). Functional coordination of personality strategies with physical strength and attractiveness: A multi-sample investigation at the hexaco facet-level. *Journal of Research in Personality*, 89:104040.

Roels, G. and Su, X. (2014). Optimal design of social comparison effects: Setting reference groups and reference points. *Management Science*, 60(37):606–627.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, 18(5):429–434.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wang, Q.-C., Wang, Y.-X., Jian, I. Y., Wei, H.-H., Liu, X., and Ma, Y.-T. (2020). Exploring the "energy-saving personality traits" in the office and household situation: an empirical study. *Energies*, 13(14):3535.

Yakovlev, E. (2018). Demand for alcohol consumption in russia and its implication for mortality. *American Economic Journal: Applied Economics*, 10(1):106–149.

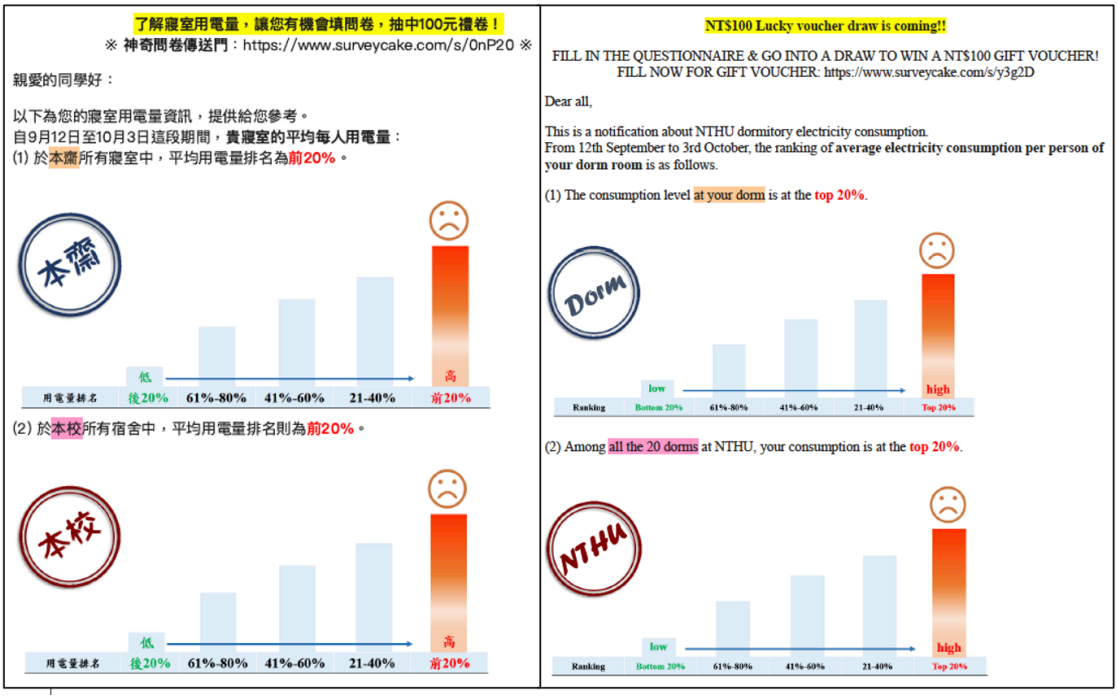


Figure 1: Electricity Feedback – Quintile Information

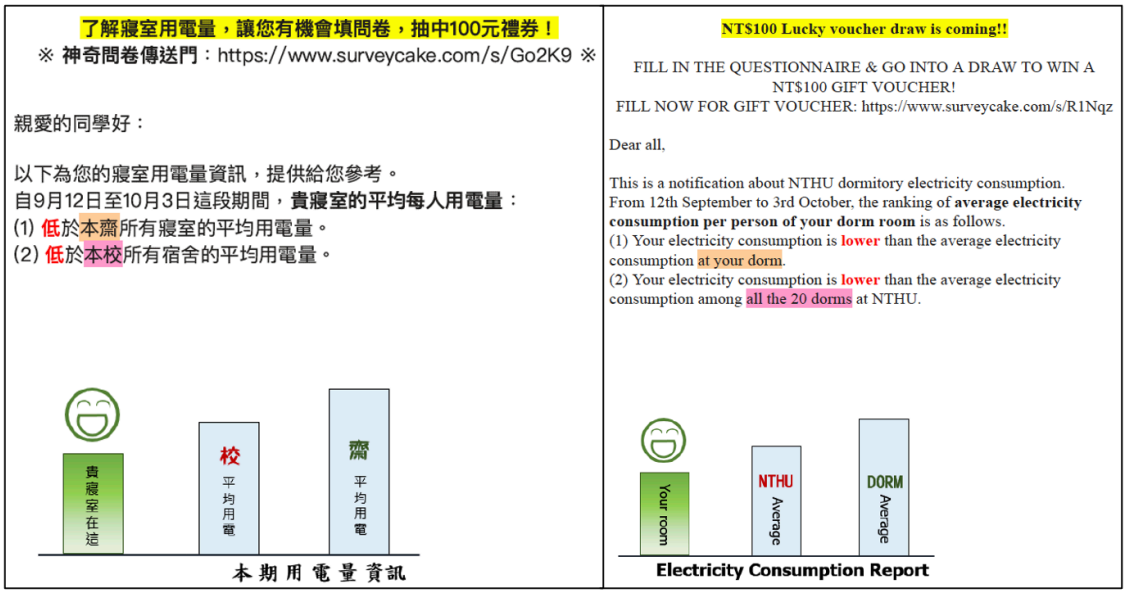


Figure 2: Electricity Feedback – Average Information

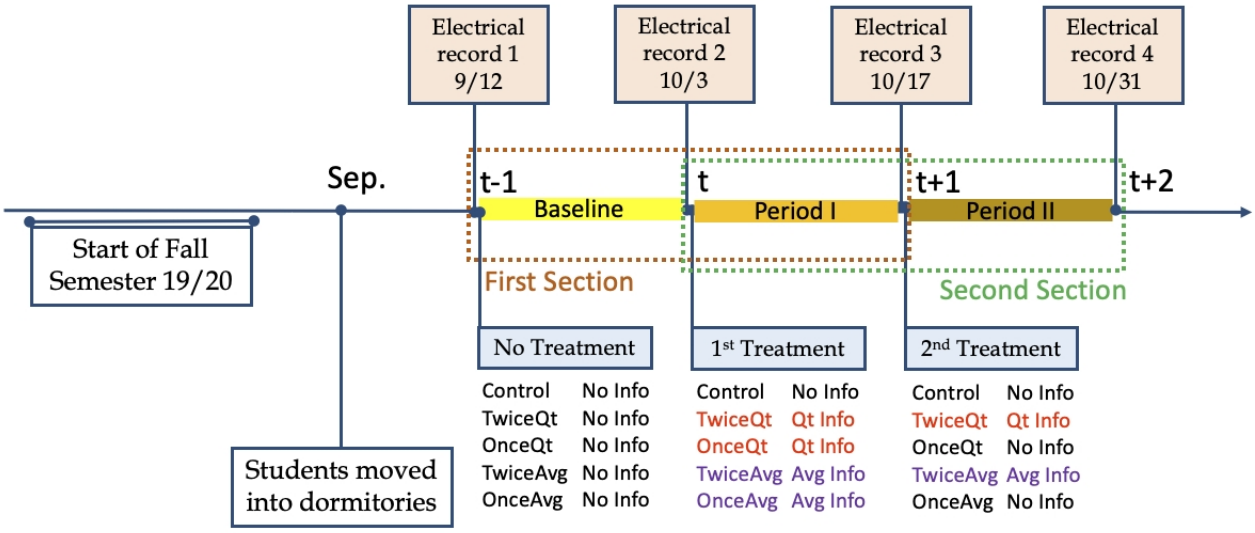


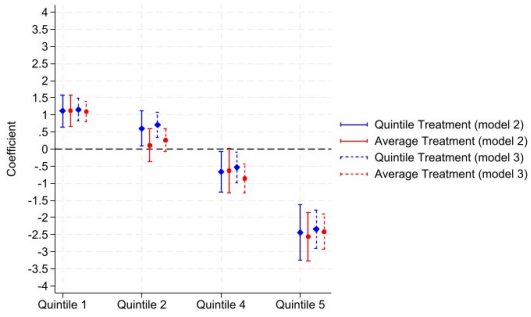
Figure 3: Model Design

Figure 4: A Watt-hour meters located outside dormitory room

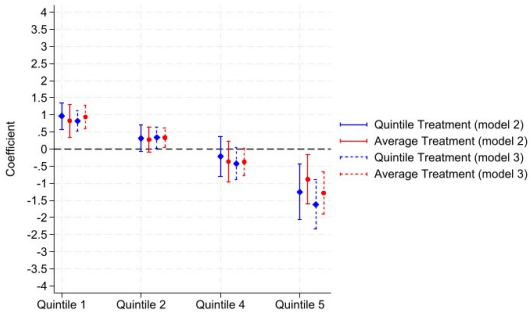
Note: The Watt-hour meters are part of the dormitory's original infrastructure and record each room's electricity consumption. We collected data on electricity usage from these meters, located outside each dormitory room.

Figure 5: An energy management device with a top-up card

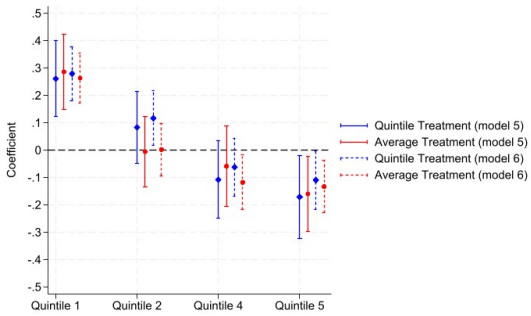
Note: The electricity top-up and consumption system is part of the dormitory's original infrastructure. Residents preload funds onto their cards, which they insert to access electricity. When the card balance is depleted, residents can reload funds at the administrative center.



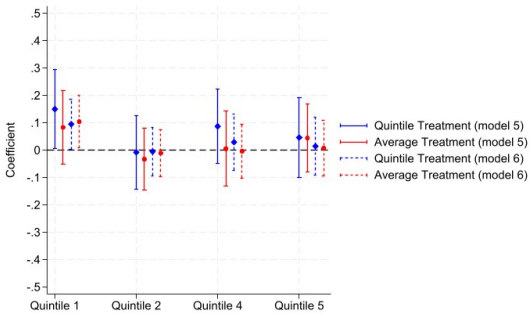
(a) One- and two-student Room (level)



(b) Three- and four-student Room (level)



(c) One- and two-student Room (log)



(d) Three- and four-student Room (log)

Figure 6: Comparison of Heterogeneous Treatment Effects Across Student Rooms

Note: The models in these graphs using the preferred models in Table 2. Panels (a) and (b) depict estimates for electricity "Usage Level" (Models 2 and 3), while panels (c) and (d) show results for "Usage Log" (Models 5 and 6). Each dot represents the estimated treatment effect for a given quintile (relative to the 3rd quintile), while the vertical arrows above and below the dots represent the 95% confidence intervals (CI).

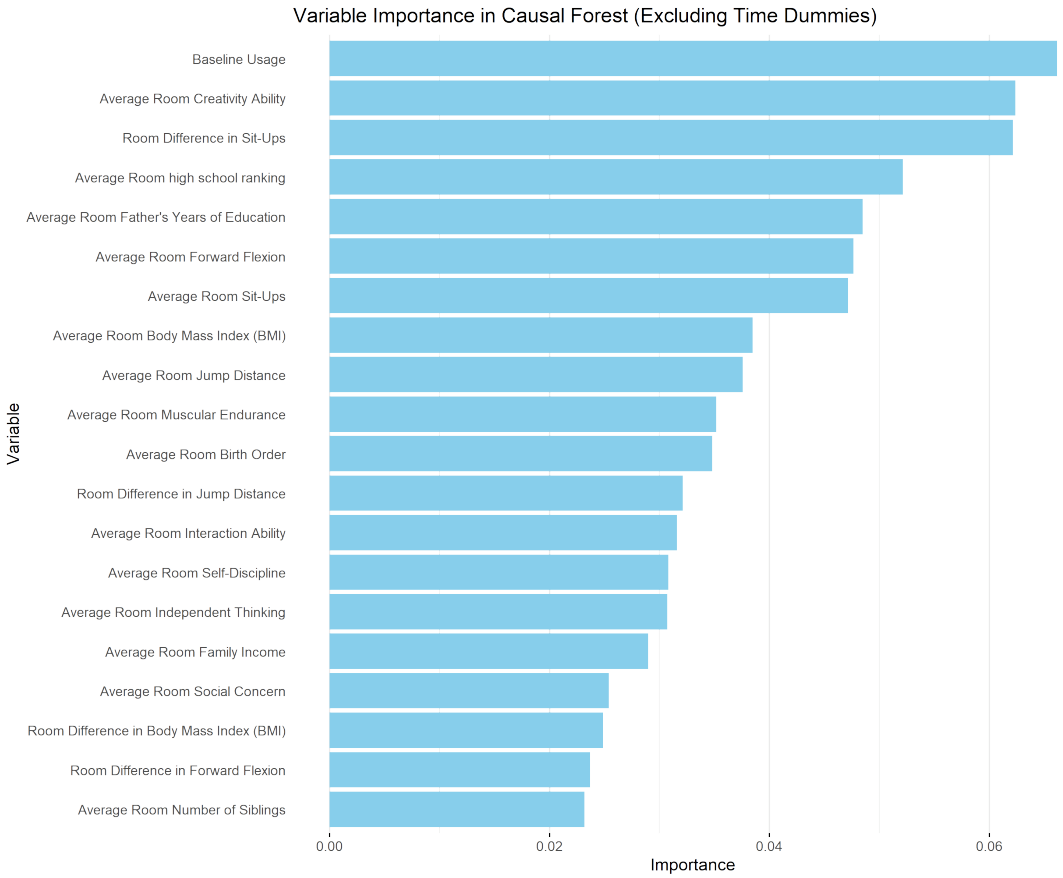


Figure 7: Variable Importance Plot

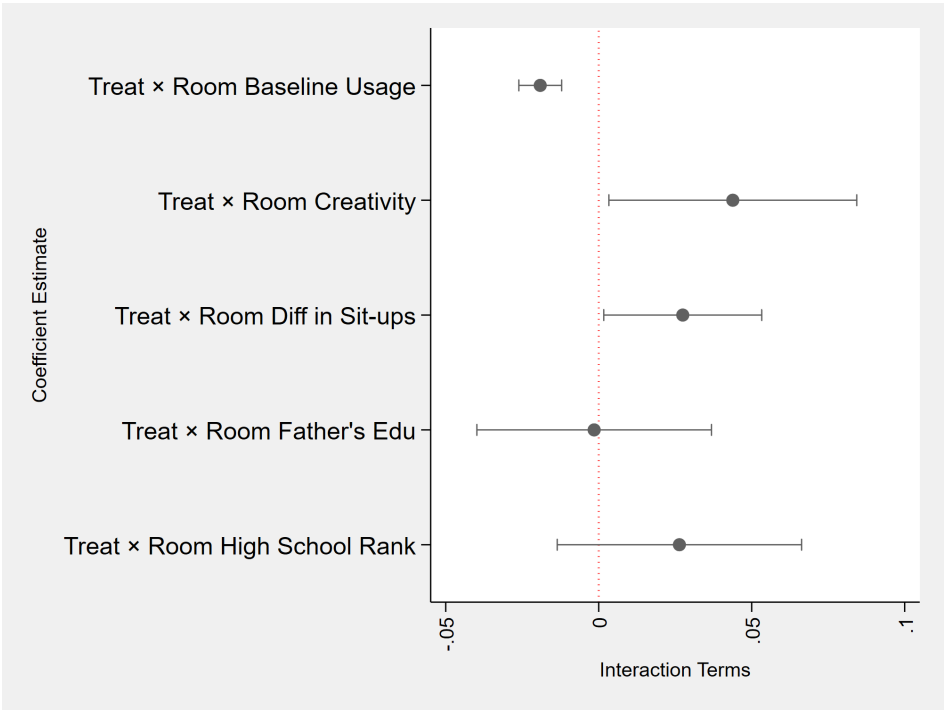


Figure 8: HTE Fixed Effects Model at Room Level

Note: The model in these graphs uses all the top five most important variables to estimate heterogeneous treatment effects at the room level. Standard errors are clustered at the room level. The dots represent the coefficients of the interaction terms, while the arrows represent the 95% confidence intervals (CI).

Table 1: Summary Statistics at the Individual Level by Treatment Status

	All	(1) Control	(2) TwiceQt	(3) OnceQt	(4) TwiceAvg	(5) OnceAvg	Multivariate means test
Variable	N=6,723	N=1,440	N=1,290	N=1,309	N=1,300	N=1,384	Prob>F
Panel A. Electricity Usage							
usage in baseline	5.646	5.697	5.737	5.487	5.783	5.530	0.298
	(4.301)	(4.282)	(4.339)	(4.051)	(4.457)	(4.366)	
Panel B. Characteristics of Rooms							
located floor	3.485	3.466	3.458	3.517	3.477	3.507	0.926
	(1.994)	(1.998)	(1.985)	(2.003)	(1.991)	(1.993)	
# students per room	2.931	2.950	2.930	2.912	2.948	2.915	0.803
	(1.025)	(1.027)	(1.024)	(1.012)	(1.024)	(1.038)	
Panel C. Residents' characteristics							
male	0.582	0.580	0.583	0.581	0.592	0.573	0.899
	(0.493)	(0.494)	(0.493)	(0.494)	(0.492)	(0.495)	
bachelor	0.703	0.722	0.702	0.691	0.694	0.704	0.401
	(0.457)	(0.448)	(0.457)	(0.462)	(0.461)	(0.457)	
master	0.243	0.235	0.242	0.250	0.253	0.236	0.748
	(0.429)	(0.424)	(0.428)	(0.433)	(0.435)	(0.425)	
phd	0.053	0.042	0.055	0.059	0.052	0.059	0.215
	(0.224)	(0.200)	(0.228)	(0.235)	(0.221)	(0.236)	
natural science major	0.692	0.678	0.705	0.681	0.685	0.714	0.167
	(0.462)	(0.467)	(0.456)	(0.466)	(0.465)	(0.452)	

Notes: The usage variables report weekly usage (kWh) per person. Standard deviations are shown in parentheses. Located building ID is the building ID of twenty dormitories. The columns (1), (2), (3), (4) and (5) display the sample means for the five groups. The multivariate means test tests whether the means across our control and four treatment groups are the same.

Table 2: Impact of Quintile and Average Treatment on Electricity Usage

	(1)	(2)	(3)	(4)	(5)	(6)
	Usage Level			Usage Log		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Quintile Treatment	-0.043	0.013	-0.072	-0.018	0.002	-0.021
	(0.109)	(0.120)	(0.106)	(0.018)	(0.024)	(0.020)
Average Treatment	-0.082	-0.033	-0.079	-0.021	-0.019	-0.019
	(0.106)	(0.119)	(0.105)	(0.017)	(0.023)	(0.020)
Individual FE	v	v	v	v	v	v
Time FE	v	v	v	v	v	v
Whole sample, pre+post 1 period	v			v		
All periods, excluding groups treated twice		v			v	
Whole sample, all periods			v			v
N	13446	12399	20169	13446	12399	20169
R-sq	0.040	0.488	0.490	0.012	0.606	0.616

Notes: The outcome variable is the usage (kWh) per person per week during each period. Standard errors in parentheses are clustered at the room level. * $p < 0.1$, **$p < 0.05$, ***$p<0.01$. Quintile treatment here includes the OnceQt group and TwiceQt group in the quintile social comparison information treatment; Average Treatment includes the OnceAvg group and TwiceAvg group in the average social comparison information treatment. The first two rows report the treatment effects from equation 1.

Table 3: Impact of Sending Social Comparison Messages Once vs. Twice on Usage by Treatment and User Group

	Usage Level			Usage Log		
	Total	High user (top 20%)	Low user (bottom 20%)	Total	High user (top 20%)	Low user (bottom 20%)
Once Quintile Treatment	0.013	-0.051	0.074	0.002	-0.007	0.014
	(0.120)	(0.406)	(0.128)	(0.024)	(0.062)	(0.049)
Twice Quintile Treatment	-0.157	-0.414	-0.174	-0.044*	-0.049	-0.092**
	(0.127)	(0.429)	(0.109)	(0.023)	(0.062)	(0.038)
Once Average Treatment	-0.033	-0.339	0.004	-0.019	-0.067	0.003
	(0.119)	(0.410)	(0.112)	(0.023)	(0.063)	(0.040)
Twice Average Treatment	-0.128	-0.299	-0.192*	-0.020	-0.028	-0.062
	(0.126)	(0.419)	(0.112)	(0.024)	(0.061)	(0.041)
Difference between once and twice (quintile) groups			**	*		**
Difference between once and twice (average) groups			*			
N	20169	3816	3894	20169	3816	3894
R-sq	0.490	0.608	0.354	0.616	0.595	0.456

Notes: The outcome variable is electricity usage (kWh) per person per week during each period. Standard errors in parentheses are clustered at the room level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Impact of Sending Social Comparison Messages Once vs. Twice on Usage in Different Post-Treatment Periods

	(1)	(2)
	Usage Level	Usage Log
Quintile treatment (once) x post 1 period	-0.005	0.003
	(0.120)	(0.020)
Quintile treatment (once) x post 2 period	0.030	0.001
	(0.187)	(0.038)
Quintile treatment (twice) x post 1 period	-0.083	-0.038*
	(0.135)	(0.021)
Quintile treatment (twice) x post 2 period	-0.232	-0.051
	(0.193)	(0.037)
Average treatment (once) x post 1 period	-0.020	-0.021
	(0.121)	(0.020)
Average treatment (once) x post 2 period	-0.047	-0.017
	(0.187)	(0.037)
Average treatment (twice) x post 1 period	-0.147	-0.021
	(0.122)	(0.020)
Average treatment (twice) x post 2 period	-0.108	-0.019
	(0.193)	(0.038)
N	20169	20169
R-sq	0.490	0.616

Notes: The outcome variable is electricity usage (kWh) per person per week during each period. Standard errors in parentheses are clustered at the room level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Heterogeneous Impact of Quintile and Average Treatment on Electricity Usage (Interaction by quintiles)

	(1)	(2)	(3)	(4)	(5)	(6)
		Usage Level			Usage Log	
Quintile Treatment	0.055	0.203	0.167	-0.010	-0.021	-0.037
	(0.185)	(0.148)	(0.132)	(0.028)	(0.041)	(0.030)
Average Treatment	-0.011	0.253*	0.204*	-0.012	-0.012	-0.018
	(0.149)	(0.150)	(0.122)	(0.025)	(0.037)	(0.029)
Quintile x Low users (1st quintile)	-0.276	1.029***	0.945***	-0.014	0.196***	0.167***
	(0.182)	(0.149)	(0.116)	(0.039)	(0.052)	(0.034)
Quintile x Low users (2nd quintile)	0.037	0.441***	0.474***	0.042	0.029	0.037
	(0.193)	(0.158)	(0.124)	(0.034)	(0.049)	(0.034)
Quintile x High users (4th quintile)	-0.088	-0.479**	-0.533***	-0.023	-0.032	-0.043
	(0.227)	(0.207)	(0.164)	(0.032)	(0.052)	(0.038)
Quintile x High users (5th quintile)	-0.150	-1.966***	-2.120***	-0.041	-0.093*	-0.089**
	(0.262)	(0.293)	(0.228)	(0.030)	(0.056)	(0.039)
Average x Low users (1st quintile)	-0.091	0.944***	1.006***	0.034	0.162***	0.170***
	(0.147)	(0.174)	(0.122)	(0.036)	(0.050)	(0.035)
Average x Low users (2nd quintile)	-0.208	0.209	0.310***	-0.027	-0.023	-0.002
	(0.155)	(0.150)	(0.110)	(0.032)	(0.044)	(0.033)
Average x High users (4th quintile)	0.041	-0.522**	-0.625***	-0.008	-0.043	-0.067*
	(0.188)	(0.226)	(0.152)	(0.027)	(0.052)	(0.037)
Average x High users (5th quintile)	-0.078	-2.185***	-2.107***	-0.044*	-0.149***	-0.117***
	(0.230)	(0.282)	(0.207)	(0.027)	(0.052)	(0.037)
Individual FE	v	v	v	v	v	v
Time FE	v	v	v	v	v	v
Whole sample, pre+post 1 period	v			v		
All periods, excluding group treated twice		v			v	
Whole sample, all periods			v			v
N	13446	12399	20169	13446	12399	20169
R-sq	0.042	0.514	0.523	0.017	0.612	0.622

Notes: The outcome variable is electricity usage (kWh) per person per week during each period. Standard errors in parentheses are clustered at the room level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Quintile and average treatments are interacted with high users (4th and 5th quintiles) and low users (1st and 2nd quintiles). The omitted category is the third quintile.

Table 6: Heterogeneous Impact of Quintile and Average Treatment on Electricity Usage (Interacting with Median)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Usage Level			Usage Log			Adjusted Rank (the lower, the better)		
Quintile Treatment	-0.047	0.822***	0.769***	0.000	0.070**	0.046**	-0.113**	-0.355***	-0.347***
	(0.114)	(0.107)	(0.098)	(0.022)	(0.028)	(0.022)	(0.047)	(0.053)	(0.044)
Average Treatment	-0.147	0.751***	0.762***	-0.011	0.047*	0.048**	-0.108**	-0.342***	-0.358***
	(0.106)	(0.111)	(0.100)	(0.022)	(0.027)	(0.023)	(0.048)	(0.054)	(0.046)
Quintile x High users (>= medium)	0.006	-1.655***	-1.697***	-0.036*	-0.140***	-0.135***	-0.262***	-0.713***	-0.729***
	(0.133)	(0.159)	(0.118)	(0.021)	(0.034)	(0.024)	(0.048)	(0.066)	(0.046)
Average x High users (>= medium)	0.132	-1.605***	-1.702***	-0.020	-0.135***	-0.137***	-0.293***	-0.723***	-0.771***
	(0.121)	(0.160)	(0.116)	(0.020)	(0.033)	(0.024)	(0.049)	(0.066)	(0.048)
Individual FE	v	v	v	v	v	v	v	v	v
Time FE	v	v	v	v	v	v	v	v	v
Whole sample, pre+post 1 period	v			v			v		
All periods, excluding groups treated twice		v			v			v	
Whole sample, all periods			v			v			v
N	13446	12399	20169	13446	12399	20169	13446	12399	20169
R-sq	0.041	0.504	0.510	0.013	0.609	0.619	0.029	0.082	0.094

Notes: The outcome variable is electricity usage (kWh) per person per week during each period. Standard errors in parentheses are clustered at the room level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Quintile and average treatments are interacted with users higher than or equal to the median electricity usage. Adjusted rank is defined as 1 if it falls within the lowest electricity usage quintile and 5 if it is in the highest usage quintile, and is time-variant.

Table 7: Heterogeneous Impact of Quintile and Average Treatment on Electricity Usage (Interacting with Median)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Usage Level			Usage Log			Adjusted Rank (the higher, the better)		
One- and Two-student Room									
Quintile Treatment	-0.041	1.410***	1.297***	0.015	0.182***	0.145***	0.153**	0.618***	0.568***
	(0.148)	(0.155)	(0.142)	(0.029)	(0.038)	(0.031)	(0.064)	(0.070)	(0.060)
Average Treatment	-0.208	1.195***	1.160***	-0.027	0.118***	0.102***	0.095	0.512***	0.484***
	(0.148)	(0.158)	(0.144)	(0.030)	(0.039)	(0.032)	(0.064)	(0.072)	(0.062)
Quintile x High users ($\geq$ medium)	0.209	-2.247***	-2.153***	-0.025	-0.294***	-0.262***	-0.329***	-1.057***	-1.001***
	(0.175)	(0.222)	(0.159)	(0.030)	(0.045)	(0.032)	(0.065)	(0.082)	(0.058)
Average x High users ($\geq$ medium)	0.257	-2.164***	-2.156***	0.006	-0.237***	-0.220***	-0.323***	-0.962***	-0.945***
	(0.180)	(0.221)	(0.158)	(0.030)	(0.045)	(0.032)	(0.066)	(0.083)	(0.061)
N	6422	5946	9633	6422	5946	9633	6422	5946	9633
R-sq	0.036	0.512	0.519	0.003	0.595	0.610	0.031	0.075	0.085
Three- and Four-student Room									
Quintile Treatment	-0.053	0.316**	0.313**	-0.014	-0.024	-0.037	0.080	0.135*	0.163***
	(0.169)	(0.143)	(0.128)	(0.032)	(0.040)	(0.031)	(0.069)	(0.076)	(0.063)
Average Treatment	-0.110	0.357**	0.399***	-0.006	-0.015	-0.003	0.112	0.199***	0.243***
	(0.151)	(0.148)	(0.131)	(0.031)	(0.037)	(0.031)	(0.070)	(0.076)	(0.065)
Quintile x High users ($\geq$ medium)	-0.220	-1.035***	-1.211***	-0.046	0.019	-0.004	-0.193***	-0.367***	-0.450***
	(0.196)	(0.217)	(0.170)	(0.030)	(0.047)	(0.033)	(0.070)	(0.098)	(0.067)
Average x High users ($\geq$ medium)	0.019	-0.816***	-1.065***	-0.027	0.007	-0.017	-0.233***	-0.407***	-0.529***
	(0.156)	(0.193)	(0.152)	(0.027)	(0.045)	(0.032)	(0.072)	(0.097)	(0.072)
N	7024	6453	10536	7024	6453	10536	7024	6453	10536
R-sq	0.057	0.578	0.569	0.036	0.657	0.658	0.032	0.186	0.180
Individual FE	v	v	v	v	v	v			
Time FE	v	v	v	v	v	v			
Whole sample, pre+post 1 period	v			v			v		
All periods, excluding groups treated twice		v			v			v	
Whole sample, all periods			v			v			v

Notes: The outcome variable is electricity usage (kWh) per person per week during each period. Standard errors in parentheses are clustered at the room level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Quintile and average treatments are interacted with users higher than or equal to the median electricity usage. The upper part includes observations from the one- and two-student rooms; the lower part includes observations from the three- and four-student rooms.

Appendix A Tables and Figures

Table A-1: Heterogeneous Impact of Quintile and Average Treatment on Electricity Usage across Different Room Types (Interaction by quintiles)

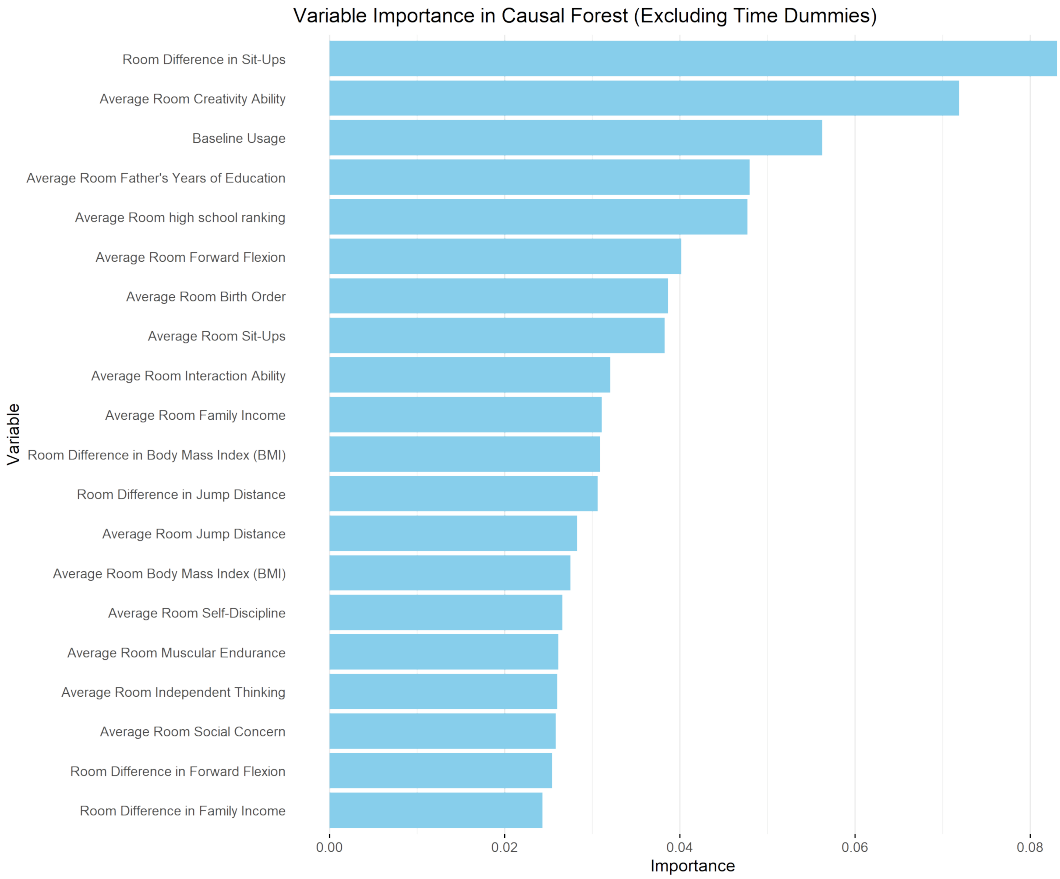
	One-Student Room						Three-Student Room					
	Usage Level			Log Usage			Usage Level			Log Usage		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	model1	model2	model3	model4	model5	model6	model1	model2	model3	model4	model5	model6
Quintile Treatment	1.010	1.466*	1.496**	0.132	0.133	0.114	−0.578	−0.742	−0.178	−0.163	−0.381**	−0.131
	(1.034)	(0.813)	(0.688)	(0.133)	(0.169)	(0.128)	(0.648)	(0.556)	(0.470)	(0.176)	(0.176)	(0.139)
Average Treatment	−1.567*	−0.471	−0.241	−0.306*	−0.341*	−0.254	0.158	−0.189	−0.047	0.019	−0.139	−0.101
	(0.853)	(0.738)	(0.635)	(0.184)	(0.199)	(0.162)	(0.835)	(0.670)	(0.565)	(0.122)	(0.148)	(0.123)
Quintile x Low users (1st quintile)	−0.767	−0.329	0.563	−0.081	−0.202	0.091	0.480	2.042***	1.431***	0.224	0.646***	0.361**
	(1.283)	(0.748)	(0.803)	(0.217)	(0.197)	(0.182)	(0.603)	(0.487)	(0.386)	(0.192)	(0.180)	(0.143)
Quintile x Low users (2nd quintile)	0.041	−0.382	0.578	0.073	−0.012	0.126	0.135	0.729	0.441	0.042	0.204	0.024
	(1.465)	(0.713)	(0.844)	(0.182)	(0.160)	(0.142)	(0.720)	(0.553)	(0.445)	(0.219)	(0.198)	(0.154)
Quintile x High users (4th quintile)	−1.784	−2.141**	−1.598**	−0.276	−0.299	−0.219	0.990	0.743	0.198	0.199	0.469	0.233
	(1.178)	(0.956)	(0.763)	(0.186)	(0.246)	(0.166)	(0.816)	(0.951)	(0.735)	(0.178)	(0.294)	(0.181)
Quintile x High users (5th quintile)	−1.331	−4.595***	−3.989***	−0.267*	−0.495**	−0.419***	1.593	2.211**	−0.796	0.193	0.787***	0.202
	(1.162)	(1.035)	(0.725)	(0.144)	(0.198)	(0.134)	(1.380)	(0.878)	(1.015)	(0.194)	(0.205)	(0.182)
Average x Low users (1st quintile)	0.865	2.438***	2.010***	0.263	0.703***	0.535***	−0.969	0.823	0.848*	−0.227	0.133	0.143
	(0.750)	(0.695)	(0.522)	(0.195)	(0.231)	(0.173)	(0.827)	(0.645)	(0.512)	(0.152)	(0.178)	(0.130)
Average x Low users (2nd quintile)	0.502	0.531	0.925	0.081	0.076	0.147	0.439	1.131	1.031*	0.166	0.243	0.267*
	(0.906)	(0.658)	(0.587)	(0.218)	(0.222)	(0.179)	(0.813)	(0.688)	(0.534)	(0.118)	(0.185)	(0.140)
Average x High users (4th quintile)	0.121	−0.481	−0.486	0.139	0.051	0.025	−0.335	−0.631	−0.657	−0.047	0.100	0.094
	(0.945)	(0.910)	(0.656)	(0.190)	(0.238)	(0.173)	(0.865)	(0.796)	(0.601)	(0.115)	(0.187)	(0.142)
Average x High users (5th quintile)	1.955	−2.151	−1.421	0.272	0.155	0.167	0.087	−0.792	−1.216	−0.025	0.374**	0.251
	(1.279)	(1.419)	(0.981)	(0.182)	(0.216)	(0.161)	(1.319)	(1.623)	(1.115)	(0.129)	(0.188)	(0.164)
N	568	534	852	568	534	852	960	888	1,440	960	888	1,440
R-sq	0.049	0.416	0.445	0.039	0.491	0.502	0.131	0.652	0.628	0.078	0.797	0.774
	Two-Student Room						Four-Student Room					
	Usage Level			Log Usage			Usage Level			Log Usage		
	model1	model2	model3	model4	model5	model6	model1	model2	model3	model4	model5	model6
Quintile Treatment	−0.155	0.514**	0.343*	−0.038	0.005	−0.055	0.165	−0.157	−0.183	−0.000	−0.053	−0.070*
	(0.253)	(0.261)	(0.207)	(0.046)	(0.062)	(0.049)	(0.270)	(0.184)	(0.178)	(0.036)	(0.049)	(0.037)
Average Treatment	0.182	0.732***	0.623***	0.023	0.022	0.009	−0.089	−0.103	−0.159	−0.031	−0.024	−0.033
	(0.233)	(0.249)	(0.191)	(0.039)	(0.059)	(0.044)	(0.200)	(0.193)	(0.162)	(0.033)	(0.047)	(0.039)
Quintile x Low users (1st quintile)	−0.101	1.264***	1.246***	0.016	0.305***	0.302***	−0.447*	0.894***	0.754***	−0.059	0.120	0.074
	(0.249)	(0.250)	(0.173)	(0.059)	(0.075)	(0.053)	(0.260)	(0.232)	(0.168)	(0.057)	(0.084)	(0.051)
Quintile x Low users (2nd quintile)	0.385	0.709***	0.785***	0.102*	0.101	0.130**	−0.128	0.343*	0.367**	0.023	0.004	0.014
	(0.274)	(0.273)	(0.192)	(0.054)	(0.071)	(0.053)	(0.267)	(0.198)	(0.166)	(0.040)	(0.065)	(0.043)
Quintile x High users (4th quintile)	0.394	−0.518	−0.407*	0.047	−0.086	−0.041	−0.521	−0.311	−0.488*	−0.079*	0.048	0.010
	(0.320)	(0.318)	(0.237)	(0.052)	(0.076)	(0.057)	(0.331)	(0.306)	(0.250)	(0.043)	(0.066)	(0.053)
Quintile x High users (5th quintile)	0.368	−2.158***	−2.123***	0.038	−0.130	−0.070	−0.962**	−1.727***	−1.726***	−0.110***	−0.037	0.006
	(0.335)	(0.443)	(0.314)	(0.047)	(0.083)	(0.060)	(0.413)	(0.423)	(0.393)	(0.041)	(0.075)	(0.054)
Average x Low users (1st quintile)	−0.296	0.981***	1.026***	−0.007	0.238***	0.240***	0.135	0.850***	0.965***	0.088*	0.085	0.105**
	(0.235)	(0.242)	(0.158)	(0.055)	(0.073)	(0.049)	(0.188)	(0.261)	(0.184)	(0.049)	(0.071)	(0.051)
Average x Low users (2nd quintile)	−0.517**	0.050	0.218	−0.094*	−0.018	−0.007	−0.089	0.176	0.250*	−0.008	−0.069	−0.045
	(0.257)	(0.256)	(0.180)	(0.052)	(0.069)	(0.051)	(0.195)	(0.192)	(0.145)	(0.041)	(0.058)	(0.045)
Average x High users (4th quintile)	−0.113	−0.667*	−0.881***	−0.038	−0.073	−0.126**	0.243	−0.202	−0.270	0.024	0.028	0.003
	(0.322)	(0.353)	(0.229)	(0.046)	(0.078)	(0.053)	(0.227)	(0.326)	(0.216)	(0.033)	(0.070)	(0.050)
Average x High users (5th quintile)	−0.114	−2.568***	−2.494***	−0.045	−0.193***	−0.160***	−0.525	−0.911**	−1.297***	−0.075*	0.003	−0.022
	(0.309)	(0.358)	(0.269)	(0.042)	(0.073)	(0.052)	(0.368)	(0.362)	(0.329)	(0.039)	(0.065)	(0.054)
N	5,854	5,412	8,781	5,854	5,412	8,781	6,064	5,565	9,096	6,064	5,565	9,096
R-sq	0.044	0.542	0.546	0.010	0.611	0.625	0.074	0.589	0.584	0.063	0.653	0.655

Standard errors in parentheses

Note: *p<0.1; **p<0.05; ***p<0.01

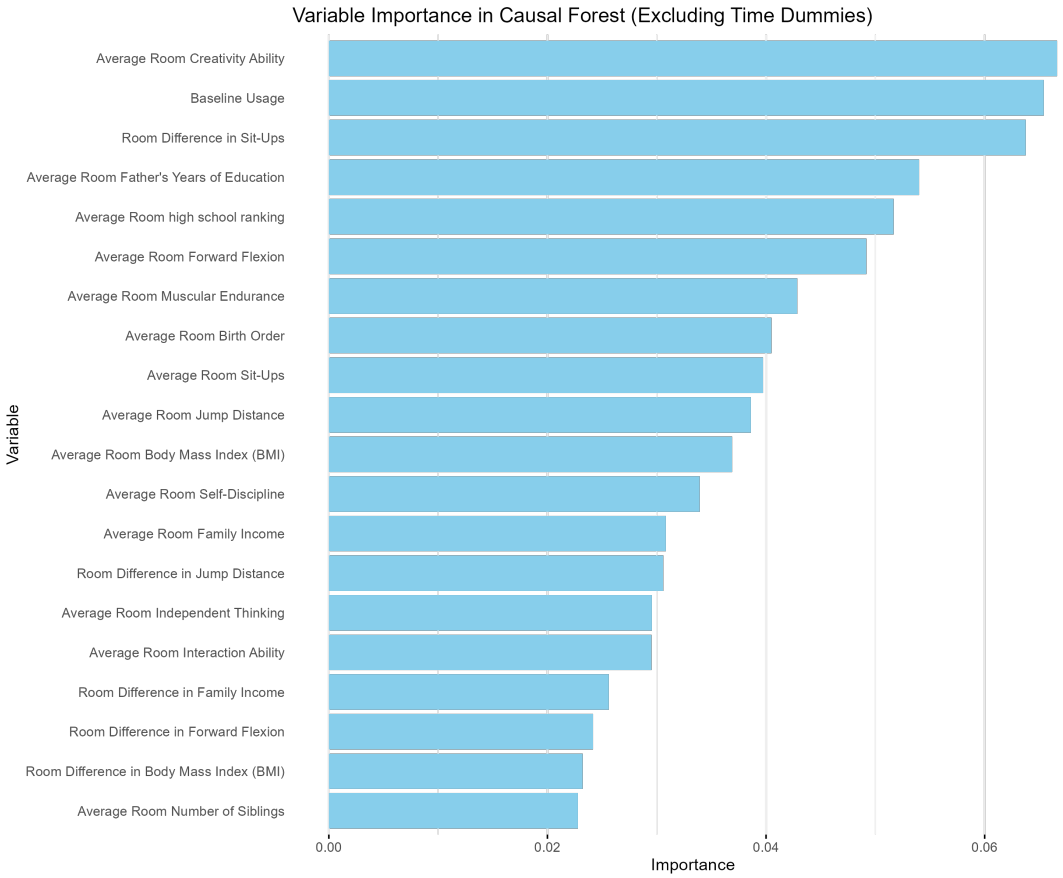
Notes: The outcome variable is electricity usage (kWh) per person per week during each period. Standard errors in parentheses are clustered at the room level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Quintile and average treatments are interacted with high users (4th and 5th quintiles) and low users (1st and 2nd quintiles). The omitted category is the third quintile. The upper part includes observations from the one- and three-student rooms; the lower part includes observations from the two- and four-student rooms.

Figure A-1: Robustness Check: Variable Importance Plot Setting Different Minimum Node Size



Note: This figure uses the same specification as in Figure 7, but with the minimum node size parameter set to 50 instead of 30.

Figure A-2: Robustness Check: Variable Importance Plot With Different Number of Trees



Note: This figure uses the same specification as in Figure 7, but with different number of trees (1000 instead of 2000).

Table A-2: Variable Definitions for ML Analysis

	mean	sd	min	max	count
Male Indicator	0.595	0.491	0.000	1.000	7206
Bachelor's Degree Indicator	0.637	0.481	0.000	1.000	7206
Average Room Independent Thinking	0.010	0.710	-3.595	1.491	6909
Average Room Social Concern	0.024	0.725	-3.132	1.530	6909
Average Room Self-Discipline	0.011	0.713	-2.937	1.458	6909
Average Room Interaction Ability	0.012	0.704	-3.025	1.487	6909
Average Room Creativity Ability	0.019	0.696	-2.828	1.774	6909
Average Room Father's Years of Education	-0.052	0.743	-2.733	2.302	6948
Average Room Number of Siblings	0.029	0.766	-1.579	3.866	6945
Average Room Birth Order	-0.004	0.680	-0.853	1.300	6945
Average Room Family Income	-0.050	0.691	-1.214	3.388	6963
Average Room Body Mass Index (BMI)	0.022	0.920	-2.113	21.130	3891
Average Room Forward Flexion	-0.032	0.745	-2.664	4.624	3894
Average Room Jump Distance	0.024	0.860	-3.472	2.469	3897
Average Room Sit-Ups	0.024	0.777	-3.460	2.822	3882
Average Room Muscular Endurance	0.049	0.948	-1.387	15.876	3660
Average Room high school ranking	0.017	0.720	-0.743	2.939	6948
Room Difference in Independent Thinking	1.229	1.010	0.000	5.086	6909
Room Difference in Social Concern	1.219	1.000	0.000	4.662	6909
Room Difference in Self-Discipline	1.211	1.016	0.000	4.945	6909
Room Difference in Interaction Ability	1.215	0.991	0.000	4.512	6909
Room Difference in Creativity Ability	1.250	1.010	0.000	4.602	6909
Room Difference in Father's Years of Education	1.193	1.016	0.000	5.035	6948
Room Difference in Number of Siblings	1.072	1.116	0.000	5.445	6945
Room Difference in Birth Order	1.171	0.997	0.000	2.153	6945
Room Difference in Family Income	1.173	1.112	0.000	4.603	6963
Room Difference in Body Mass Index (BMI)	0.863	1.232	0.000	35.535	3891
Room Difference in Forward Flexion	1.211	0.959	0.000	8.605	3894
Room Difference in Jump Distance	0.892	0.788	0.000	5.809	3897
Room Difference in Sit-Ups	1.168	0.920	0.000	4.875	3882
Room Difference in Muscular Endurance	0.544	0.838	0.000	17.043	3660
Number of Departments in the Room	1.726	0.718	1.000	4.000	7206
Room Difference in high school ranking	1.121	1.094	0.000	3.683	6948
Number of Students in the room	2.704	0.919	2.000	4.000	7206
Number of STEM students in the room	1.749	1.159	0.000	4.000	7206
Baseline usage	5.800	4.363	0.000	33.167	7206

Table A-3: Robustness Check: Treatment-effect heterogeneity: include vs. exclude interaction terms

	(1) Usage Log	(2) Usage Log	(3) Usage Log	(4) Usage Log	(5) Usage Log	(6) Usage Log
Treatment	-0.333***	-0.462***	-0.333***	-0.297***	-0.336***	-0.333***
	(0.0284)	(0.0213)	(0.0285)	(0.0217)	(0.0285)	(0.0282)
Treatment × Baseline Usage	-0.0191***		-0.0191***	-0.0199***	-0.0188***	-0.0191***
	(0.00357)		(0.00358)	(0.00356)	(0.00356)	(0.00356)
Treatment × Average Room Creativity Ability	0.0438**	0.0427**		0.0410**	0.0430**	0.0437**
	(0.0207)	(0.0217)		(0.0205)	(0.0205)	(0.0206)
Treatment × Average Room High School Ranking	0.0275**	0.0401***	0.0253*		0.0274**	0.0274**
	(0.0132)	(0.0133)	(0.0131)		(0.0131)	(0.0132)
Treatment × Average Room Father's Years of Education	0.0264	0.0161	0.0250	0.0262		0.0264
	(0.0204)	(0.0211)	(0.0203)	(0.0204)		(0.0204)
Treatment × meanroom_fyosstd	-0.00149	0.00514	0.00144	-0.000283	-0.00225	
	(0.0195)	(0.0203)	(0.0197)	(0.0196)	(0.0196)	
Constant	1.690***	1.690***	1.690***	1.690***	1.690***	1.690***
	(0.00616)	(0.00630)	(0.00618)	(0.00617)	(0.00617)	(0.00616)
Observations	3870	3870	3870	3870	3870	3870
Within R-squared	0.112	0.108	0.111	0.112	0.112	0.112

Notes: The outcome variable is electricity usage (kWh) per person per week during each period. Standard errors in parentheses are clustered at the level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Treatments pool together the quintile-based and average-based treatments. Additional controls corresponding to the covariates included in the interaction terms are included in the regressions but not reported. We sequentially exclude interaction terms based on their variable importance in the causal forest analysis. Specifically, column (2) omits the most important variable, while column (6) omits the fifth most important variable used for splitting in the heterogeneous treatment effect (HTE) trees.

Table A-4: Robustness Check: Heterogeneous Impact of Quintile and Average Treatment on Electricity Usage (Interaction by quintiles)

	(1)	(2)	(3)	(4)	(5)	(6)
		Usage Level			Usage Log	
Quintile Treatment	0.226	0.864	0.681	-0.042	-0.091	-0.153
	(0.759)	(0.638)	(0.548)	(0.115)	(0.173)	(0.121)
Average Treatment	-0.037	0.820*	0.697*	-0.041	-0.038	-0.063
	(0.510)	(0.488)	(0.419)	(0.084)	(0.121)	(0.100)
Quintile x Low users (1st quintile)	-0.761	2.373***	2.016***	-0.017	0.550***	0.466***
	(0.715)	(0.635)	(0.490)	(0.126)	(0.189)	(0.121)
Quintile x Low users (2nd quintile)	0.051	0.964	1.241**	0.138	0.111	0.152
	(0.745)	(0.637)	(0.504)	(0.123)	(0.185)	(0.123)
Quintile x High users (4th quintile)	-0.401	-2.657**	-2.637***	-0.138	-0.254	-0.276
	(1.080)	(1.242)	(0.822)	(0.151)	(0.290)	(0.183)
Quintile x High users (5th quintile)	-0.768	-11.686***	-11.828***	-0.250*	-0.612*	-0.570***
	(1.354)	(2.507)	(1.699)	(0.145)	(0.325)	(0.212)
Average x Low users (1st quintile)	-0.225	2.152***	2.419***	0.097	0.412***	0.453***
	(0.471)	(0.554)	(0.422)	(0.103)	(0.147)	(0.111)
Average x Low users (2nd quintile)	-0.571	0.438	0.729*	-0.068	-0.057	0.005
	(0.499)	(0.479)	(0.373)	(0.097)	(0.135)	(0.105)
Average x High users (4th quintile)	0.181	-2.276**	-2.732***	-0.054	-0.257	-0.350**
	(0.816)	(1.127)	(0.694)	(0.113)	(0.248)	(0.161)
Average x High users (5th quintile)	-0.429	-11.864***	-10.653***	-0.253**	-0.882***	-0.647***
	(1.118)	(2.088)	(1.367)	(0.123)	(0.275)	(0.180)
N	13446	12399	20169	13446	12399	20169
Individual FE	v	v	v	v	v	v
Time FE	v	v	v	v	v	v
Whole sample, pre+post 1 period	v			v		
All periods, excluding group treated twice		v			v	
Whole sample, all periods			v			v

Notes: The outcome variable is electricity usage (kWh) per person per week during each period. Standard errors in parentheses are clustered at the room level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Quintile and average treatments are interacted with high users (4th and 5th quintiles) and low users (1st and 2nd quintiles). The omitted category is the third quintile. The IV variable uses the original treatment status, and the endogenous variable for those actually treated includes those who have answered the survey.

Table A-5: Robustness Check:Heterogeneous Impact of Quintile and Average Treatment on Electricity Usage (Interacting with Median)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Usage Level			Usage Log			Adjusted Rank (the lower, the better)		
Quintile Treatmen	-0.133	2.422***	2.192***	0.001	0.207**	0.131**	0.321**	1.047***	0.990***
	(0.326)	(0.336)	(0.291)	(0.063)	(0.083)	(0.063)	(0.135)	(0.162)	(0.132)
Average Treatment	-0.397	1.961***	2.057***	-0.029	0.123*	0.130**	0.290**	0.894***	0.966***
	(0.286)	(0.305)	(0.281)	(0.059)	(0.072)	(0.062)	(0.131)	(0.148)	(0.130)
Quintile x High users	-0.075	-7.228***	-7.010***	-0.188**	-0.608***	-0.595***	-1.094***	-3.112***	-2.970***
	(0.643)	(1.007)	(0.673)	(0.088)	(0.168)	(0.111)	(0.217)	(0.384)	(0.246)
Average x High users	0.325	-6.306***	-6.607***	-0.122	-0.570***	-0.559***	-1.189***	-2.827***	-2.966***
	(0.566)	(0.855)	(0.607)	(0.078)	(0.148)	(0.103)	(0.212)	(0.347)	(0.241)
N	13446	12399	20169	13446	12399	20169	13446	12399	20169
Individual FE	v	v	v	v	v	v	v	v	v
Time FE	v	v	v	v	v	v	v	v	v
Whole sample, pre+post 1 period	v			v			v		
All periods, excluding groups treated twice		v			v			v	
Whole sample, all periods			v			v			v

Notes: The outcome variable is electricity usage (kWh) per person per week during each period. Standard errors in parentheses are clustered at the room level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Quintile and average treatments are interacted with users higher than or equal to the median electricity usage. The IV variable uses the original treatment status, and the endogenous variable for those actually treated includes those who have answered the survey.

Appendix B Causal Forest

With the growing application of machine learning in economics, causal forests have been widely used to estimate Conditional Average Treatment Effects (CATE) in various contexts. In this study, we apply the Generalized Random Forest (GRF) algorithm introduced by Wager and Athey (2018) to estimate heterogeneous treatment effects (HTE) in response to an electricity consumption feedback intervention.

As our study is based on a randomized controlled trial (RCT), we can directly estimate heterogeneous treatment effects using causal forests, a machine learning approach that extends standard random forests to treatment effect estimation. Instead of estimating the overall average treatment effect (ATE), we focus on the interaction term, which captures how treatment effects differ for high users relative to low users—a distinction motivated by previous findings. The intervention consisted of social comparison messages providing feedback on electricity usage. Since the two message types (quintile ranking and mean comparison) had similar ATEs, we merge them into a single treatment group for this analysis. This choice improves statistical power by increasing the sample size, leading to more efficient tree splitting in the causal forest model. Larger sample size for only merged one treatment can prevent overfitting as growing separate causal forests for each treatment group could lead to small sample sizes and unstable estimates. In addition, given the lack of strong theoretical priors on how covariates in our model influence heterogeneity, this serves as an exploratory analysis for future policy targeting. For example, our covariates include novel traits and physical characteristics, which may reflect students' dorm dynamics—such as social interactions, self-discipline, and room composition. By consolidating the treatment groups, we provide a broader perspective on which room-level factors matter most, making the findings more generalizable for policy decisions.

We first estimate a single treatment effect and introduce an interaction term capturing differences in treatment effects for high electricity users:[16]

$$Y_i = c_i + \gamma_1 W_i + \varepsilon_i, \tag{8}$$

where W_i is an indicator for whether an individual was assigned to either one of the treatments; γ_1 estimates the average treatment effect.

The heterogeneous effect is of our interest, and is defined as:

$$\tau(x) = \mathbb{E}\left[Y_i(W_i = 1) - Y_i(W_i = 0) \mid X_i = x\right]. \tag{9}$$

where $\tau(x)$ captures the difference in treatment effects conditional on covariates.

For GRF implementation, it does not directly predict the outcome variable as the random forest algorithm. Instead, GRF partitions the data adaptively to maximize heterogeneity in treatment effects and estimates CATE. In our case, we focus on maximizing the heterogeneity of the interaction term. This approach ensures that the algorithm prioritize splits that creates the greatest distinction in treatment effects between these subgroups, and to identify the key

covariates driving the variation in the impact of the behavioral responses to the intervention. Under this framework, the GRF calculates solutions based on the following equation:

$$\hat{\tau}(x), \hat{c}(x) \in \arg\min_{\tau,c} \left\| \sum_i \alpha_i(x) \cdot (Y_i - \tau(x)(W_i) - c(x)) \begin{pmatrix} 1 \\ W_i \end{pmatrix} \right\|_2 \tag{10}$$

$\tau(x)$ is the heterogeneous treatment effect (HTE); $c(x)$ is the constant term; $\alpha_i(x)$ is an adaptive weighting function ensuring efficient estimation. This formulation ensures that the residuals are orthogonal to the weighting function, leading to a treatment effect estimator that is robust to misspecification and optimally partitions the feature space to maximize heterogeneity in treatment responses. This moment condition ensures that treatment effect heterogeneity is estimated efficiently with a locally adaptive weighting scheme, using similar notations and conditions as described in Murakami et al. (2022).

For the implementation of the algorithm, GRF aggregates multiple trees grown on bootstrap subsamples to form an ensemble estimator. Each tree is constructed by recursively splitting the covariate space, where, at each node, a subset of candidate variables is randomly selected, and the optimal split is chosen to maximize heterogeneity in treatment effects (in our case, the heterogeneity between the differential treatment effects of high users and low users).

Unlike Random Forest, which seeks to improve prediction accuracy, GRF intends to identifying features that create the most distinct subgroup in terms of treatment effects. The tree-growing process follows an honest estimation approach–one part of data is used for determining where to split trees, and another independent part is used to estimate treatment effects. This prevents overfitting and ensures that estimated effects remain valid for inference. The imeplementation for the CF under the GRF framework is as follows:

(i) Draw a total of S random subsamples from the full dataset without replacement. This ensures that each tree is trained on a different subset of the data, introducing variability and reducing overfitting. The number of trees, denoted as B, is set to 2000 in our case.

(ii) Each subsample is divided into two seperate parts. The first subset S_1 is used for growing trees and determining the optimal splits. The second subset, S_2, is for estimating effects in the leaves. This honest estimation procedure ensures that the same observations are not used for both splitting and effect estimation, thereby reducing overfitting. By default, the estimation fraction is set to 50%.

(iii) The trees are grown recursively by selecting optimal covariates for splitting. The GRF maximize heterogeneity in the estimated treatment effects when deciding splits. This ensures that each leaf node contains units that are as homogeneous as possible in their estimated treatment effects.

(iv) Once the trees are grown, the estimated sample S_2 is used to compute the adaptive weights $\alpha_i(x)$. The weight for each observation depends on how frequently it appears in the same leaf across trees. These weights ensure that the treatment effects are estimated efficiently and reduce bias in the final results

(v) The above steps are repeated $B = 2000$ times, generating 2000 causal trees. Each tree produces an estimate of the heterogeneous treatment effect $\tau(x)$. The final treatment effect estimate is then obtained by averaging over all trees, ensuring robustness and stability in the estimation.

This honest estimation approach makes sure that treatment effects are not overfitted to the data used for splitting, resulting in valid statistical inference. The combination of honest estimation, adaptive weighting, and tree-based heterogeneity maximization allows GRF to flexibly estimate heterogeneous treatment effects, particularly in our settings where interaction between treatment and subgroup (e.g. high vs. low electricity users) are of primary interest.

Appendix C Nash Bargaining Model Illustration

Conceptual framework and interpretation. To clarify how bargaining power and reservation conditions interact in our setting, we outline a simple Nash bargaining model that can be interpreted in terms of roommates jointly determining an outcome such as comfort temperature or collective energy use.

Two-person model. Consider two individuals, A and B, each with an ideal outcome T_i^* (for instance, their most comfortable temperature) and utility

$$u_i(T) = -(T - T_i^*)^2.$$

Each has a reservation utility $d_i = u_i(T_{0i}) = -(T_{0i} - T_i^*)^2$, representing the payoff they would obtain under the status quo or their own set point T_{0i}. The Nash bargaining solution chooses $T^\star$ to maximize the weighted Nash product

$$\max_T (u_A(T) - d_A)^{\alpha_A} (u_B(T) - d_B)^{\alpha_B},$$

where α_A and $\alpha_B = 1 - \alpha_A$ are bargaining weights that capture the relative influence of each individual in negotiation.

Predictions from bargaining power. In this symmetric quadratic case, the interior target outcome is the weighted average of both ideals,

$$T^{\text{target}} = \alpha_A T_A^* + (1 - \alpha_A) T_B^*,$$

which moves linearly with α_A. Thus, α_A determines the *direction and strength* of the compromise: when α_A increases, the joint outcome shifts toward A's preferred level, and when $\alpha_A = 0.5$ both parties have equal influence and the outcome is the midpoint between their ideals.

Reservation points and feasibility. The negotiated outcome must make each party at least as well off as at their reservation point, $u_i(T^\star) \geq d_i$. Define each individual's "comfort interval" as

$$I_i = [T_i^* - r_i,\ T_i^* + r_i], \qquad r_i = |T_{0i} - T_i^*|.$$

Cooperation is feasible only if the two intervals overlap, that is, if

$$I_A \cap I_B \neq \varnothing \quad \Longleftrightarrow \quad |T_A^* - T_B^*| \leq r_A + r_B.$$

If the overlap is empty, no mutually beneficial agreement exists and the outcome remains at the status quo set point. When cooperation is feasible, the actual Nash outcome $T^\star$ equals the target T^{target} projected onto the intersection $I_A \cap I_B$. In this sense, bargaining power determines *where* the group aims to settle, while the reservation utilities determine *whether* such a settlement is attainable.

Comparative statics and axioms. The model exhibits clear comparative statics:

$$\frac{\partial T^{\text{target}}}{\partial \alpha_A} = T_A^* - T_B^*, \quad \frac{\partial T^{\text{target}}}{\partial T_A^*} = \alpha_A, \quad \frac{\partial T^{\text{target}}}{\partial T_B^*} = 1 - \alpha_A.$$

Larger bargaining power for one party pulls the outcome toward that party's ideal, and wider tolerance ranges r_i enlarge the feasible region. The Nash solution further satisfies the standard axioms of *Pareto efficiency*, *individual rationality*, *symmetry* (if $\alpha_A = \alpha_B$ and $d_A = d_B$), *affine invariance* (invariance to linear scaling of utilities), *independence of irrelevant alternatives*, and *uniqueness*. These axioms imply that the Nash outcome is fair, efficient, and robust to how utilities are measured.

Extension to three or more individuals. When there are n roommates, the same reasoning extends naturally. Each individual i has bargaining weight $\alpha_i > 0$ with $\sum_i \alpha_i = 1$ and ideal T_i^*. The interior target becomes

$$T^{\text{target}} = \sum_{i=1}^{n} \alpha_i T_i$$

and cooperation is feasible if the overlap of all comfort intervals is nonempty, $\bigcap_i I_i \neq \varnothing$. As the number of participants increases, it becomes more difficult for all intervals to overlap: a single highly constrained or extreme roommate can act as a veto player. When the intersection is nonempty, $T^\star$ equals the projection of T^{target} onto this common overlap. Power thus governs the direction of adjustment, while the feasibility region, shaped by reservation utilities, determines whether cooperation succeeds.

Empirical interpretation and link to machine learning variables. This theoretical framework clarifies how the model maps into our empirical variables. In our data, *baseline usage* and *room-average creativity* capture the underlying conditions that influence the reservation utilities d_i—analogous to each roommate's set point or status-quo comfort level. Rooms with higher baseline usage or greater creative activity start from different equilibrium conditions, affecting how much further improvement is feasible through cooperation. The variable capturing *room difference in sit-up behavior* reflects within-room heterogeneity and thus the relative bargaining weights α_i: larger differences suggest unequal influence or negotiation power among roommates. This difference may also affect the tolerance ranges r_i (physical or behavioral flexibility), linking it to both the power and the reservation-point dimensions. Hence, in our empirical framework, bargaining power (α_i) shapes the direction of behavioral adjustment, while reservation-related factors (d_i) determine its attainable magnitude. The joint interaction of these forces depends on its correlation with reservation conditions: when influential individuals also have higher comfort or baseline capacity, their preferences dominate both through power and through favorable outside options.