# Electricity Use Application: User Study Report

Lothaire Aubergeon Ian Hutter Sophie Leichtle Marco Schöb Markus Wagner

## 1 Introduction

The goal of our project is to create an app that incentivizes its users to consume less electricity in their daily life. The motivation for this is to provide an approach to solve the problem of electricity overuse in the world, as it is a key contributor to the ongoing climate change crisis [1]. An assumption for this project is that we would have access to futuristic smart home technology that can get exact usage information from all devices in a users home. This is not entirely unrealistic as the concept of "Internet of Things" becomes more and more relevant [2].

Our approach had to solve two key problems: the first being that in general, most households are not actually aware of how much electricity they consume on a daily basis and especially not of which devices the major factors contributing to the overall uses are [3]. Studies have shown that people tend to either vastly underestimate or greatly overestimate an objects power consumption, and that just making users aware of the actual values already helps a great deal with lowering electricity consumption [3].

The second problem was how to further incentivise lower consumption. In our research, we came across multiple cases where some sort of peer or social pressure positively influenced power usage [4], so we decided to try out a gamification approach. This was done by defining a metric that acts as both a score for social comparison and a currency to buy cosmetics for the app with. Every day the user starts with the max currency and as their home uses more electricity, they get deducted currency, which finally is handed out to be used at the end of each day. As previously mentioned, this currency is used as a score users can compare themselves with, producing a ranking that further incentivizes them through competition.

Our app layout contains five pages, those being a dashboard, a rankings page, a shop page, an analytics page and a settings page. For our A-B Test we made two prototypes: A being the "fancy" one and B being a straightforward simple one. Only the first four pages are relevant for the user study, so we will briefly go into more detail on them in the following sections.

### 1.1 The Dashboard

The dashboard is the landing page of the app, so it needs to contain the most basic information which the user needs. In our case, this is the currency the user has banked at the moment as well as basic tips on how to improve their score/consumption, which we decided to implement as "challenges" that reward the user with currency on completion. The difference between the two prototypes A and B on this page consists of how the challenges are displayed. In the fancy prototype, the challenges are in the form of a list of stylized cards that contain all information and can be accepted on the fly, whereas in the second prototype they are a simple list where challenges can be accepted in a popup.

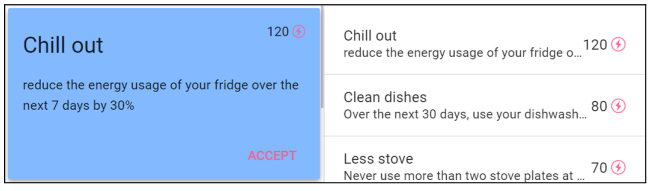


Figure 1: A vs B Dashboard

### 1.2 The Rankings

This page is where users compare themselves with friends, family and global users, their scores being displayed in a ranked list. Furthermore, the two prototypes differ in how a specific ranking is selected from different timescales, e.g. weekly, daily, etc., and user groups. In prototype A, users switch between user groups and timescale using arrows, whereas in prototype B they can select them with two dropdown menus.

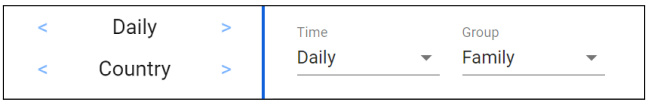


Figure 2: A vs B Rankings

### 1.3 The Shop

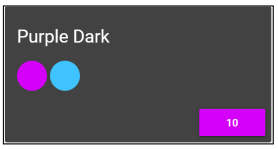


Figure 3: Shop Item

The shop is where users can spend their earned currency on buying cosmetics, which in the current prototype just include themes for the app but would be expanded to profile customization and more in a final product. The basic idea of the shop is that items are represented by a card that contains its name, basic information about the item and a buy/equip button. The difference between prototypes concerns bought items. Version A has all items, purchased or not, in the same list that can be filtered and searched through. Version B on the other hand has a tab for purchasable and one for purchased items. While this is a subtle change, the idea was that version A might lead the user to discover more items they would like to purchase, as they always have to interact with new items. In version B, when just equipping an already bought item, the user can ignore the shop as the purchases are separate.

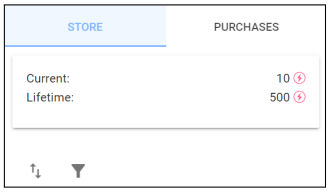


Figure 4: Shop Interface in B. Same Interface in A but without the Tabs.

### 1.4 The Analytics

The analytics page is what, in conjunction with the tips/challenges, informs the user of their consumption and tells them how to improve. Both versions include a graph, called consumption graph, that informs the user about overall usage over time but differ in how device specific information is displayed. Version A has a "bubble" view, which shows categories as bubbles of different sizes, with larger bubbles corresponding to larger consumptions. Clicking on a category bubble would then lead to another bubble view with a list

of devices in that specific category. Version B uses the same concept, but visualizes it using a bar chart instead.

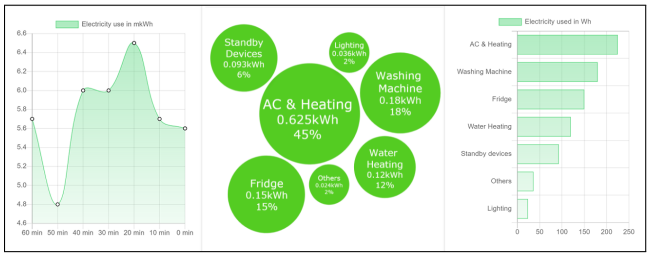


Figure 5: From Left to Right: Consumption Graph, Bubble View in A, Bar Chart in B

## 2 Study Design

### 2.1 Apparatus

During our study, the users were given an Android phone - with a screen size of at least 6 inches - with our application running in a browser. After scrolling, the navigation bar of the browser disappeared, thus giving the users an experience similar to a native application. The participants were seated at a table during the entirety of the study.

### 2.2 Independent Variable

The independent variable of our study is the interface layout of our application. The interface changes between the A and B versions were described in details in section 1.

### 2.3 Dependent Variables

During our study, we decided to automatically track a number of metrics during each round of the experiment. The time taken by the user to complete all tasks for a given version of the application, the time spent by the user on each screen of the application and the global amount of clicks until the completion of all tasks were recorded. Moreover, each participant was asked to fill three questionnaires (one after each round of the study to get feedback on each version, and one at the end to compare both versions) to measure user satisfaction as precisely as possible.

### 2.4 Hypotheses

The null hypotheses that we used for our user study are listed below.

- Null Hypothesis 1: The content structure has no effect on the task completion time.
- Null Hypothesis 2: The content structure has no effect on the amount of clicks needed to complete the tasks.
- Null Hypothesis 3: The content structure has no effect on the user satisfaction.

### 2.5 Experimental Procedure

The recruitment of users for our study was done using convenience sampling. The participants were thus friends and family members that agreed to participate in our study. The study was run with one participant at a time over multiple days.

The exact study procedure is described below followed by the task list for prototype A and B.

- After welcoming each participant, they were attributed a unique ID and asked to fill out a demographic questionnaire.
- Each participant was then quickly introduced to the prototype; the goal of our application was presented and all the screens of our prototype were explained along with their basic functionalities. We then explained the users that we would be measuring the way they interact with the prototype with a list of tasks to execute, which are different for each version of the prototype.
- Subsequently, we gave an answer sheet to each participant with empty spaces into which they could write the values they found after they had completed each task.
- Next, the participants were given four tasks to fulfil for the current prototype. They are listed below.

The behavior of the participants was recorded in the background while they were performing the tasks. The data recorded can be found in section 3. After completing all four of the tasks of the first version of the prototype, the participants were asked to fill out a System Usability Survey (SUS) questionnaire [5] about the given version. They then repeated the task list for the other version of the prototype. After doing that, they were asked to fill out a fourth questionnaire comparing the two versions of the prototype. Finally, each participant was given the opportunity to make comments about the prototypes or ask any questions. The comments were subsequently noted.

Task list of prototype A:

1. Locate the challenges section on the dashboard, locate and accept the dishwasher challenge and solar panel challenge, fill the reward into the sheet.
2. Move on to the rankings, find your place on the monthly worldwide ranking and the yearly friends ranking and fill it into the sheet.
3. Move on to the usage page, find the electricity usage of your fridge over the last 24h and the usage of your chargers over the last month and fill it into the sheet.
4. Move on to the shop, find, buy and equip the Purple Dark theme.

Task list of prototype B:

1. Locate the challenges section on the dashboard, locate and accept the stove challenge and the TV challenge, fill in the reward on the sheet.
2. Move on to the rankings, find your place on the yearly country ranking and the weekly worldwide ranking and fill it into the sheet.
3. Move on to the usage page, find the electricity usage of your lighting over the last 24h and the TV box over the last month and fill it into the sheet.
4. Move on to the shop, find, buy and equip the Purple Light theme.

### 2.6 Participants

We selected a total of 14 participants for our study. Slightly more people identified as male than female. Furthermore, almost half of the participants were in the age group of 20-24. The complete demographic makeup can be seen in Figure 6.

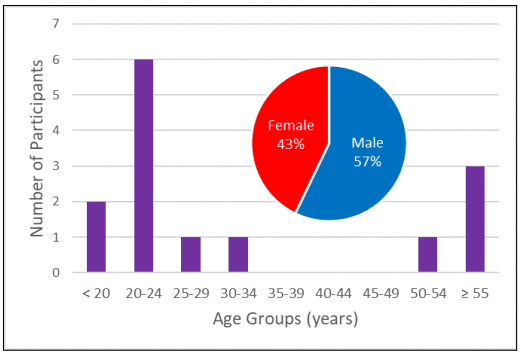


Figure 6: Age and Gender Distribution of the AB-Test Participants

## 3 Results

While the participants were using the app, the number of clicks and the amount of time they needed for each task and for the app in total was recorded and afterwards written into an Excel sheet. For the analysis, this data was split into two subgroups: version A and version B data, which both were further subdivided into round 1 and 2 data, indicating, if the participant first used version A or B. Data from version A or B was only compared if it was recorded in the same round. This was done on the presumption that the participants generally know where to find the different task items after completing their first round.

### 3.1 Numerical Data

During the data analysis all null hypotheses underwent a significance test with a 95% confidence interval.

*First Null Hypothesis*
In the case of the first null hypothesis, the average task time difference of version A minus version B, as well as their respective 95% confidence intervals were calculated and subsequently plotted for round 1 and round 2 (Figure 7).

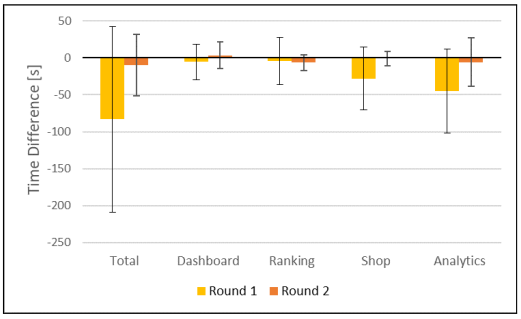


Figure 7: First Null Hypothesis: Average Task Time Difference (A-B) with 95% CI's

It can be observed that irrespective of the round, all confidence intervals include zero, even if some of the means changed signs. This indicates that the null hypothesis is valid i.e. the amount of time the participants spent on the tasks is independent of the version they used. However it can also be stated that the confidence intervals of round 2 have shrunk significantly compared to round 1, which strengthens our previously stated presumption that the participants were generally more able to orient themselves in the app as they were in the first round.

*Second Null Hypothesis*
The same procedure was subsequently also applied to the second null hypothesis (see Figure 8).

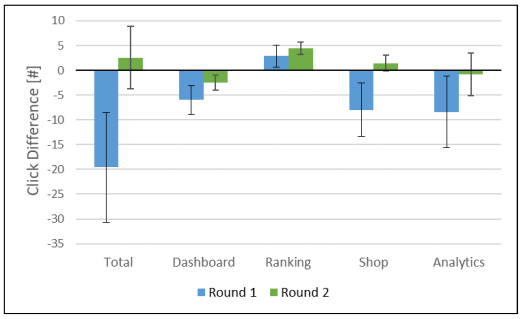


Figure 8: Second Null Hypothesis: Average Amount of Clicks Difference (A-B) with 95% CI's

For the second hypothesis, the confidence intervals do not include zero for round 1, which could have indicated an invalidation of the null hypothesis if the results of the second round did not partially contradict them. However in both rounds, the difference between the amount of clicks for the dashboard (more clicks needed for version B) and for the ranking (more clicks needed for version A) proved to be statistically relevant. Similarly to the first, the absolute values of the confidence intervals decreased for the second round.

*Third Null Hypothesis*
To check the third null hypothesis, the average difference of the SUS scores also including the 95% confidence interval of both rounds was calculated and plotted (see Figure 9, left).

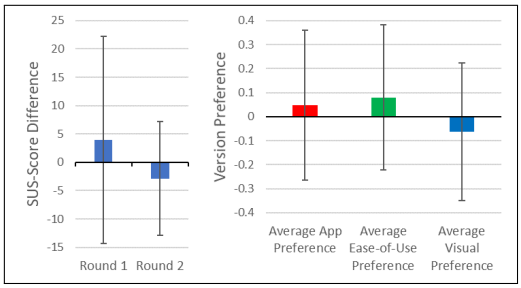


Figure 9: Third Null Hypothesis: Average SUS-Score Difference (A-B, left), Normalized AB-Version Preference (A: pos | B: neg, right); Both with 95% CI's

Like with the first null hypothesis, we can also see here that both CI's include zero, while the mean changes the sign from plus to minus (indicating that the participants were more in favour of version B when they used version A first). The same holds true for the absolute values of the CI's: as with the two hypotheses before, they shrunk for round 2.

*Contraindication*
However, the slight preference of version B in the second round stands in contrast to the data from the third questionnaire. Here, participants were asked which version they preferred (overall, on an ease-to-use basis and visually) on a scale from 1 (version A) to 10 (version B). By subtracting the average of this scale (5.5) from all the provided answers and normalizing and inverting the results, a new scale with a range from -1 to 1 was created on which negative values indicate a preference for version B and positive values a

preference for version A. Like with the null hypotheses, the means and its CI's were calculated and plotted (see Figure 9, right).

Similar to the SUS-Score differences, these results also include zero in their confidence intervals. However, the results indicate that the participants liked version A overall more and found it easier to use, while version B was just more visually appealing to them.

*Summary*
Overall, all null hypotheses were validated by the results. However, the amount of clicks recorded for the dashboard sub task was lower for version A and for the one regarding the ranking lower for version B. These differences can, however, be explained by the different setups of the task functionalities for these pages, which necessitate more clicks to complete the test tasks. There have also been contradicting results for the user preference of either version. In the end, both results are not conclusive though, again validating the third null hypothesis.

### 3.2 Verbal Feedback

In addition to the numerical results, we collected verbal feedback, which were very mixed. As it was expected for a hi-fi prototype, a lot of the verbal feedback was focused on very specific details since not all (sub)functionalities were already fully implemented.
In the dashboard, a common comment was that, while the colored cards in A did not look good, the instant accepting of challenges without a popup felt more intuitive. Other than that, participants were divided on whether the card view or list view was better. Unrelated to the A-B test, another common comment was that the sand-watch like visualization for banked currency, while intuitive, needed polishing and, if possible, some sort of history.
In the rankings page, participants generally agreed that the dropdown menu was easier to use in order to navigate to a specific page. However, some still preferred the arrows as they felt nicer to use, even if the participants were slightly slower in fulfilling the task.
In the shop, the most common feedback was that the filters were unnecessary as it was much easier to just scroll down to the desired theme and select it. This was, of course, one of the limitations of our study, and is not representative of how the finished app would work. Additionally, some participants were slightly confused about the colors of the cards and buttons in the shop. We could mitigate this by giving better indication that the shop item colors reflect the theme it represents.
The analytics section was by far the most criticised one. This was what we expected, as it was the one with both the biggest difference between the versions as well as the one, which was hardest to implement for our prototype. In both versions, finding information for a specific device was perceived as unintuitive, mainly because it was not made clear what parts of the graphs were interactable. It did not help that, due to the limited scope of this project, not all bubbles were functional. While the bubble graph was generally the preferred visualization, participants often criticised the way it looked, likely because it was not an actual implementation but a static image used as it could not be implemented directly with a reasonable amount of work.
When looking at the statistics and the verbal feedback together, the idea that the two versions were not that drastically different both regarding efficiency and user experience was reinforced. Which layout was better and/or preferred differed on a user to user basis, so no clear winner could be chosen. However, the feedback was still very useful as it helped us better understand some of the statistics, as well as hone in on specific details that would otherwise not show up in simple time and click statistics. When refining the prototype, this will help us choose the best features and designs from either prototype.

## 4 Limitations

The first evident limitation to our user study is the limited number of participants. Indeed, to obtain more precise and meaningful results, the number of participants needs to be greatly increased. Secondly, our prototype was run on a mobile browser, which caused some issues with the browser navigation bar. This would not have happened, had we used a native application. We also used different models of phones for our study, which could lead to some minor differences in results.
The number of themes available in our prototype (20) was too low to truly test our search system, as we are planning for around 1000 themes to be available in the final product. This lead to users bypassing the system entirely, as they skipped the possibility of using filters and simply scrolled down until they found what they were looking for. This did not generate useful feedback. Furthermore, the rough implementation of the analytics page lead to some confusion during the testing. This muddled the feedback as a lot of it was concerned with the specifics of what was still missing instead of the intended function.
Moreover, the dependant variables that we chose are no perfect indicators of efficiency. Finally, the standard deviation of the age of the participants is too big, which probably had an impact on our results. Indeed, we can distinguish two groups of participants: young adults spending multiple hours each day on their phones and older adults less experienced with mobile user interfaces.

## 5 Future Work

Regarding the future and how work on this or similar projects could continue, the main takeaway is that this type of testing does not lend itself well to our kind of project. While the verbal feedback was useful for tweaking small usability aspects of the app, the time and click statistics do not say much about the success of the implementation. We believe that the standard usage of the app is not influenced much by whether individual tasks take a second longer or not, as those would generally only be executed a few times each day or even week. Since the concept of the app is to motivate users on a long-term basis, success could only properly be measured after a study covering a much longer time-frame. This in turn does not lend itself well to simple A-B testing. Therefore, a different approach would need to be used.

## 6 Conclusion

All in all, the results from this user study were rather encouraging. While the numerical comparisons between the two versions were inconclusive, we got valuable feedback from testers that will help us choose features from each version to keep for a final prototype. Furthermore, as discussed in the results section, testers liked the general layout and idea of the app which reinforces our belief in its potential usefulness.
Unfortunately, since this was just a prototype, we did not get to test anything portending to how effective the app would actually be. While our assumptions, i.e. perfect device data through the Internet of Things, do not currently hold, it is not too far fetched that within the next years this kind of technology becomes viable. Once that happens, we could definitely see this kind of app becoming a successful tool to motivate people to reduce energy usage, even if it is likely that it would not work as a sole motivator.
In conclusion, we think that we have succeeded both regarding the stated goal of our app as well as the goal of this user study. Our app provided a good proof of concept for a future possibility, and the user study helped us refine this into something that could easily be expanded upon in future iterations. It also helped us better understand what kind of tests and studies are useful for certain metrics and use cases, which will help us improve user testing for future projects.

## REFERENCES


[1] X. Li and D. J. Sailor, "Electricity use sensitivity to climate and climate change," *World Resource Review*, vol. 7, 9 1995.

[2] L. Tan and N. Wang, "Future internet: The internet of things," in *2010 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE)*, vol. 5, pp. V5–376–V5–380, 2010.

[3] D. Brounen, N. Kok, and J. Quigley, "Energy literacy, awareness, and conservation behavior of residential households," *Energy Economics*, vol. 38, p. 42–50, 07 2013.

[4] F. Abdallah, S. Basurra, and M. Gaber, *An Agent-Based Collective Model to Simulate Peer Pressure Effect on Energy Consumption: 10th International Conference, ICCCI 2018, Bristol, UK, September 5-7, 2018, Proceedings, Part I*, pp. 283–296. 01 2018.

[5] J. Brooke, "Sus: A quick and dirty usability scale," *Usability Eval. Ind.*, vol. 189, 11 1995.