Energy & Buildings 232 (2021) 110652

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enb

Residential electricity conservation in response to auto-generated, multi-featured, personalized eco-feedback designed for large scale applications with utilities

Christoph J. Meinrenken [a,b,*], Sanjmeet Abrol [a], Gaurav B. Gite [a], Christopher Hidey [d], Kathleen McKeown [a,d], Ali Mehmani [a], Vijay Modi [a,c], Elsbeth C. Turcan [d], Wanlin Xie [d], Patricia J. Culligan [a,b,e,f]

[a] Data Science Institute, Columbia University, New York, USA
[b] Earth Institute, Columbia University, New York, USA
[c] Dep. of Mechanical Engineering, Columbia University, New York, USA
[d] Dep. of Computer Science, Columbia University, New York, USA
[e] Dep. of Civil Eng. and Eng. Mechanics, Columbia University, NY, USA
[f] Dep. of Civil & Env. Eng. & Earth Sciences, U. of Notre Dame, Notre Dame, USA

A R T I C L E I N F O

Article history:
Received 6 October 2020
Revised 28 November 2020
Accepted 7 December 2020
Available online 11 December 2020

Keywords:
Energy conservation
Electricity usage
Eco-feedback
Social comparisons
Natural language processing
Boomerang effect
Sentiment
Mean reversion

A B S T R A C T

While past research has shown that providing residents with feedback about their electricity usage can reduce demand and its associated environmental burdens, some questions remain regarding what makes such feedback most effective. We followed the electricity usage of 36 residents who each received 14 feedback messages over 2 months. Using approaches borrowed from Natural-Language-Processing, feedbacks were generated automatically, using 10 features in random combinations. Unlike in previous studies, each resident received varying types of messages over time. In 504 observations, the average prompted reduction in electricity usage was 11 ± 3%, compared to a control group of 89 residents who received no messages. Feedback types prompting the largest reductions were self-comparisons with one's own earlier usage (average reduction 14%) and messages of high variety from one feedback-cycle to the next (average reduction 16%). Comparisons with neighbors did not prompt higher reductions on average. Instead, they prompted reductions only when a resident's recent usage happened to be higher than the average usage of neighbors, and increases when the reverse was true. This behavior was exhibited by all residents and is likely explained by a norm-conforming mean reversion of residents to their neighbors' average usage, rather than an anti-conform "boomerang" behavior previously suggested in similar contexts.

1. Introduction

1.1. Background and previous studies

In 2017, residential buildings used 32% of US electricity [1], an amount responsible for 9% of domestic greenhouse-gas emissions [2]. To reduce usage, residents can install efficient technology (e.g., LED lights or smart devices [3]) or change their behavior (e.g., turning off unneeded lights), with the former sometimes deemed less effective [4] due to the associated out-of-pocket costs.

Most strategies to prompt electricity usage reductions fall into one or more of three categories [5]: (i) Pecuniary incentives such as higher electricity prices [6]; (ii) in-house displays to continuously inform residents about their usage, separately reviewed by Faruqui et al. [7] and recently employed by Schultz et al. [8] and Wood et al. [9]; and (iii) periodic feedbacks via, e.g., door hangers in apartment buildings [10], monthly or quarterly hard copy mailings distributed by utilities such as *OPower's* "home energy report" [11,12], or emails (as in this study). While most studies meter the electricity usage for an apartment or a single family home as a whole, some also consider the consumption of individual devices (e.g., refrigerator, washer/dryer, television set) and/or their link to self-reported household activities [13]. This is done to prompt higher electricity savings by providing residents with more granular information [14–16], and/or to study via which devices residents are able to save the most electricity [17,18]. In this context, consumers' misconceptions about which of their devices contribute the most to total electricity usage have been discussed

* Corresponding author at: 500 West 120 Street (Mudd 918), New York, NY 10027, USA.
E-mail address: cmeinrenken@ei.columbia.edu (C.J. Meinrenken).

https://doi.org/10.1016/j.enbuild.2020.110652
0378-7788/

Nomenclature

R_{date}	Average treatment vs. control on specific date (any date in baseline or feedback period) [Table 2]
BA_r	Baseline adjustment for resident r in treatment group [Table 2]
BA_r	Baseline adjustment for entire treatment group [Table 2]
$R_{date,net}$	Average treatment vs. control on specific date (during feedback period only) [Table 2]
$R_{r,f}$	Effect of treatment of resident r, in response to feedback f [Table 3]
$R_{r,all\ f}$	Effect of treatment on resident r, averaged across all 14 feedback rounds [Table 3]
$r_{t,all\ f}$	Effect of treatment on resident r, averaged across all 14 feedback rounds [Table 3]
$p_{r,f}$	Usage of resident r (relative to control) during days ahead of feedback f [Table 4]
$p^*_{r,f}$	Usage of resident r (relative to other treated apartments) during days ahead of feedback f [Table 4]
$\Delta p_{r,f}$	Change in usage of resident r (relative to control) from prior to after receiving feedback f [Table 4]
SEM	Standard Error of the (Sample) Mean

as a crucial reason for households' inefficient action to reduce usage [19].

With respect to periodic feedbacks, various reviews and meta studies have analyzed which feedback features tend to reduce usage the most [5,7,17,20–27], including how this can inform regulatory policies [28]. For example, past experiments have found that the context in which electricity usage is presented can affect the prompted degree of reduction: Asensio and Delmas found that sending participants in the US health and environment-oriented feedback prompted a 8–10% reduction, which persisted throughout the 14-week experiment (14 emailed feedbacks) while cost-oriented feedback lost its effectiveness after just a few weeks [17]. The higher effectiveness of health-oriented feedback was also shown by Chen et al. in a similar study in India [29]. Similarly, when varying the metric of expressing energy use, Jain et al. found that invoking environmental externalities (e.g., CO_2 emissions) was more effective than direct energy units (kWh) [25].

Overall, previous studies suggest that feedbacks "need to be targeted to be most effective" [26]. But previous research has also pointed to 3 sets of nuanced findings that warrant further study to allow for said targeting:

(i) *Social comparisons*: Many strategies to reduce residential energy usage employ some form of "social marketing" [30] – even if this effect is not always acknowledged by the recipients [31] – and recent qualitative studies indicate that people's social relationships affect their usage [32]. Social comparisons may be particularly useful in settings where a direct financial incentive for energy conservation is lacking because residents do not pay for their own electricity [33]. As discussed in Jain et al. [34] and mirrored in a meta analysis of 156 studies [5], feedbacks that include normative social influence – such as informing residents how their usage fares relative to that of peers – often prompts steeper reductions (11.5%) than information only about their own usage. However, in contrast, Harries et al. [35] have shown with 316 residents, who received real time feedback about their own current and past usage, that adding peer comparisons to such feedback did not significantly increase reductions.

(ii) *"Boomerang" effect*: Feedback about electricity consumption does not always yield the intended effect with all recipients – a phenomenon sometimes referred to as "boomerang effect" [10,12,36]. Asensio and Delmas found that health and environment-oriented feedback led to a 1.9% increase in the lowest decile of baseline electricity users whereas other users reduced their usage [37]. Schultz et al. found that residents who received door hangers comparing their usage to that of neighbors increased their usage if their baseline usage had been below average (again, other residents reduced their usage) [10]. In contrast, Kažukauskas et al., in a similar study using in-house displays, did not observe a boomerang effect in electricity usage [38].

(iii) *Mitigation of boomerang effect:* Alleviating the effect is desirable because it would lead, all else being equal, to larger reductions of the cohort as a whole (by preventing the increases of some cohort members). Schultz et al. found that adding smiley faces to the door hangers for low baseline users mitigated the boomerang effect [10]: Pairing the peer comparison with injunctive norms – i.e., norms "that characterize the perception of what most people approve or disapprove" [39] – modified the peer comparison's effect. In contrast, Allcott [11], in a study with tens of thousands of participants, found that such injunctive elements were an insignificant factor in prompting low consumers not to increase their usage.

1.1. Objective and differentiation of current study

We introduce a scalable approach to generating feedback as a means of prompting residential electricity usage reductions and measuring its effectiveness, aimed at enabling large scale applications of eco-feedback and optimizing the content of each feedback message so it can be as effective as possible in prompting usage reductions. Borrowing techniques from Natural Language Processing, messages are generated automatically, using 10 message features in randomly varied combinations to create a multitude of message types. As a novelty vis-à-vis previous approaches, residents are not pre-partitioned into subsets that would henceforth receive the same message type and whose electricity reduction would then be compared (e.g., compare the subset of residents that always received information about their neighbors' usage with the subset that never received such information). Instead, each resident receives different feedback types during the course of the experiment. This approach provides the ability to analyze not only which feedback types are more effective than others for the average resident (as studied previously), but also how a feedback's effect on the same resident differed depending on (i) the resident's most recent usage; and (ii) the variability of messages from one feedback to the next (not studied previously). We applied the approach in a proof-of-concept case study, which followed the electricity usage of 36 residents who each received 14 different message types over the course ~2 months and compared their usage to that of a control group of 89 residents during the same time period (504 observations). We discuss potential advantages of our approach and how it might be used in large scale field experiments that could be carried out in collaboration with utilities.

2. Methods

We first describe the experimental setup, its rationale, and how observed changes in electricity usage were adjusted for (i) exogenous effects (such as weather, weekday/weekend, and holidays); and (ii) idiosyncratic effects such as personal differences in lifestyle and consumption behavior that were present already during the baseline, prior to receiving feedback (such as, e.g., predominantly cooking at home versus dining out). The following sections provide an overview of the novel feedback scheme, details on how the feedbacks were generated and delivered, and how the electricity data were analyzed in order to determine the efficacy of different feedback features and how these varied depending on a resident's recent usage prior to each feedback.

In all reported results and figures, error intervals illustrate ±1 standard error of the mean (SEM [40]) of the sample average. Differences between averages are tested for statistical significance via two-tailed student t-tests (unequal variances) at the reported p-value [40]. Because of the varying sample sizes across the different analyses, we tested for three levels of statistical significance: weakly significant ($p < 0.10$), significant ($p < 0.05$), and strongly significant ($p < 0.01$).

2.1. Overview and rationale of experimental design

The experiment was carried out in 2017 in a New York City residential rental building that comprises predominantly single-occupied studio apartments of similar size. The electricity usage, metered for each apartment separately, comprises air conditioning (controlled by individual residents via their own thermostat), lights, as well as plug-loads such as the refrigerator, electric oven, microwave, entertainment system, computers and electronics, and router/WiFi equipment. Apartments do not feature their own washer/dryer. Rather, these are located centrally and thus do not run on a resident's own electricity meter (hence not captured in our data and excluded from this study). After obtaining approval from Columbia University's Institutional Review Board, researchers approached residents in the building lobby about receiving feedback on their electricity consumption. Residents were told that the electricity data from previously installed smart meters was available to be shared with residents, via emails. Any further explanation was kept to a minimum, in an effort not to bias participants.

44% of residents gave their consent to participate in the study, by sharing their first name, preferred email address, and apartment number (so we could identify the correct meter). About 1 in 20 of these participants later unsubscribed while the study was ongoing, putting the final voluntary participation rate at 42%. The apartments were divided into 4 groups: (i) Studios, whose residents received the randomized, multi-featured feedbacks (n = 36; henceforth "treatment group"); (ii) studios whose residents received other feedbacks as part of a different study (n = 34, not reported here); (iii) studios whose residents had not volunteered to participate and which were hence used as the control group (n = 91, including 2 outliers (below)); and (iv) all remaining apartments, which were excluded either because of their atypical size/layout (1- or 2-bedroom apartments and studios located in the souterrain; n = 45) or because their residents later unsubscribed (n = 5).

In order to determine how electricity usage changed as a result of the feedback (as opposed to above exogenous and idiosyncratic effects), data analyses observed the following principles (see equations in Tables 2–4):

(1) *Control group adjustment:* As in previous work [25], any change in electricity usage in an apartment whose resident received feedback was measured relative to the average usage in the control group during the same time period. This was necessary because both the treatment group and the control group used more electricity during the feedback period than during the baseline period, most likely due to increased air conditioning in warmer weather (see section 3.1). Similarly, during the feedback period itself, usage from one feedback to the next may have further varied naturally as a function of weekday/weekend or holidays. The control group adjustment thus separated the effect of receiving feedback on the one hand from the effects of said exogenous factors on the other (assuming that, on average, the exogenous factors would have affected the treatment group and the control group equally if no feedback had been sent).

(2) *Baseline adjustment:* This corrects for any usage difference between the treatment and the control group that may have pre-dated the first feedback and would thus have to be netted out from the effects attributed to feedback. In this study, when measured for the treatment group on average, such baseline difference happened to be negligible (0.04%, not statistically different from zero; see section 3.1), meaning that low electricity consumers were equally likely to opt in to receiving feedback as high consumers. But the baseline adjustment was still needed when analyzing the feedback effect on individual residents whose electricity usage may have started out significantly above or below control even before the first feedback, because of their personal consumption behavior.

Feedbacks were emailed to only one resident per apartment. For simplicity of language, throughout our analyses we refer to the electricity usage in a specific apartment (and any changes in that usage) as the "resident's usage" and the "resident's reduction/increase", even though, on occasion, someone other than the designated recipient of the feedback may certainly have contributed to the electricity usage in the respective apartment.

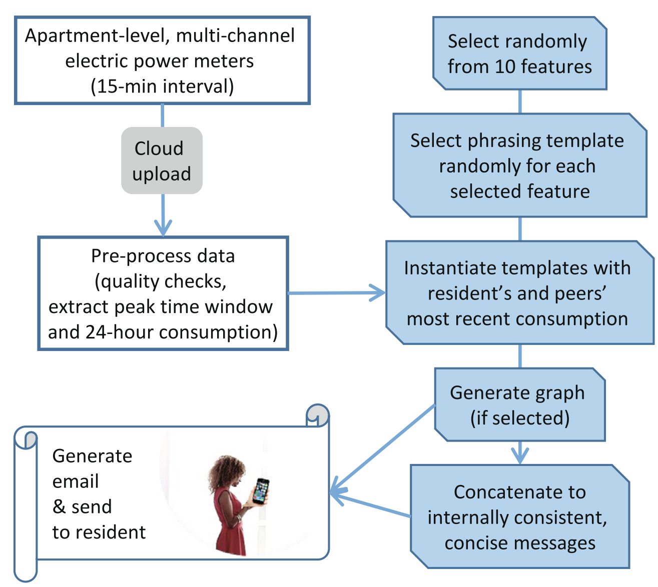

Fig. 1. Overview of end-to-end system, from electricity metering at apartment-level to emailing feedbacks. Feedback features and their phrasing were selected via random number generation for each of the 14 feedbacks to the 36 residents (504 feedbacks). Feedbacks varied not only between residents but the same resident received different features from one feedback to the next, thus providing a detailed dataset to investigate which feedback features were the most effective in prompting a change in electricity usage – and whether this feedback efficacy differed, depending on when and to whom the feedback was sent.

Electricity usage was metered and feedback messages were generated and emailed to residents automatically. Fig. 1 shows an overview of the system, covering all steps from metering the electricity usage to sending the emailed feedback messages. The system's individual steps are explained below.

2.2. Metering and processing of electricity usage

Apartment-level multi-channel electric power meters with scan transponders (*Quadlogic Controls Corporation®*) had previously been installed in the building. Every 15 minutes, an automated cloud-based driver collected and stored data from these meters into a *Microsoft Azure®* server using the *Hypertext Transfer Protocol* (HTTP). Data was quality-checked to ensure that (i) no special characters were present, which would indicate a transfer error; and (ii) kWh readings increased monotonously over each 24 h (meters recorded cumulative kWh and reset at midnight). After determining the time-window of peak usage for each apartment (see below), only the total usage in each apartment for each day (0–24 h local time) was retained for subsequent processing.

2.3. Overview of feedback scheme

In the approach developed for this study, residents in the "treatment" group (i.e., those who received feedback) received automatically generated, multi-featured, randomized feedbacks via email approximately twice a week, at 10 am, on varying days of the week. The design of each message varied in ten dimensions, henceforth referred to as "features". Features constituted, for example, whether the feedback included a graph or was text-only, in what equivalent consumption unit the usage was contextualized (e.g., "CO_2 emissions" or "trees cut down"), whether the recipient's usage was compared to his/her own previous consumption or that of peers, and in the sentiment of the text (e.g., "Great job!" for a positive and "Unfortunately, ..." for a negative sentiment). Crucially, for every feedback message individually, the presence/absence of each of the features was determined randomly. Accordingly, the features varied not only between residents, but also, for the same resident, from one feedback to the next. The same resident therefore observed a diversity of feedbacks over time, not only because the data (i.e., their own electricity usage and/or that of their peers) changed from one feedback to the next, but also because the feedback features that communicated said data changed. An overview of the 10 possible features and an example feedback are shown in Fig. 2. Each of the 10 features and their combination into concise messages are explained in the next section.

2.4. List of 10 feedback features and their possible expressions

The expressions for eight of the ten features were binary (true/false), simply denoting the presence or absence of a piece of information in the feedback email. These binary features were: (1) whether a graph was provided; (2) whether the resident's own usage was compared to the peer average in the treatment group (see *Results* for definition of peers); (3) whether the resident's own usage prior to the current feedback was compared to his/her usage in the previous feedback; (4) whether a projected monthly electricity cost was provided (calculated at USD 0.20/kWh, based on analysis of electricity invoices from Consolidated Edison, New York); (5) whether the resident's cost was compared to the peer-average in the treatment group; (6) whether the cost projected from the current feedback was compared to that projected from his/her previous feedback; (7) whether a link to and summary of a relevant news article was provided (details, see below); and finally (8) whether the participant was told when their peak usage time had occurred ("day", if 9 am to 6 pm; or "night", otherwise). Including information about their peak usage time may encourage residents to load-shift some of their electricity usage, which is typically higher during the day time [41], to the night time, with the underlying aim to alleviate stress on the grid (as an alternative to other load profile optimization schemes such as market-based mechanisms [42], wide-spread availability of building-level batteries [43], connected vehicle storage [44], grid level storage [45], central optimization [46] or apartment-redistribution of air conditioning loads [47], or more efficient electric consumer products [3,48]).

In contrast to the above eight binary features, the two remaining features, sentiment and equivalent consumption unit, determined the overall framing of the message and were always present, but in one of multiple different expressions: (9) Sentiment referred to the expressed attitude of the text, with three expres-

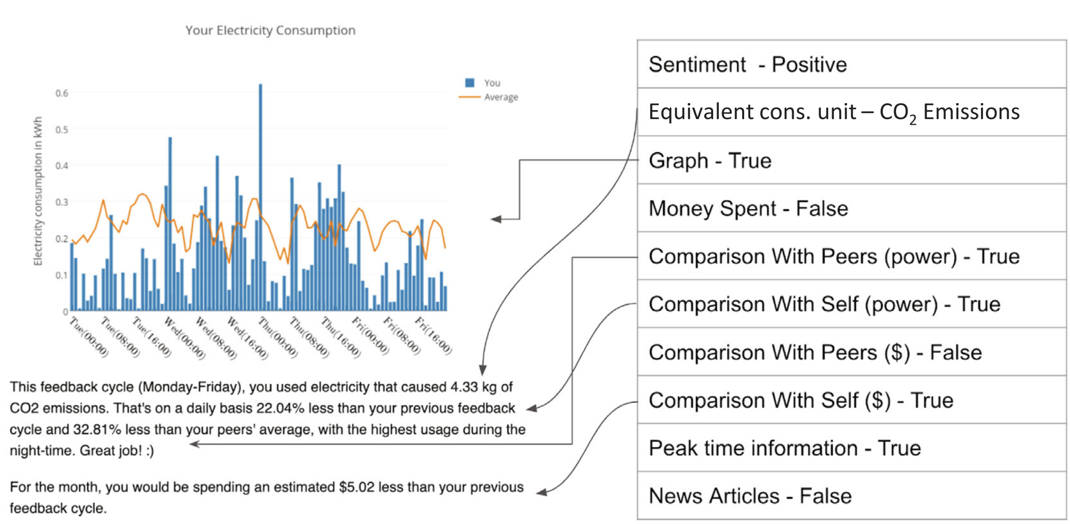

Fig. 2. Overview of 10 message features (right) and their specific expressions and phrasing in one example (left). Eight features were binary (i.e. "true" if present, "false" otherwise). Two features, sentiment and equivalent consumption unit were always present, but in one of multiple expressions (1 of 3 different sentiments; 1 of 5 different consumption units). In the example message, 7 of the possible 10 features are present.

sions: positive (using words like "congratulations" and "good job"), negative ("unfortunately", "worse"), or neutral (neither positive nor negative). (10) The equivalent consumption unit determined the context in which to communicate the amount of electricity a resident had used during the days since the most recent feedback. One of five expressions was possible: kilowatt-hours, equivalent trees cut down annually (a unit shown to be successful in previous work [25]), CO_2, greenhouse gases, or equivalent miles driven in a 20 mpg vehicle. Conversion factors from kWh to each of these equivalent consumption units were obtained based on US EPA guidance [49] for the trees and miles driven metrics, and from New York City's greenhouse gas inventory [50] for the CO_2 and greenhouse gas metrics (Table 1).

2.5. Feedback generation

Template-based systems are a common way of generating grammatically correct textual sentences from raw data [51]. A flexible, template-based natural language generation system was designed to generate textual feedbacks for this study. Several possible templates were defined for each feature. While each feature template was manually specified, the system was able to automatically combine templates for all possible configurations. For example, the feedback message in Fig. 2 included seven features. But other messages included as few as two features or as many as ten. Any combination of features was possible, and the system determined the appropriate phrase template for each combination. Sometimes the message was generated as one sentence per feature, but more often, when multiple features were included, a single sentence was generated that comprised two or more features. Given this flexibility, our approach yielded a total of 55,560 possible different message types, which were then instantiated with the respective values (e.g., the exact amount of electricity used by the particular resident in the 3–5 days prior to the feedback). This is a dramatic increase in message variety over an approach that would have used a single, manually crafted sentence per each feature, and the resulting messages tended to be much more fluent.

For example, if the consumption was expressed in the metric of equivalent trees cut down, the template sentence conveying that information was:

> "This feedback cycle ($duration), your electricity consumption rate corresponded to a yearly cut down of \$trees trees",

where \$duration and \$trees were replaced with the appropriate values. An example of a single sentence containing multiple features was:

> "Unfortunately, this is $power_prev_diff your consumption last week and $power_avg_diff similar apartments in your building, with the worst consumption in the $peak_time",

where the three parameters were replaced with the appropriate values. In this example, which featured the comparison with peers, the comparison with self, and the peak time info, and which used a negative sentiment, appropriate instantiations of \$power_prev_diff were adjusted to be "*1.3 kWh more than*" or "*the same amount of energy as*", depending on the other features chosen for the message. In case the $peak_time instantiation was "*day time*", the sentence was extended by the phrase "*..., a time when the grid is most stressed.*", in order to raise the issue of resiliency of energy availability, which has been shown to resonate with consumers [52].

Table 1
Equivalent units of consumption and respective conversion factors used in feedback messages.

Consumption unit	Conversion factor
Plain kWh	n/a
Trees cut down yearly	0.273 trees per year/kWh
CO_2 emissions	0.293 kg/kWh
Miles driven in a 20 mpg vehicle	1.009 miles/kWh
Greenhouse gas emissions	0.293 kg/kWh

The final feedback text was created by concatenating the templated and instantiated sentences for the specific combination of features. The two features sentiment and equivalent consumption unit modified all other templates – each with a variation of the three sentiments and the five consumption units (see above). Furthermore, if the randomly selected sentiment conflicted with a resident's actual usage in a particular feedback, thus resulting in an inconsistent message (e.g., the resident used less electricity than previously but the randomly selected sentiment was negative), the sentiment feature for that particular feedback message was changed to neutral.

As the final feature, a summary of a news article along with the respective URL could be added at the end of the feedback, as follows: A story selected via standard search engines using predefined, energy and environment-related keywords (such as, e.g., "power generation", "climate change") was summarized into 1–2 sentences. The summarizer used a recently developed approach based on convolutional neural networks that selected salient news article sentences such that all important topics in the article were covered [53]. An example of such a summary, which was included in a feedback was:

We think this news article may interest you: Temperatures 30 to 50 degrees Fahrenheit above normal hit parts of the region this past winter. Several years ago New York City-based photographer Diane Tuft decided she had to visit the Arctic to see the vanishing polar environment for herself. Read more at https://www.scientificamerican.com/article/the-ominous-beauty-of-the-arctic-meltdown/.

Feedbacks that included this feature were on average less effective in prompting reduction. The average $R_{t,f}$ was 8% (n = 248) for feedbacks with a news story vs. 14% (n = 256) without (not shown in *Results*). While not statistically significant, this result could indicate that the news story simply made the message longer rather than adding impactful information, or that the topics and keywords based on which the news stories were selected were less relevant to the recipients than we would have liked.

2.6. Feedback delivery

Our laboratory's local SMTP server automatically delivered each feedback to the resident's preferred email address. Graphs, if present, were embedded in the html at the beginning of the email via the *Python®* package *Plotly*, along with an "unsubscribe" button at the end. Over the course of the experiment, 504 emails were sent (14 emails to each of the 36 residents in the treatment group).

2.7. Removing outlier apartments

To remove statistical outliers, including unoccupied apartments, we first analyzed the distribution of the time-averaged electricity usage over the course of the experiment ($\bar{P}$) across all treatment and control apartments. As expected from zero-bound metrics, the distribution was asymmetric, akin to a lognormal distribution. In our study the minimum load was 6 Watt, geometric mean 168 Watt, arithmetic mean 192 Watt, maximum 634 Watt, and the skew was 1.5 (n = 127 apartments). Because of the lognormal shape, we then transformed each load $\overline{P}$ to $ln(\overline{P})$ and subsequently treated as outliers any apartments whose $ln(\overline{P})$ was outside a ± 2 standard deviation tolerance around the average of all $ln(\overline{P})$ [40]. Two of the 127 apartments were thus removed, the one with 634 Watt time-average load and the one with 6 Watt.

Table 2
Glossary and equations for $R_{date,(net)}$. Metrics denoted with *R* are sign-inverted: A positive metric means that residents reduced their usage relative to the control group (i.e., the effect of the feedback was as intended).

Metric	Explanation, use, and numerical example	Symbol [sample size]	Equation
Average treatment vs. control on specific date (any date in baseline or feedback period)	• Monitor daily variation in the relative usage in the treatment versus the control group • Not yet baseline-adjusted • Expl: If for a certain 24 h period the average apartment load in the treatment group is 380 W versus 400 Watt in the control group, then R_{date} = +5%	R_{date} [n = 65; 52 for the feedback period and 13 for the baseline period]	$1-\frac{1/36\sum_{r=1}^{36}\overline{P_{r,date}}}{1/89\sum_{c=1}^{89}\overline{P_{c,date}}}$ Where: (1) • $\overline{P_{r[c],date}}$ [Watt] is the time-average load of resident r [c] in the treatment [control] group during the entire day (0–24 h local time) of the specified date
Baseline adjustment for resident *r* in treatment group	• Used to isolate the effect of the feedback from pre-existing differences in electricity usage for resident *r* that were present during the baseline already • Expl.: If resident *r*'s time-average load during the baseline is 150 Watt but the average of this measure in the control group is 200 Watt, then this metric for resident *r* is 0.75	BA_r	$\frac{\overline{P_{r,basel.}}}{1/89\sum_{c=1}^{89}\overline{P_{c,basel.}}}$ Where: (2) • $\overline{P_{r[c],basel.}}$ as in Eq. (1), however measured over the entire baseline period
Baseline adjustment for entire treatment group	• Used to isolate effects of feedback from pre-existing differences in usage for the treatment group that were present during the baseline already and therefore not a result of the feedback • In this experiment, the baseline time-average load of treated residents was 132.88 Watt whereas it was 132.93 Watt for the control group. This metric is therefore 0.9996	BA_r	$\frac{1/36\sum_{r=1}^{36}\overline{P_{r,basel.}}}{1/89\sum_{c=1}^{89}\overline{P_{c,basel.}}}$ (3) Where: • $\overline{P_{r[c],basel.}}$ as in Eq. (2)
Average treatment vs. control on specific date (during feedback period only)	• Monitor downward trend in the daily difference between treatment and control group in the 3–5 days following each feedback ("response-relapse") • Baseline-adjusted • Expl.: If for a certain 24 h period the average treated resident's load is 175 W versus 184 W in the control group (i.e., 0.9500), and since that ratio during the baseline period is 0.9996, then $R_{date,net}$ = +4.96% (here, W_{date} = 1) • The weighting (via W_{date}) provides the advantage that the average of all 52 $R_{date,net}$ equals the average of all 36 $R_{t,all\ feedbacks}$, thus facilitating interpretation of results across all metrics, as shown in Eq. (11)	$R_{date,net}$ [n = 52 for the feedback period]	$W_{date}\cdot\left[BA_r+R_{date}-1\right]$ Where: (4) • BA_r as in Eq. (3) • W_{date} is the consumption-weight for that date, used to adjust for variations in electricity usage in the control group from one day to the next: Wdate=Pctrl., date/Pctrl., datewith $P_{ctrl.,date}$ [Watt] denoting the time-average load per resident in the control group on the specified date (0–24 h local time), and $P_{ctrl.,date}$ denoting the average $P_{ctrl.,date}$ across all 52 days of the feedback period

The latter apartment was most likely unoccupied, judged from the small electricity load. Both outlier apartments were in the control group. This left 89 control apartments ($\overline{P_c}$ with range 54–508 Watt) and 36 treatment apartments ($\overline{P_r}$ with range 61–400 Watt), which were used in all subsequent analyses.

2.8. Reduction metrics

We defined four metrics, designed to capture different aspects of the efficacy of feedbacks to prompt reductions in electricity usage. All four metrics measure the change in electricity usage in the treatment group as a fraction of the average usage in the control group during the same time intervals (control group adjustment; see Section 2.1). The metrics further correct for any differences in electricity usage between residents in the treatment group and the control group that pre-existed already during the baseline period (baseline adjustment; see Section 2.1). The four metrics reflect different cohorts (from a single resident to an average across all 36 residents) and different time intervals (a specific day in the feedback period, any set of 3–5 days that followed a specific feedback round, or all 52 days in the entire feedback period), as follows:

$R_{date,(net)}$: Quantifies the average electricity usage reduction achieved by residents in the treatment group on a specific day of the 52 day feedback period. This metric was used to track the reductions achieved by the average resident over time (Fig. 3) as well as the average resident's response-relapse in-between feedbacks (Fig. 4).

$R_{r,f}$: Quantifies the electricity usage reduction that a specific resident *r* achieved in response to a specific feedback *f*. This metric was used for two purposes: (i) To determine which specific feedback types were more effective than others (Fig. 5); and (ii) to determine whether the efficacy of the feedbacks varied depending on which resident received them and when (Figs. 6–9).

$R_{r,all\ f}$: Quantifies the electricity usage reduction achieved by a specific resident *r*, averaged across the entire feedback period. This metric was used to determine whether some residents exhibited more pronounced reductions than others (sometimes referred to as heterogeneity [5,37]).

$\Delta p_{r,f}$: Quantifies the *change* in electricity usage of a specific apartment from before to after receiving a specific feedback. This metric was used to investigate mean reversion in residents' usage.

Equations for the metrics and numerical examples to illustrate their interpretation are described in Tables 2–4. We further point to an important mathematical identity: Whether the feedback effect is analyzed by day (n = 52 observations), by resident for all feedbacks combined (n = 36 observations), or by resident and by feedback (n = 504 observations), the averages of these metrics are identical to the overall achieved reduction of 11.1%. This identity is used in order to determine whether there were certain subsets of days, residents, feedback types, or pairings of residents and feedback types whose usage reductions were higher or lower than the overall average of 11.1%. Note that the average of the 504 $\Delta p_{r,f}$ is 0.0% and that $\Delta p_{r,f}$ (as opposed to the various *R* metrics) is not weighted. Therefore, results in Figs. 8 and 9 (which focus on short

Table 3
Glossary and equations for $R_{r,f}$, $R_{r,all\ f}$, and $r_{r,all\ f}$.

Metric	Explanation, use, and numerical example	Symbol [sample size]	Equation	
Effect of treatment of resident r, in response to feedback f	• Investigate how different residents react to specific feedbacks with different features • Baseline-adjusted • Expl.: If resident r consumes 5% below control <u>during the baseline</u> but 10% above control in the days following, e.g., feedback #12 (and if W_{12} = 1), this metric is −15% (i.e., increase) for f = 12 • The weighting (via W_f) provides the advantage that the average across all 504 $R_{r,f}$ equals the average across all 36 $R_{r,all\ feedbacks}$, thus facilitating interpretation of results across all metrics, as shown in Eq. (11)	$R_{r,f}$ [n = 504, 36 apts. receiving 14 feedbacks each]	$W_f \cdot \left[BA_r - \frac{\overline{P_{r,f}}}{1/89\sum_{c=1}^{89}\overline{P_{c,f}}}\right]$Where: • BA_r as in eq (2) • $\overline{P_{r[c],f}}$ as in Eq. (1), however measured over the 3–5 days <u>following</u> feedback f (until the next feedback was sent) • W_f is the consumption weight for feedback f, used to adjust for variations in electricity usage in the control group from one feedback to the next: Wf=Pctrl., f/Pctrl., fwith $P_{ctrl.,f}$ [Watt] denoting the time-average electric load per apartment in the control group during the 3–5 days following feedback f, and $P_{ctrl.,f}$ denoting the average $P_{ctrl.,f}$ across all 14 feedback rounds	(5)
Effect of treatment on resident r, averaged across all 14 feedback rounds	• Investigate heterogeneity amongst residents re. how they respond to feedbacks overall • Baseline-adjusted • Expl.: If resident r consumes 20% above control during the baseline period but 5% below control during the feedback period, this metric is +25%	$R_{r,all\ f}$ [n = 36]	$1/14\sum_{f=1}^{14} R_{t,f}$Where: • $R_{r,f}$ as in Eq. (5)	(6)
Effect of treatment on resident r, averaged across all 14 feedback rounds	• Investigate heterogeneity amongst residents re. how they respond to feedbacks overall • Baseline-adjusted • Expressed as fraction of a resident's own consumption (as opposed to the control group's average), hence determines each resident's net change relative to his/her <u>individual</u> baseline use Expl.: If resident r consumes 50% below control during the baseline period and 60% below control during the feedback period, this metric is +20%	$r_{t,all\ f}$ [n = 36]	$1 - \left[\frac{\overline{P_{r,feedb.}}}{1/89\sum_{c=1}^{89}\overline{P_{c,feedb.}}}/BA_t\right]$Where: • $\overline{P_{r[c],feedb.}}$ as in in Eq. (1), however measured over the entire feedback period BA_r as in Eq. (2)	(7)

Table 4
Glossary and equations for $p_{r,f}$, $p^*_{r,f}$, and $\Delta p_{r,f}$.

Metric	Explanation, use, and numerical example	Symbol [sample size]	Equation	
Usage of resident r (relative to control) during days <u>ahead of</u> feedback f	• Investigate mean reversion behaviour • Expl.: If resident r consumes 10% above control during the days <u>ahead of</u> feedback e.g. #12 (i.e., between feedback *#11* and *#12*), this metric is +10% f = 12	$p_{r,f}$ [n = 540, 36 apts.; f = 1…15	$\frac{\overline{P_{r,f-1}}}{1/80\sum_{c=1}^{89}\overline{P_{c,f-1}}}$Where: • $\overline{P_{r[c],f-1}}$ as in Eq. (5) • When used in Eq. (10), $p_{r,1}$ uses the usage P during the baseline, and $p_{r,f=15}$ the usage during the 5 days after the last feedback (#14). This yields 14 $\underline{\Delta} p_{r,f}$ for each of the 36 residents	(8)
Usage of resident r (relative to other treated apartments) during days <u>ahead of</u> feedback f	• Investigate mean reversion behaviour • Expl.: If resident r consumes 10% above the average of all treated apartments during the days <u>ahead of</u> feedback e.g. #12 (i.e., between feedback *#11* and *#12*), this metric is +10% f = 12 • p^* was chosen differently from p, in order to be the same as the actual peer comparison metric communicated to residents in feedbacks; the correlation between $p_{r,f}$ and $p^*_{r,f}$ is 0.97	$p^*_{r,f}$ [n = 504, 36 apts. receiving 14 feedbacks each]	$\frac{\overline{P_{r,f-1}}}{1/70\sum_{r=1}^{70}\overline{P_{r,f-1}}}$Where: • $\overline{P_{r,f-1}}$ as in Eq. (5) • $p^*_{r,f=1}$ uses the consumption P during the 5 days prior to feedback #1 (i.e., the period reported on in feedback #1)	(9)
Change in usage of resident r (relative to control) from <u>prior to after</u> receiving feedback f	• Investigate mean reversion behaviour • Expl.: If resident r consumes 5% below control during the days following feedback e.g. #11 but 10% above control in the days following feedback #12, this metric is +15% change in consumption for f = 12 (i.e., increase relative to control)	$\Delta p_{r,f}$ [n = 504, 36 apts. receiving 14 feedbacks each]	$p_{r+1,f} - p_{r,f}$Where: • $\Delta p_{r,f}$ as in Eq. (8)	(10)

term behavioral aspects from one feedback to the next) are not directly comparable to those in Figs. 3–6 (which focus on net reductions versus the baseline usage).

The 3 metrics are related as follows, using the same symbols as in Eqs. (1)–(6):

$$\frac{1/36\sum_{r=1}^{36}\overline{P_{r,basel.}}}{1/89\sum_{c=1}^{89}\overline{P_{c,basel.}}} - \frac{1/36\sum_{r=1}^{36}\overline{P_{r,feedb.}}}{1/89\sum_{c=1}^{89}\overline{P_{c,feedb.}}} = 11.1\%$$
$$= \frac{\sum_{date=1}^{52} R_{date,net}}{52} = \frac{\sum_{r=1}^{36} R_{r,allf}}{36} = \frac{\sum_{f=1}^{14}\sum_{r=1}^{36} R_{r,f}}{504} \quad (11)$$

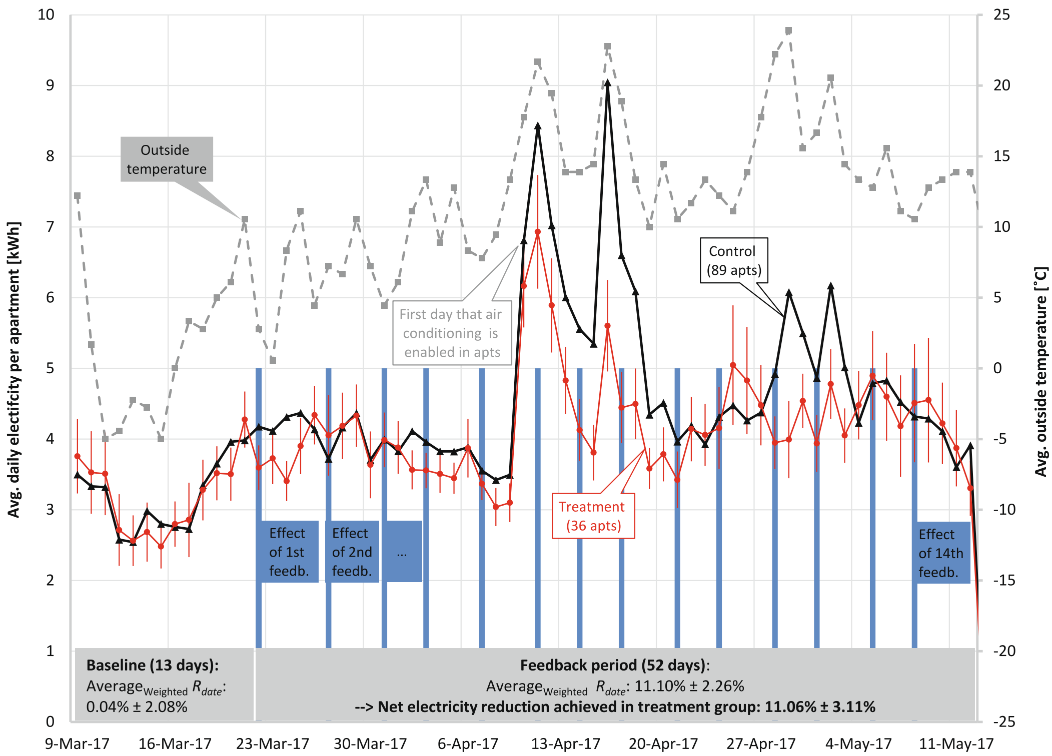

Fig. 3. Overview of experimental design and determination of electricity reduction (11.1 ± 3.1%) of the 36 residents that received feedbacks. Averages of daily reductions (R_{date}; see *Methods*) are weighted by the control group's usage the same day. Blue vertical bars indicate the date on which each of the 14 feedbacks were emailed to residents in the treatment group (at 10am local time). Black triangles and red circles show the average apartment-level electricity usage each day for the control and the treatment group, respectively. Error bars for the treatment group (n = 36) show ± 1 SEM. SEM for the control group (n = 89) are smaller but not shown, to preserve clarity of the graph. Average outside dry-bulb temperature (grey squares, right axis) is shown to infer approximate air conditioning loads (www.ncdc.noaa.gov; New York City Central Park station). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3. Results

3.1. Total achieved reductions

Fig. 3 illustrates the experiment, its duration, and the purpose of monitoring the electricity usage of a control group. During the baseline period (days 1 to 13), the apartments' daily electricity usage was monitored but no feedbacks were sent yet. On day 14 of the experiment (at 10 am), the first feedback was sent to the 36 participants. This was followed by 13 more feedbacks, always at 10 am, each containing information about the respective resident's electricity usage in the 3–5 days prior to the feedback. The experiment ended on the 4th day after sending the 14th feedback. After that, a large portion of the residents vacated their apartments because of annual lease agreement conditions, thus concluding the study.

During the baseline period, the average daily electricity usage in the treatment group (n = 36) was 3.189 kWh per apartment, statistically indistinguishable from that in the control group (3.190 kWh; n = 89). During the 52-day feedback period, usage in both groups increased, largely driven by air conditioning in response to rising outside temperatures (also shown in Fig. 3). In order to separate these exogenous effects from the effect of the feedback, we analyzed the treatment group's usage *relative* to that of the control group during the same period. We then counted as attributable to feedback only those changes in the treatment group that deviated from concurrent changes in the control group (see *Methods*).

During the feedback period, the average daily usage in the treatment group was 4.195 kWh per apartment, 0.524 kWh lower than that in the control group (4.719 kWh). As seen from the error bars

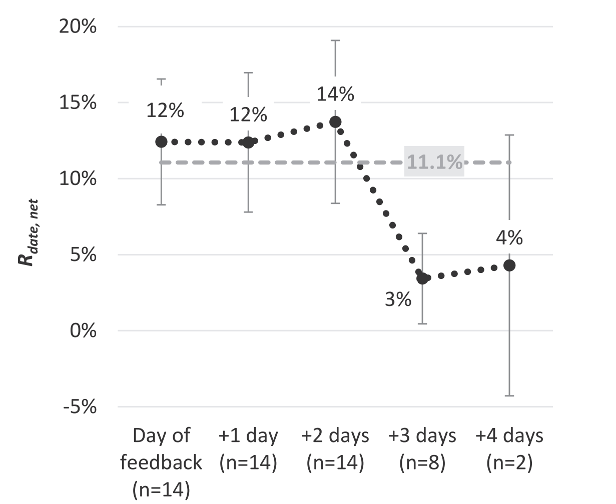

Fig. 4. Response-relapse. Electricity usage reductions on specific days ($R_{date,net}$; 0–24 h local time; see *Methods*) averaged across all 36 residents in the treatment group as a function of the time passed since the most recent feedback (feedbacks were sent at 10 a.m.). Error bars show ± 1 SEM, accounting for the varying sample size. The reduction for the "+3 days" group is statistically significantly lower than for the 3 earlier days combined ($p < 0.05$). Grey dashed line shows average $R_{date,net}$ across all 52 days of the feedback period.

in Fig. 3, variation between residents added substantial noise to the daily average usage in both groups. But when averaged over all days, the experiment shows a statistically significant net reduction

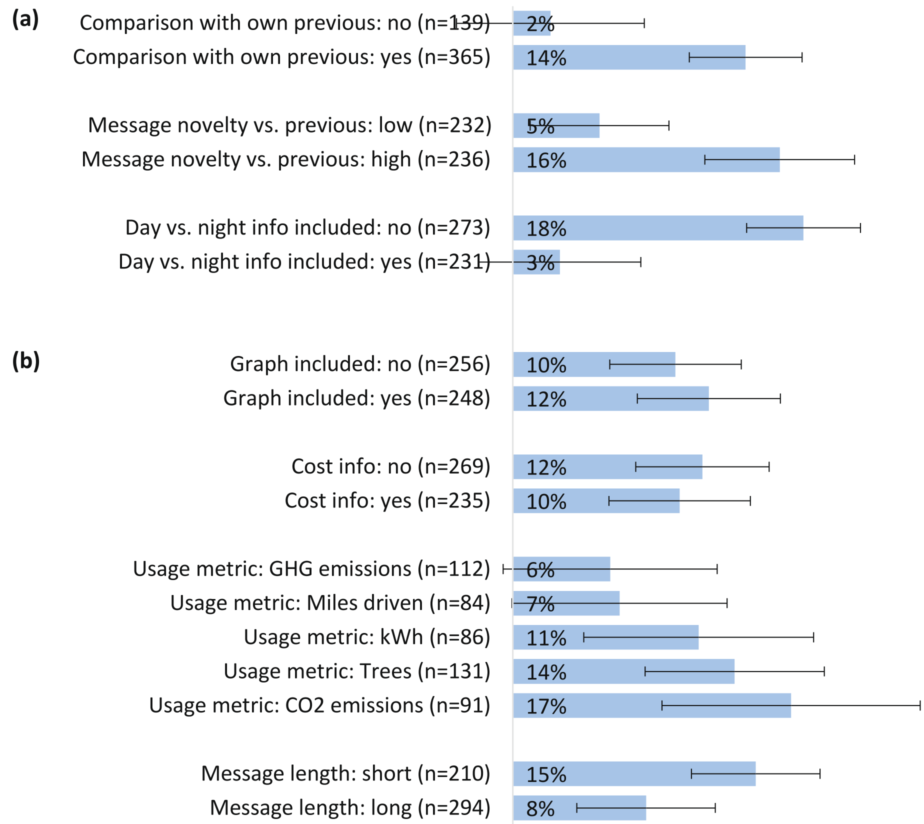

Fig. 5. Efficacy of feedback types. Average electricity usage reductions as % of the control group ($R_{r,f}$; see *Methods*) as function of feedback type (average across all 504 $R_{r,f}$ is 11.1%). Error bars show ± 1 SEM, accounting for the varying sample size. Results are shown in two tiers: (a) Types that had a (weakly) statistically significant effect on the observed reductions (p < 0.10); (b) Types whose effect was directionally as expected based on previous literature, however not statistically significant given the sample size.

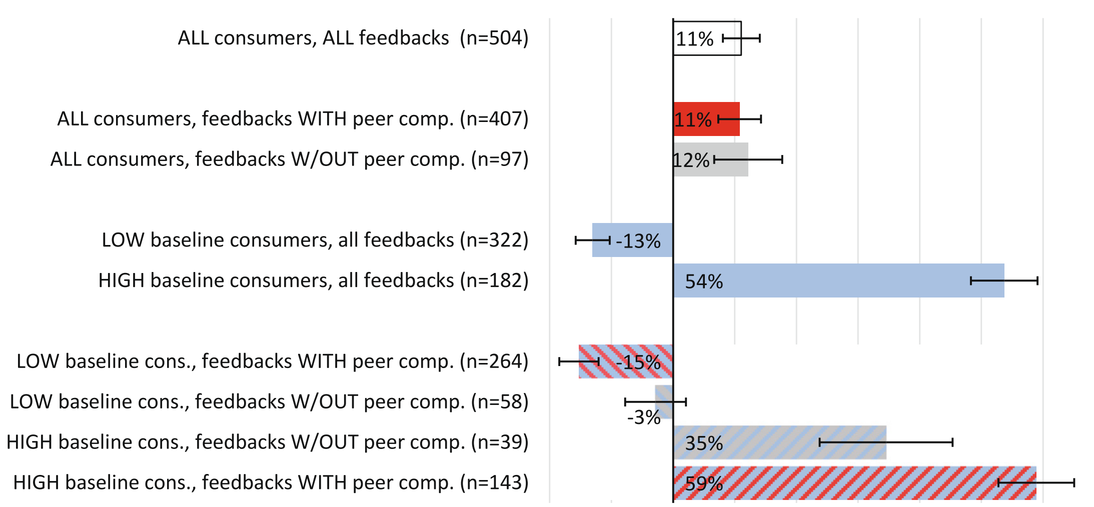

Fig. 6. "Boomerang effect" and role of peer comparisons. Average electricity usage reductions as % of the control group ($R_{r,f}$; see *Methods*) as function of (i) whether a peer comparison was included in the feedback; and (ii) whether feedback was sent to a resident with high [low] baseline consumption (i.e., above [below] average). Negative reductions denote usage increases. Error bars show ±1 SEM. Differences between "low" and "high" baseline consumers were statistically significant (p < 0.01).

between the baseline and the feedback period (p < 0.01). Interestingly, there is a correlation between R_{date} and the control group's use $P_{ctrl.,date}$ on the same date (ρ = 0.62, n = 52, p < 0.01). This correlation can be seen in Fig. 3: On days on which the electricity use in the control group was higher, e.g., on days with average outside temperatures above 20 °C, the relative difference between treatment and control group tended to be larger as well. In other words, on those days tenants in the treatment group were able to save a larger relative portion of their usage. We interpret this as indicating that the higher electricity usage on those days comprised a larger portion of discretionary use (e.g., the optional use of air conditioning in contrast to the non-optional electricity for the refrigerator), thus enabling a larger reduction relative to the total usage.

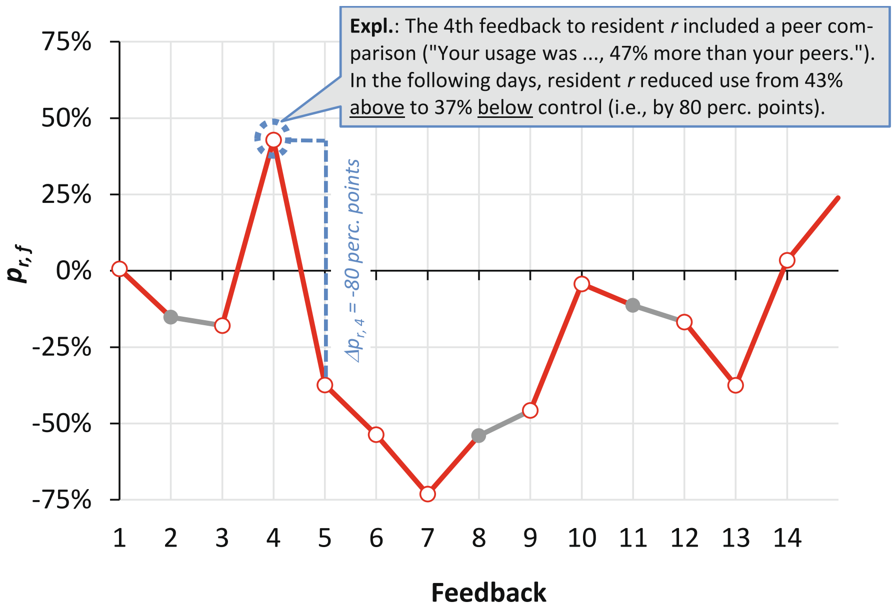

Fig. 7. Electricity usage of single resident over time (example). One particular resident's usage in the days prior to receiving a particular feedback, as deviation from control ($p_{r,f}$; see *Methods*). Open red circles indicate that the feedback included a peer comparison. Solid grey circles indicate other feedbacks. Changes in usage following a feedback with peer comparison are highlighted by red lines. As an example, the blue dashed circle illustrates the effect of one particular feedback on the resident. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In summary, the feedbacks prompted residents in the treatment group to reduce their usage, on average by 11.1 ± 3.1% relative to what residents in the control group consumed during the same time period (all error intervals are given as *sample average ± 1 standard error of the mean*, SEM [40]). Improvements in reductions above and beyond this general average are investigated in the following sections, and their possible applications are described in *Discussion*.

3.2. Response-relapse

Since feedback features varied randomly between residents, and since the same resident received different features from one feedback to the next, we can investigate which features yielded a more pronounced electricity usage reduction than others. However, this first raises the question as to whether a resident's reduction, if observed, e.g., during the days after receiving the 4th feedback, is caused mostly by this 4th feedback – or should the observed behavior be attributed to all 4 feedbacks received so far? In other words, at how many days (on average) after receiving a particular feedback did a resident's usage relapse back to his/her normal usage (referred to as in-treatment persistence, durability, or response-relapse [5,54]).

To quantify this, we analyzed the 52 daily reductions ($R_{date,net}$; see *Methods* for all metrics and respective equations), grouped based on how many days had passed since the most recent feedback was received. The average $R_{date,net}$ for each of the 5 groups is shown in Fig. 4, with "days passed" ranging from zero days (i.e., same day as the 10 a.m. feedback) to "+4 days". $R_{date,net}$ was statistically significantly larger than zero on the day of receiving the feedback (sent at 10am) and for another 2 days thereafter. However, on the following day ("+3 days"), the feedback effect is no longer statistically different from zero. The result for "+4 days" is inconclusive because only two of the 14 feedbacks were followed by 4 days without a new feedback, resulting in a small sample and thus large SEMs for the "+4 days" measurement. SEMs are too large to discern a more detailed pattern – such as an expected gradually declining usage reduction from one day to the next. Still, we interpret these results as indicating that the majority of any observed change in usage can be attributed to the most recent feedback only. This finding is consistent with a previous study, which also found a "response-relapse" from day + 3 after feedbacks [25]. Therefore, in all subsequent analyses, we consider the relationship of observed usage changes on one hand and the features of only the most recent feedback on the other, while disregarding the much smaller effect that previous feedbacks may have contributed.

In summary, the response relapse allows for the analysis of more complex factors contributing to feedback effectiveness such as message variety over time and a specific resident's most recent energy usage (see Sections 3.3, 3.5 and *Discussion*).

3.3. Efficacy of different feedback types to prompt reduction

We investigated the 504 observed reductions ($R_{r,f}$; see *Methods*), grouped by the type of each feedback. Feedback types were defined by which feature or combination of features was present in each feedback. Fig. 5 shows the 17 feedback types and the average $R_{r,f}$ prompted by each type. For example, the average reduction for the 365 feedbacks that included a self-comparison to a resident's own previous usage (whether in terms of the energy feature, the cost feature, or both) was analyzed separately from the average reduction of the 139 remaining feedbacks that did not have any of the two possible self-comparison features. Since the presence or absence of features was determined by a random number generator, the other feature expressions in the respective groups were the same on average. For example, about half of the 365 feedbacks of type "comparison with own previous: yes" included a graph, whereas the other half did not, and the same is true for the 139 feedbacks of type "comparison with own previous: no". Therefore, any observed difference in reductions prompted by two groups can be attributed to the particular type by which they were grouped.

Three message types had (weakly) statistically significant effects on the electricity usage reduction behavior (Fig. 5a). Feedbacks that included a self-comparison (i.e., "..., *x% higher [lower]*

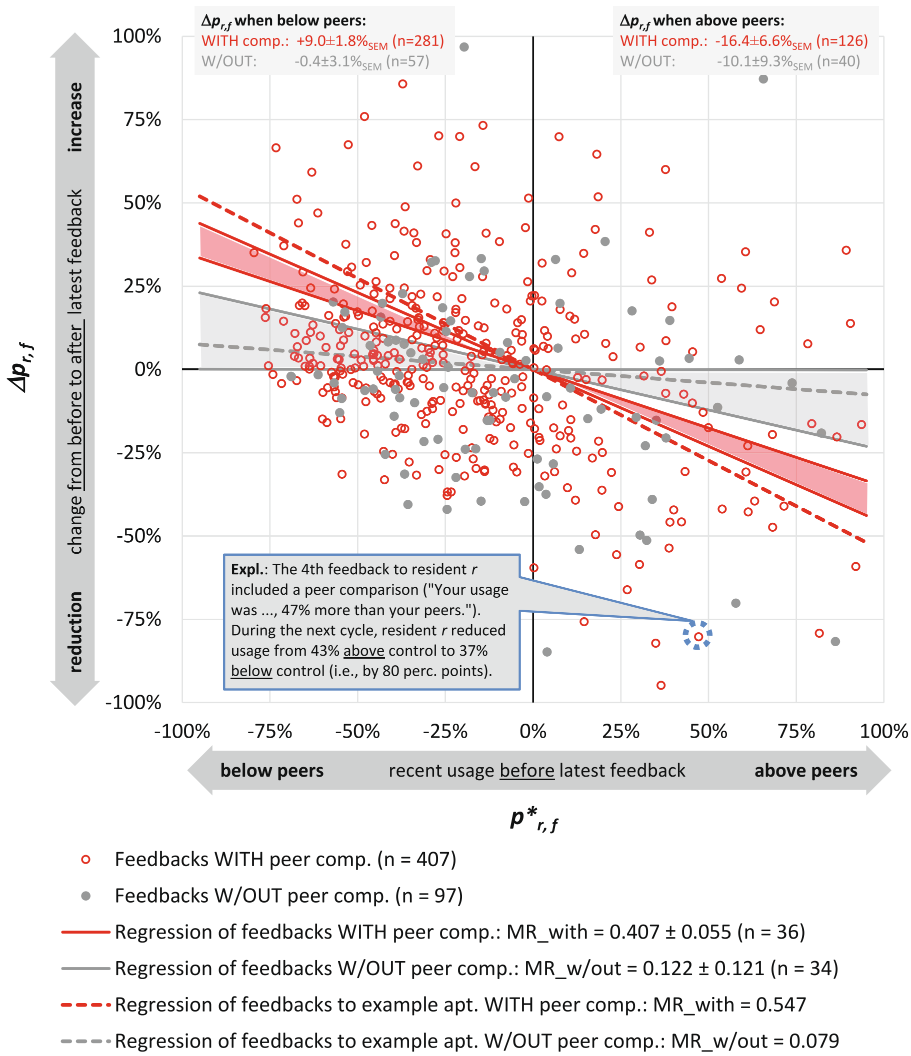

Fig. 8. Mean reversion of electricity usage and amplification by peer comparison. Usage change of resident r from before to after receiving feedback f ($\Delta p_{r,f}$) as function of resident r's recent usage relative to peers before fedback f ($p^*_{r,f}$; see *Methods*). Dashed regression lines show the mean reversion (MR) behavior for one example resident (see legend). Shaded areas between the solid regression lines show the ±1 SEM range of the mean reversion behavior of the average resident . The example resident and the specific observation highlighted by the blue dashed circle are the same as in Fig. 7. Note that ~ 5% of the 504 observations are outside the plotted range (but were included in the regression results and shown averages). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

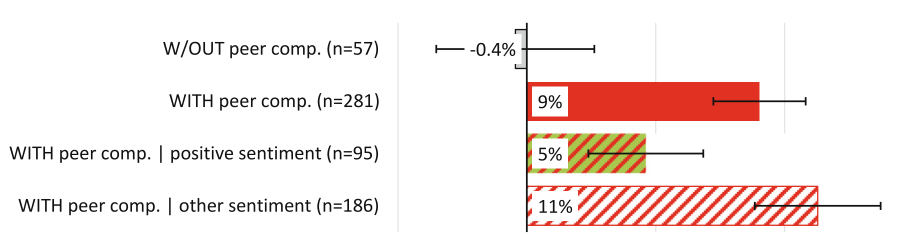

Fig. 9. Mitigation of boomerang effect in residents with recent electricity usage below peers. Observed average usage increases from one feedback to the next in % of the control group ($\Delta p_{r,f}$; see *Methods*) as function of (i) whether a peer comparison was included in the feedback; and (ii) whether the feedback's sentiment was positive or negative/neutral. Because of the focus on the boomerang effect, only feedbacks to residents are shown whose most recent usage was below that of peers (i.e., 338 of the 504 observations). Error bars show ± 1 SEM.

than in your previous period") had a strong average reduction effect (14%), whereas feedbacks that lacked a self-comparison had a smaller effect (2%, statistically indistinguishable from zero). Feedbacks of "high novelty" prompted ~3 times larger reductions than low-novelty feedbacks (16% vs. 5%). The "novelty" type (low or high) was defined as whether the number of features whose expression changed vis-a-vis the particular resident's previous feedback was lower or higher than average (determined for all feedbacks except the first). Feedbacks that added information at what time (day or night time; see *Methods*) a resident had used the most electricity were less effective than those without this information (see *Discussion*).

Other feedback types behaved directionally as expected based on past literature. However, the respective differences in prompted reductions were not large enough to be statistically significant (Fig. 5b). Including the graph (see example in Fig. 2) tended to be more effective than not including the graph. Including cost information – in our study the retail cost of electricity that residents had consumed, extrapolated to a full month – was less effective, possibly because residents in the building were not separately billed for electricity. With regards to in which equivalent metric the electricity usage was expressed, "greenhouse gases" prompted the lowest electricity usage reductions. "Number of trees cut down" was more effective than direct kWh, directionally in agreement with previous studies [25]. Interestingly, referring to climate-change relevant emissions simply as "CO_2" was more effective than "greenhouse gases" even though the amounts in kg were the same. This may indicate that the term "CO_2" is more immediately associated by residents as something to avoid than the term "greenhouse gases", but such nuances would have to be re-tested with larger sample sizes (see *Discussion*). Short feedbacks (defined here as those with a below average number of present features) tended to be more effective in prompting reduction than long feedbacks, suggesting a benefit for automatically generated feedback messages to be as concise as possible. The effect of peer-comparison and message sentiment are investigated in more detail in the following sections. Including a climate-related news story in the feedback did not show a statistically significant effect (see *Methods*).

In summary, of all features and message types, message novelty showed the most pronounced positive impact on message efficacy, pointing to strategies for large scale applications of eco-feedback (see *Discussion*).

3.4. "Boomerang effect" and role of peer comparisons

Fig. 6 again shows different feedback types along with their average $R_{r,f}$. However, in contrast to Fig. 5, Fig. 6 further differentiates the feedback groups depending on whether the feedback recipient was a "high [low] baseline consumer". For this analysis, in line with previous studies [10,37], we defined a "high [low] baseline consumer" as a resident having had higher [lower] electricity usage during the baseline than the average usage in the control group.

Feedbacks to high baseline consumers prompted a usage reduction of 54% ± 5.4% (n = 182). In contrast, feedbacks to low baseline consumers prompted an increase of 13% ± 2.8% (n = 322), a phenomenon referred to in similar studies as a boomerang effect (see *Discussion*). There was a statistically significant positive correlation between a resident's usage during the baseline period prior to any feedback messaging on the one hand and their reduction during the feedback period on the other (n = 36, $p < 0.01$). This was the case whether the reduction was measured relative to a resident's own baseline usage ($r_{r,all\ f}$, $\rho = 0.66$) or relative to the baseline usage of the control group ($R_{r,all\ f}$ $\rho = 0.81$). A possible explanation is that high baseline consumers may have had higher portions of discretionary usage than low baseline consumers.

We also analyzed the effect of peer comparisons. A feedback with a peer comparison included the phrase "*..., x% less [more] than similar apartments in your building*", with *x* replaced with the deviation of the resident's own recent usage from that of his/her peers' recent usage ($\left|p^*_{r,f}\right|$; where peers were all other residents who lived in similar size apartments in the building and received feedbacks). For example, when $p^*_{r,12}$ for a specific resident *r* was 15% during the days leading up to feedback *f* = 12, then feedback *f* = 12 to that resident, provided a peer comparison was chosen by the randomized feature scheme, contained the phrase "*..., 15% less than similar apartments in your building*".

Peer comparisons did not a have significant effect on the average electricity usage reduction of the study group (11% reduction for feedbacks with peer comparison versus 12% without). However, peer comparisons did affect the boomerang effect: Only those feedbacks to low baseline consumers that included a peer comparison (whether in terms of energy, cost or both) led to a statistically significant boomerang effect (15% increase, $p < 0.01$). The feedbacks to low baseline consumers without peer comparison led to a much weaker increase (3%, statistically indistinguishable from zero). For the high baseline consumers, the effect was reversed, i.e. feedbacks with peer comparison led to stronger usage reductions (59%) than those without (35%) ($p < 0.10$).

In summary, peer comparisons did have an effect on a resident's electricity usage, but this effect varied as a function of how much electricity the resident consumed during the baseline. For the cohort as a whole, the effect averaged out. This could inform details of large scale applications of eco-feedback (see *Discussion*).

3.5. Mean reversion and its amplification by peer comparison

Results in Fig. 6 raise the question whether low baseline consumers behaved inherently differently from high baseline consumers (sometimes referred to as heterogeniety [5,37]), with respect to their usage reductions and to how peer comparisons affected this reduction. In our study, the majority of residents in the treatment group (78%) consumed less than their peers during at least some portions of the feedback period but more than their peers during other portions. In other words, a typical resident that consumed <u>below</u> average during the baseline period still received at least some feedbacks reading "*..., y% <u>more</u> than similar apartments in your building*", and vice versa. To illustrate this, Fig. 7 shows an example of a single resident's electricity usage during each of the 14 feedback cycles (relative to the control group's usage in the same cycle), and how this resident's usage changed frequently from above control to below control.

To analyze how peer comparisons affected each resident's electricity usage in response to the specific peer comparison they actually received ("*..., x% more [or less] than similar apartments in your building*"), we modified the analysis in Fig. 6. Instead of separating "low" from "high" consumers according to their baseline usage (which had been the approach in previous studies [10,37]), we analyzed each resident *r*'s change in usage, from immediately before to after receiving a particular feedback *f* ($\Delta p_{r,f}$), and as a function of that resident's usage immediately before feedback *f* relative to peers ($p^*_{r,f}$; henceforth "recent usage"). The created 504 observations of $\Delta p_{r,f}$ versus $p^*_{r,f}$ (281 for feedbacks with peer comparisons and 57 without). The observations are shown in Fig. 8, revealing the following effects: Feedbacks with peer comparison sent to residents who exhibited recent usage above peer average (right quadrants) had the strongest impact (reduction of 16.4 ± 6.6%, n = 126; $p < 0.01$). Feedbacks with peer comparison sent to residents with recent usage below peer average (left quadrants) also had a strong impact (increase of 9.0 ± 1.8%, n = 281; $p < 0.01$) – however, the impact was in the opposite direction. Feedbacks without peer comparison sent to residents with recent usage above peer average also prompted a reduction (10.1%), but this was not statistically distinguishable from zero.

The observed $\Delta p_{r,f}$ of each individual resident could be partially explained by a linear regression describing a mean reversion, a concept commonly used for asset prices or consumption behavior in economics [55]:

$$\Delta p_{r,f} = MR_{r,pc} \cdot p^*_{r,f} + \varepsilon_{r,f} \tag{12}$$

where $\Delta p_{r,f}$ and $p^*_{r,f}$ are as defined above; *pc* is 1 if feedback *f* to resident *t* included a peer comparison feature and 0 otherwise; $MR_{r,pc}$ is the mean reversion coefficient for resident *r* (separately for *with* vs. *without* peer comparison; *pc = 1* or 0, respectively); and ε is a not further specified error term that captures other resident behavior beyond the mean reversion. Each of the 2·36 $MR_{r,pc}$ was then obtained by minimizing least square errors in the respective regression. A negative $MR_{r,pc}$ means that the particular resident tended to increase his/her usage when receiving a feedback after a period of usage below peer average, and vice versa. Residents tended to exhibit a pronounced mean reversion behavior when receiving peer comparisons: The average $MR_{r,pc=1}$ across residents was $\overline{MR_{r,pc=1}} = -0.41 \pm 0.06$ (n = 36; red shaded area in Fig. 8). Mean reversion was weaker (and statistically indistinguishable from zero) when feedbacks did not include peer comparisons: $\overline{MR_{r,pc=0}} = -0.12 \pm 0.12$ (n = 34, excluding the 2 residents all of whose feedbacks included a peer comparison and whose $MR_{r,pc=0}$ could therefore not be determined via regression; grey shaded area in Fig. 8). Note that some degree of mean reversion (i.e., $MR < 0$) is to be expected even in the absence of any feedback. This is because any resident's electricity usage will remain in line with his/her long term average usage (as opposed to a pure random walk (MR = 0), which would deviate increasingly away (above or below) from that average). A pairwise, two-tailed student *t*-test showed that the difference between $\overline{MR_{r,pc=1}}$ and $\overline{MR_{r,pc=0}}$ was statistically significant ($p < 0.05$, n = 34).

In summary, not just the group as a whole, but each resident individually tended to exhibit a more pronounced mean reversion behavior in response to feedbacks that included a peer comparison than in response to other feedbacks. Implications for devising the most effective eco-feedback in large scale applications are described in *Discussion*.

3.6. Mitigating the boomerang effect in residents with recent usage below peers

Previous studies vary regarding whether and how the boomerang effect can be mitigated (see *Introduction*). Focusing only on residents with recent usage below peers and again tracking their usage from before to after receiving a particular feedback, a boomerang effect is observed in response to feedbacks with peer comparisons but not to feedbacks without peer comparison (9% increase vs. 0%, $p < 0.01$; Fig. 9). Further analyzing the feedbacks with peer comparison as a function of the feedbacks' sentiment, we find that a positive sentiment lowers the boomerang effect but does not fully neutralize it (5% increase vs. 11% increase, $p < 0.05$).

In summary, simply avoiding a peer comparison for residents with recent below-peer electricity usage is more effective in keeping usage low than pairing the peer comparison with additional normative elements (*Discussion*).

4. Discussion and conclusions

While we used our approach only in a relatively small case study (504 observations of 36 residents over ~2 months; 89 control residents), the observed reduction (11.1 ± 3.1% versus control) is consistent with the average reduction found in previous studies that employed feedbacks with social comparisons (11.5%) [5], and higher than the average reduction in other interventions (7%; with the exception of audits and consulting (14%)) [5]. From a policy perspective, since 42% of all residents we approached opted to receive feedback, the energy conservation of the initiative "offer all residents in a building feedback" was (11.1 ± 3.1%)·42% = 4.6 ± 1.3% (residents who declined started with the same average usage as the treatment group). This conservation is in line with the high end of the 1.4–3.3% range reported for the *O-Power* experiments [11] (which assigned feedback recipients randomly).

Our approach offers three advantages over previously employed approaches that use feedback to prompt energy conservation. First, sending *different* feedback types to the *same* resident over time allows one to statistcially isolate the effect of different message features (e.g., self-comparison analyzed separately from peer comparison), not only on different subgroups of residents but on the same resident. Second, the approach allows for the analysis of more complex contributing factors such as message variety over time and the role of a specific resident's most recent energy usage. Third, because feedbacks are generated automatically, the approach is scalable to large field experiments in which utilities could use the data from more and more common smart meters to send feedback messages to tens of thousands of energy users. This would be akin to the *OPower* experiments [11,12], but feedbacks could be sent more frequently than once per month and with more complex and varying feedback types. Such experiments could validate previous work by means of larger sample sizes and extend it to include different demographics and possibly interaction with message features (e.g., is the effect of peer comparisons a function of age?). In settings where appliance end-use can be determined in addition to total electricity use, e.g., via load disaggregation [14], the underlying reasons for the observed higher portion of electricity savings during cooling days (see section 3.1) could also be investigated further.

Based on the 504 observations in our study, the two feedback types prompting the largest average reduction were (i) comparison of a resident's most recent usage with their own usage in the previous feedback (14% reduction); and (ii) feedbacks of high novelty vis-à-vis the previous feedback (16% reduction). As anecdotal evidence for this, one resident told us: "I like the feedbacks – I always try to beat myself." Another told us "I stopped paying attention to the feedbacks whenever I felt they were the same, email after email", underlining the benefit of automatic feedback systems that can produce high message variety. In contrast, comparisons with peers, on average, did not prompt larger reductions. Instead their effect was more nuanced, prompting an increase in low consuming residents (see below). Finally, feedbacks that added information at what time (day or night time) a resident had used the most electricity were less effective than those without this information. We were surprised by this because we had anticipated the opposite, namely that information about when the most usage occurred may help residents to identify the source (e.g., which appliance) and then reduce it. Note however that this feature required a definition of what constitutes "day time" (we chose 9 am–6 pm). We therefore cannot exclude the possibility that other and/or more granular time windows may have made this feature more effective. We therefore interpret our finding regarding "is use time information helpful?" as inconclusive.

Our approach allowed us to evaluate the effect of peer comparisons separately from self-comparisons and other message effects. The mean reversion analysis indicates that the peer comparison prompts not only different subgroups of residents (e.g., energy-efficient residents with low baseline usage), but each resident individually, to follow the social norm of the average. Such behavior, observed empiricially in our analysis, was previously postulated by, e.g., Schultz et al. [10] and interpreted by Knittel et al. [12] as a type of social comparison process, a framework introduced in the 1950ies by Festinger [56]. To use an image: Throughout the experiment, the peer comparisons appear to "shepherd" each individual resident back towards the peers' average, whenever that resident consumed below or above that average.

In line with previous work in energy conservation [10,12,36], we refer to this as a "boomerang effect", but only in the narrow sense that some feedbacks achieved the opposite of their intended effect [11,57] (namely, a usage increase rather than the intended reduction). However, this narrow sense differs in connotation from the broader social/behavioral context in which the boomerang effect, originally coined by Hovland et al. [58], has at times been discussed. Brehm and Brehm's seminal book [59] on the subject is nuanced, e.g., noting the distinction of actual "boomerang effects as opposed to reduced positive influence". Similarly, reviewing respective experimental evidence, Montgomery pointed to the complex overlap in related research on (anti)conformity on one hand versus attitude change on the other [60]. These important distinctions notwithstanding, the boomerang effect has been referred to in the context of "defiance" [60], and Brehm and Brehm's original work [59] has been summarized to state that "the theory of psychological reactance – that people act to protect their sense of freedom – is supported by experiments showing that attempts to restrict a person's freedom often produce an anticonformity 'boomerang effect'" [61]. In contrast to this connotation, a behavioral interpretation of our data might be that none of the observed residents behaved in an anti-conformist manner. Rather, all residents simply strived to conform to their peers. This interpretation would be consistent with previous research, which showed that comparison feedbacks are more effective when residents are shown the electricity usage of friends (in the same building), rather than that of other building residents [54], indicating that the instinct to conform is stronger with friends than with one's neighbors.

How should feedbacks be designed in order to achieve the highest savings? Our findings suggest three possible strategies: (i) Devise feedbacks with deliberate variation from one feedback to the next, to further increase the variation already yielded by the random feature selection employed in our study; (ii) for residents with low recent usage, either do not include a peer comparison at all or (iii) pair the peer comparison with injunctive norms [39], such as a positive sentiment.

CRediT authorship contribution statement

Christoph J. Meinrenken: Funding acquisition, Methodology, Formal analysis, Writing - original draft, Visualization. **Sanjmeet Abrol:** Data curation. **Gaurav B. Gite:** Methodology, Software, Investigation. **Christopher Hidey:** Validation. **Kathleen McKeown:** Funding acquisition, Conceptualization, Methodology, Writing - original draft. **Ali Mehmani:** Data curation. **Vijay Modi:** Validation. **Elsbeth C. Turcan:** Software, Methodology, Writing - original draft. **Wanlin Xie:** Writing - original draft. **Patricia J. Culligan:** Funding acquisition, Conceptualization, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE), under the Building Technologies Office's, BENEFIT program, Award Number DE-EE-0007864, and we thank Erika Gupta and her team for outstanding support during our project. Further we would like to thank the building's management and superintendents for assistance with access to the building and the meter enclosures. We also thank Scott Barrett, Peter T. Coleman, Aven Satre-Meloy, and Erez Yoeli for excellent discussions, especially with regards to placing our results into the context of previous studies around social norms and behavior.

References

[1] IEA, International Energy Agency https://www.iea.org/data-and-statistics/data-tables?country=USA&energy=Electricity&year=2017, 2017.
[2] US-EPA, United States Environmental Protection Agency https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks, 2017.
[3] L. Nicholls, Y. Strengers, J. Sadowski, Social impacts and control in the smart home, Nat. Energy 5 (3) (2020) 180–182.
[4] S.Z. Attari, M.L. DeKay, C.I. Davidson, W. Bruine de Bruin, Public perceptions of energy consumption and savings, Proc. Natl. Acad. Sci. 107 (37) (2010) 16054–16059.
[5] M.A. Delmas, M. Fischlein, O.I. Asensio, Information strategies and energy conservation behavior: a meta-analysis of experimental studies from 1975 to 2012, Energy Policy 61 (2013) 729–739.
[6] R.C. Battalio, J.H. Kagel, R.C. Winkler, R.A. Winett, Residential electricity demand: an experimental study, Rev. Econ. Stat. (1979) 180–189.
[7] A. Faruqui, S. Sergici, A. Sharif, The impact of informational feedback on energy consumption—a survey of the experimental evidence, Energy 35 (4) (2010) 1598–1608.
[8] P.W. Schultz, M. Estrada, J. Schmitt, R. Sokoloski, N. Silva-Send, Using in-home displays to provide smart meter feedback about household electricity consumption: a randomized control trial comparing kilowatts, cost, and social norms, Energy 90 (2015) 351–358.
[9] G. Wood, R. Day, E. Creamer, D. Van Der Horst, A. Hussain, S. Liu, A. Shukla, O. Iweka, M. Gaterell, P. Petridis, Sensors, sense-making and sensitivities: UK household experiences with a feedback display on energy consumption and indoor environmental conditions, Energy Res. Social Sci. 55 (2019) 93–105.
[10] P.W. Schultz, J.M. Nolan, R.B. Cialdini, N.J. Goldstein, V. Griskevicius, The constructive, destructive, and reconstructive power of social norms, Psychol. Sci. 18 (5) (2007) 429–434.
[11] H. Allcott, Social norms and energy conservation, J. Public Econ. 95 (9–10) (2011) 1082–1095.
[12] C.R. Knittel, S. Stolper, Using machine learning to target treatment: The case of household energy use, Working paper 26531 of the National Bureau of Economic Research (http://www.nber.org/papers/w26531), 2019.
[13] A. Satre-Meloy, M. Diakonova, P. Grünewald, Cluster analysis and prediction of residential peak demand profiles using occupant activity data, Appl. Energy 260 (2020) 114246.
[14] K.C. Armel, A. Gupta, G. Shrimali, A. Albert, Is disaggregation the holy grail of energy efficiency? The case of electricity, Energy Policy 52 (2013) 213–234.
[15] S. Darby, The effectiveness of feedback on energy consumption, A review for DEFRA of the literature on metering, billing and direct displays, 486 (2006) 2006
[16] K. Ehrhardt-Martinez, K.A. Donnelly, S. Laitner, Advanced metering initiatives and residential feedback programs: a meta-review for household electricity-saving opportunities, American Council for an Energy-Efficient Economy Washington, DC (Report E105), 2010
[17] O.I. Asensio, M.A. Delmas, The dynamics of behavior change: Evidence from energy conservation, J. Econ. Behav. Organ. 126 (2016) 196–212.
[18] A. Satre-Meloy, M. Diakonova, P. Grünewald, Daily life and demand: an analysis of intra-day variations in residential electricity consumption with time-use data, Energ. Effi. (2019) 1–26.
[19] O.I. Asensio, Correcting consumer misperception, Nat. Energy 4 (10) (2019) 823–824.
[20] J. Froehlich, Promoting energy efficient behaviors in the home through feedback: The role of human-computer interaction, HCIC Workshop & University of Washington tech note #09-02-01, 2009.
[21] K. Buchanan, R. Russo, B. Anderson, The question of energy reduction: the problem (s) with feedback, Energy Policy 77 (2015) 89–96.
[22] J. Froehlich, L. Findlater, J. Landay, The design of eco-feedback technology, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2010, 1999–2008.
[23] C. Fischer, Feedback on household electricity consumption: a tool for saving energy?, Energ. Effi. 1 (1) (2008) 79–104.
[24] K.S. Wolske, K.T. Gillingham, P.W. Schultz, Peer influence on household energy behaviours, Nat. Energy 5 (2020) 202–212.
[25] R.K. Jain, J.E. Taylor, P.J. Culligan, Investigating the impact eco-feedback information representation has on building occupant energy consumption behavior and savings, Energy Build. 64 (2013) 408–414.
[26] D.L. Costa, M.E. Kahn, Energy conservation "nudges" and environmentalist ideology: evidence from a randomized residential electricity field experiment, J. Eur. Econ. Assoc. 11 (3) (2013) 680–702.
[27] M. Bergquist, A. Nilsson, W.P. Schultz, A meta-analysis of field-experiments using social norms to promote pro-environmental behaviors, Global Environ. Change 59 (2019) 101941.
[28] E. Yoeli, D.V. Budescu, A.R. Carrico, M.A. Delmas, J.R. DeShazo, P.J. Ferraro, H.A. Forster, H. Kunreuther, R.P. Larrick, M. Lubell, E.M. Markowitz, B. Tonn, M.P. Vandenbergh, E.U. Weber, Behavioral science tools to strengthen energy & environmental policy, Behav. Sci. Policy 3 (1) (2017) 69–79.
[29] V.L. Chen, M.A. Delmas, S.L. Locke, A. Singh, Information strategies for energy conservation: a field experiment in India, Energy Econ. 68 (2017) 215–227.
[30] T. Dietz, G.T. Gardner, J. Gilligan, P.C. Stern, M.P. Vandenbergh, Household actions can provide a behavioral wedge to rapidly reduce US carbon emissions, Proc. Natl. Acad. Sci. 106 (44) (2009) 18452.

[31] J.M. Nolan, P.W. Schultz, R.B. Cialdini, N.J. Goldstein, V. Griskevicius, Normative social influence is underdetected, Pers. Soc. Psychol. Bull. 34 (7) (2008) 913–923.
[32] T. Hargreaves, L. Middlemiss, The importance of social relations in shaping energy demand, Nat. Energy 5 (3) (2020) 195–201.
[33] R.J. Bator, K. Phelps, J. Tabanico, P.W. Schultz, M.L. Walton, When it is not about the money: social comparison and energy conservation among residents who do not pay for electricity, Energy Res. Social Sci. 56 (2019) 101198.
[34] R.K. Jain, R. Gulbinas, J.E. Taylor, P.J. Culligan, Can social influence drive energy savings? Detecting the impact of social influence on the energy consumption behavior of networked users exposed to normative eco-feedback, Energy Build. 66 (2013) 119–127.
[35] T. Harries, R. Rettie, M. Studley, K. Burchell, S. Chambers, Is social norms marketing effective? A case study in domestic electricity consumption, Eur. J. Market. 47 (9) (2013) 1458–1475.
[36] I. Ayres, S. Raseman, A. Shih, Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage, J. Law, Econ., Organ. 29 (5) (2013) 992–1022.
[37] O.I. Asensio, M.A. Delmas, Nonprice incentives and energy conservation, Proc. Natl. Acad. Sci. 112 (6) (2015) E510–515.
[38] A. Kažukauskas, T. Broberg, J. Jaraite, Social comparisons in real time: A field experiment of residential electricity and water use, (2017).
[39] R.B. Cialdini, C.A. Kallgren, R.R. Reno, A focus theory of normative conduct: a theoretical refinement and reevaluation of the role of norms in human behavior, Adv. Exp. Soc. Psychol. 24 (1991) 201–234.
[40] B.S. Everitt, The Cambridge Dictionary of Statistics, Cambridge University Press, 2003.
[41] C.J. Meinrenken, N. Rauschkolb, S. Abrol, T. Chakrabarty, V.C. Decalf, C. Hidey, K. McKeown, A. Mehmani, V. Modi, P.J. Culligan, MFRED, 10 second interval real and reactive power for groups of 390 US apartments of varying size and vintage, Sci. Data 7 (1) (2020) 375.
[42] Y. Song, Y. Ding, P. Siano, C. Meinrenken, M. Zheng, G. Strbac, Optimization methods and advanced applications for smart energy systems considering grid-interactive demand response, Appl. Energy 259 (2020) 113994.
[43] M. Zheng, C.J. Meinrenken, K.S. Lackner, Electricity storage in buildings for residential sector demand response: Control algorithms and economic viability evaluation, National Institute of Standards and Technology (GCR 14-978), 2014.
[44] C.J. Meinrenken, Z. Shou, X. Di, Using GPS-data to determine optimum electric vehicle ranges: a Michigan case study, Transp. Res. D Transp. Environ. 78 (2020) 102203.
[45] M.L. Zhen, J. Sun, C.J. Meinrenken, T. Wang, Pathways toward enhanced techno-economic performance of flow battery systems in energy system applications, J. Electrochem. Energy 16 (2) (2019) 1–11.
[46] C.J. Meinrenken, A. Mehmani, Concurrent optimization of thermal and electric storage in commercial buildings to reduce operating cost and demand peaks under time-of-use tariffs, Appl. Energy 254 (2019) 113630.
[47] S. Abrol, A. Mehmani, M. Kerman, C.J. Meinrenken, P.J. Culligan, Data-enabled building energy savings (D-E BES), Proc. IEEE 106 (4) (2018) 661–679.
[48] C.J. Meinrenken, D. Chen, R.A. Esparza, V. Iyer, S.P. Paridis, A. Prasad, E. Whillas, Carbon emissions embodied in product value chains and the role of Life Cycle Assessment in curbing them, Sci. Rep. 10 (1) (2020) 6184.
[49] US-EPA, Greenhouse Gases Equivalencies Calculator United States Environmental Protection Agency (www.epa.gov/energy/ghg-equivalencies-calculator-calculations-and-references; accessed February 2017), 2017.
[50] CityOfNewYork, Inventory of New York City Greenhouse Gas Emissions in 2014, Mayor's Office of Sustainability, 2016
[51] K.V. Deemter, M. Theune, E. Krahmer, Real versus template-based natural language generation: a false opposition?, Comput Linguist. 31 (1) (2005) 15–24.
[52] S. Baik, A.L. Davis, J.W. Park, S. Sirinterlikci, M.G. Morgan, Estimating what US residential customers are willing to pay for resilience to large electricity outages of long duration, Nat. Energy 5 (3) (2020) 250–258.
[53] G.B. Gite, Extractive Summarization of Individual News Articles – Using Structured Prediction Energy Networks, Columbia University Master Thesis, 2017.
[54] G. Peschiera, J.E. Taylor, J.A. Siegel, Response–relapse patterns of building occupant electricity consumption following exposure to personal, contextualized and occupant peer network utilization data, Energy Build. 42 (8) (2010) 1329–1336.
[55] M.J. Kim, C.R. Nelson, R. Startz, Mean reversion in stock prices? A reappraisal of the empirical evidence, Rev. Econ. Stud. 58 (3) (1991) 515–528.
[56] L. Festinger, A theory of social comparison processes, Human Relat. 7 (1954).
[57] M.A. Clee, R.A. Wicklund, Consumer behavior and psychological reactance, J. Consumer Res. 6 (4) (1980) 389–405.
[58] C.I. Hovland, I.L. Janis, H.H. Kelley, Communication and Persuasion, Yale University Press, New Haven, CT, USA, 1953.
[59] S.S. Brehm, J.W. Brehm, Pschological reactance – A theory of freedom and control, Academic Press, Harcourt Brace Jovanovich Publishers, New York, 1981. ISBN: 0-12-129840-X
[60] R.L. Montgomery, social influence and conformity: a transorientational model, in: D. Granberg, G. Sarup (Eds.), Social Judgment and Intergroup Relations. Springer, New York, NY, 1992. DOI: 10.1007/978-1-4612-2860-8_8
[61] D.G. Myer's, Social Psychology –- Chapter 6: Conformity and obedience McGraw Hill, New York, 2010