Building and Environment 46 (2011) 1891–1898

Contents lists available at ScienceDirect

Building and Environment

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

Usability of residential thermostats: Preliminary investigations

Alan Meier [a,*], Cecilia Aragon [a], Therese Peffer [b], Daniel Perry [b], Marco Pritoni [c]

[a] Lawrence Berkeley National Laboratory, Berkeley, California, USA
[b] University of California, Berkeley, California, USA
[c] University of California, Davis, California, USA

ARTICLE INFO

Article history:
Received 30 November 2010
Received in revised form
21 March 2011
Accepted 22 March 2011

Keywords:
Programmable thermostat
User interface
Amazon mechanical turk
Usability test procedure
Energy star
Energy conservation

ABSTRACT

Residential thermostats control 9% of the total energy use in the United States and similar amounts in most developed countries; however, the details of how people use them have been largely ignored. Five parallel investigations related to the usability of residential thermostats were undertaken. No single investigation was representative of the whole population, but each gave insights into different groups or usage patterns.

Personal interviews revealed widespread misunderstanding of thermostat operation. The on-line surveys found that most thermostats were selected by previous residents, landlords, or other agents. The majority of occupants operated thermostats manually, rather than relying on their programmable features and almost 90% of respondents reported that they rarely or never adjusted the thermostat to set a weekend or weekday program. Photographs of thermostats were collected in one on-line survey, which revealed that about 20% of the thermostats displayed the wrong time and that about 50% of the respondents set their programmable thermostats on "long term hold" (or its equivalent). Low-income families were visited and their thermostats photographed. Even though 85% of the respondents declared that they use programming features to automatically raise or lower the temperature, the photos indicated that 45% were in hold. Laboratory tests were undertaken to measure usability of thermostats. A measurement protocol was developed and a metric was created that could quantitatively distinguish usability among five thermostats. This metric could be used to establish minimum levels of usability in programmable thermostats and other energy-using devices with complex controls.

1. Introduction

Residential thermostats have been a key element in controlling heating and cooling systems for over sixty years. During this period, consumer expectations regarding the quality of the indoor thermal environment have increased. People expect thermostats, by controlling the heating and cooling systems, to carefully regulate temperatures, respond rapidly to changes in preferences or outside conditions, all with only infrequent input by the occupants. Modern, programmable, thermostats are typically marketed as "energy-saving" and consumers typically justify their purchase with this goal in mind.

Residential thermostats have been relatively ignored as a focus of research. This may be surprising given that they control 9% of the total energy use in the United States [1] and similar amounts in most developed countries. With such a large amount of energy in play, it is essential to understand the thermostat's technology and the way the occupants interact with them. Furthermore, thermostats themselves are undergoing a dramatic change in capabilities. Today's thermostats generally control only temperature; however, in the near future they may control ventilation and humidity, and take into consideration occupancy and the price of the energy. Finally, thermostats are being connected to the Internet and expanded networks inside homes, suggesting that controls are likely to become much more complex.

In 1995, the Energy Star Program established technical specifications for "energy-saving" programmable thermostats. Many building codes and government programs require installation of programmable thermostats because of their assumed energy savings. Nevertheless, there have been few careful studies of the energy savings attributable to these thermostats. Several recent field studies have found no significant savings in households equipped with programmable thermostats compared to households with manual thermostats [2–5]. Two other studies found that homes relying on programmable thermostats actually consumed more energy than those where the occupants set the thermostats manually [6], especially in homes equipped with heat pumps [7]. Anecdotal evidence suggested that the thermostats were overly complex and

* Corresponding author. Energy Analysis Department, Lawrence Berkeley National Laboratory, MS90R2000, Berkeley, CA 94720, USA. Tel.: +1 510 486 4740.
E-mail address: akmeier@lbl.gov (A. Meier).

0360-1323/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.buildenv.2011.03.009

that consumers were unable to operate them in a way that obtained energy savings compared to manually operated thermostats. As a result, Energy Star terminated the thermostat endorsement program in 2009.

We describe below the results from five parallel investigations related to the usability and actual use of residential thermostats. They focus on programmable thermostats because programmable thermostats now represent about 40% of thermostats in existing US homes and nearly 100% of thermostats in new homes. The studies were designed to assess the extent to which the occupants were able to successfully exploit the new features of programmable thermostats. None of the studies attempts to be comprehensive, yet each offers different insights into the way in which people interact with thermostats.

2. Earlier studies of thermostat usability in the literature

The performance of specific components – the furnace, compressor, heat exchanger, fan, etc. – of heating and cooling systems has been studied in great detail over the years. Curiously, the thermostat, that is, the control of the heating and cooling system, has received relatively less attention. A survey of the literature broadly dealing with thermostats was recently undertaken by researchers at Lawrence Berkeley National Laboratory [8].

The *usability* of thermostats has been the subject of even less research even though it is a popular complaint and topic for anecdotes. To be sure, thermostat manufacturers have undertaken research into the effectiveness of their designs, but the results have been mostly confined to proprietary reports. Manufacturers consider any insights gained through their usability studies to be a competitive advantage. Furthermore, manufacturers tend to focus on their own products rather than examining generic effectiveness of the devices.

Researchers have periodically commented about usability problems associated with thermostats both when specifically examining thermostats or in the course of other research. Table 1 summarizes the usability problems identified in the literature. Surprisingly few comments have been made over the past twenty years, especially compared to investigations of other components in heating and cooling systems.

An important concept is the mental model assumed by thermostat users. Kempton [24] used ethnographic methods to interview occupants and building supervisors to derive insights. For example, many occupants treated thermostats more like a valve rather than a switch. Thus, the occupants expected heat to be delivered faster when they set higher temperatures. This led to energy-wasteful operating outcomes because indoor temperatures would overshoot desired temperatures. (Our own research indicates that this remains a popular mental model [25].)

Table 1
Usability problems associated with programmable thermostats identified in the literature (Note: "PT" = Programmable Thermostat.)

Programmable Thermostats Complaints/Issues	References
PTs are too complicated to use	[9–18], [4], [19]
Buttons/fonts are too small	[10], [20], [12], [13], [21], [18]
Abbreviations and terminology are hard to understand; lights and symbols are confusing	[20], [12], [13], [16], [22], [18]
The positioning of interface elements is illogical	[20], [12], [18]
PTs are positioned in an inaccessible location	[16], [21]
Setting the thermostat is troublesome	[14], [17], [4], [21]
It is difficult to set time and date	[10]
PTs give poor feedback on programming	[16], [18]
PTs are not attractive to use	[23]

Problems with thermostats are not limited to North America and the unique heating systems found there. In Finland, Karjalainen [26] conducted qualitative and quantitative surveys on thermostat use in homes and offices. He concluded that many people had misconceptions about how thermostats and their heating systems actually operate (such as treating the thermostat as a valve rather than a switch) and that they found thermostats too complicated to use with confidence.

In the UK, Rathouse and Young [19], conducted six focus groups to investigate issues in use of heating controls. Based on the users' experiences and complaints, Rathouse and Young formulated recommendations for manufacturers and installers including that manufacturers offer a variety of products of different complexity to suit different needs.

Consumer magazines occasionally evaluate thermostats. Usability is typically one of the factors considered in the overall ratings. These evaluations generally took place in conditions where usability problems would be minimized. For example, when *Consumer Reports* [10] evaluated fifteen thermostats, the tests were conducted in a well-illuminated room, by highly-trained panelists comfortably seated at a table (a situation rarely encountered in homes). Even then, the panelists found some of the thermostats difficult to use. Consumer magazines in other countries, notably Germany [27] and Sweden [28] have also reviewed thermostats. Both investigations included usability as a consideration but only in a qualitative sense. Heating controls are somewhat different in Europe because the heating technologies are different; in addition, few residential systems include cooling.

In spite of the relatively sparse literature describing usability problems associated with thermostats, many attempts to design more usable thermostats have been undertaken by manufacturers, researchers, and students. In Human Factors courses at universities, designing a more user-friendly thermostat is a popular assignment. This is another indication of the observed poor usability of these devices. Nevertheless, few groups have tried to document the extent of poor usability before embarking on new designs.

There appears to have been an upsurge in activity related to designing new thermostats. Many small firms—often with roots in Silicon Valley—have entered the market. We attribute this to declining costs of key components (logic circuits, displays, and communications), expertise in design processes developed for smart phones, easier connections to the Internet, and the prospect of time-of-use pricing for electricity. Thermostats are also gradually becoming less like an appendage to the home's heating and cooling system and more like a new category of consumer electronics.

3. Field evaluations of programmable thermostats

3.1. Approach to evaluations

We undertook a wide range of studies to determine the extent to which occupants were able to successfully use the features of programmable thermostats. We chose them in order to learn what kinds of data could be collected, how useful a larger survey would be, and to give us insights to specific groups (such as low-income users). These studies included:

1. Personal interviews with people regarding their thermostat habits
2. An on-line survey
3. An on-line survey supplemented with respondent-supplied photographs of their thermostats
4. A survey of homes participating in a weatherization program
5. Laboratory tests of people's ability to perform tasks on thermostats

No single study was comprehensive and all were able to collect limited data, but each sought to capture different groups or answer specific questions. Details and results of the studies are given below.[1]

3.2. Personal interviews regarding thermostat habits

We administered six semi-structured qualitative interviews in Berkeley and San Francisco (California, USA) to assess whether people generally understand how thermostats work. The age of interviewees (students or professionals) spanned from mid-twenties to late-thirties. Interviews were recorded in the interviewees' houses.

Many of the complaints and problems listed in Table 1 also emerged in our interviews. For example, improper placement of thermostats such as on another floor or in the warmest/coolest room affects the accuracy of sensors and has consequences on settings. We found that in the majority of the households thermostats were improperly positioned. Moreover, location can affect the readability of the device. For example, in one household the thermostat was installed in a dark hallway and rotated 90°. Many people admitted to using their programmable thermostat as an on/off switch instead of programming it, as reported in previous studies. The users in the remaining households rarely changed the scheduled settings, often only once per year, and instead frequently used override modes. Most of the interviewees showed little knowledge of the thermostat and a few were also worried about touching it: "I don't touch it because I don't understand it" or "I don't want to mess it up". A further complicating factor in assessing the effects of controls on energy consumption is that supplementary heating/cooling systems, such as fans, additional air conditioning units and fireplaces, were often used and not controlled by the thermostats.

3.3. An on-line survey of thermostat usage patterns

We created several on-line surveys, each seeking to probe different aspects or targeting a different audience. We posted a 15-question survey through Facebook, Craigslist, and other on-line distribution channels. No compensation was provided for completing the survey. Questions ranged from brand and placement of thermostats within the home to the respondents' perceived effectiveness and specified temperature settings. We also asked the respondents about their interactions with the device, such as frequency of adjustments and use of hold modes. We collected limited demographic information, including geographic location and primary household language, to help us better understand the respondents.

The on-line surveys yielded 81 respondents from ten US states and 57 cities. The surveys revealed that occupants often operated thermostats manually rather than relying on their programmable features. About 89% of respondents reported that they rarely or never adjusted the thermostat to set a weekend or weekday program. About 54% of respondents used the on/off switch at least weekly.

Only 19% of thermostats were purchased by the respondents or someone in their household. Thus, we can conclude that most thermostats were selected by previous residents, landlords, or other agents. This has important implications regarding delivery of instructions for use of the devices and availability of operating manuals. In addition, most occupants have thermostats with features that are not necessarily matched to the occupants' needs and abilities.

[1] Additional studies are still in progress and will be reported elsewhere.

3.4. An on-line survey supplemented with respondent-supplied photographs

Amazon Mechanical Turk [29] is an electronic marketplace that matches requestors and suppliers for web-related tasks, such as adding tags to figures, searching figures, and proofreading. Tasks are typically very small—no more than a minute—and the payments are correspondingly small (typically only a few cents). The workers are anonymous and paid through a third party (Amazon). In November 2010, Amazon claimed that over 77,000 "Human Intelligence Tasks" (or HITs) were on offer. Amazon has not released the number of "Workers" registered or active although it almost certainly exceeds the number of active HITs. The demographics of Mechanical Turk Workers have been examined in some detail [30]. The population is surprisingly diverse but is mostly women. The average age is about 36, which is slightly younger than the U.S. as a whole. The Workers are also slightly better educated but earning a lower income than the population as a whole.

The Mechanical Turk is an attractive source of information because one can obtain a rapid and geographically dispersed response. The service also provides automatic tabulation of the results. All of this is available at a very modest cost.

We used Mechanical Turk as another source of thermostat users but also as a means of obtaining additional information. We offered Mechanical Turk Workers the job of filling out a one-page form about their thermostat practices and paid them $2.00 to upload photographs of their thermostats. In this way, we were able to determine the settings. This appears to be the first research study where Mechanical Turk Workers were asked to supplement a survey with a photograph.

We obtained 63 responses to our offer in only 24 h, 15 of which included photographs. The geographic distribution of responses is shown in Fig. 1. Most of the photographs uploaded by the Workers were clear enough to allow us to make determinations of the settings on the thermostats. We relied on the Workers to compare the thermostat's displayed time to the actual time.

Based on the Workers' responses, 20% of the thermostats displayed the wrong time (half of them by more than 1 h). This is important because programming is unlikely to be effective when the thermostat relies on the incorrect time. Alternatively, the incorrect time setting may signal that the thermostat is not programmed. We also found that about 50% of the respondents set their programmable thermostats on "long term hold" (or its equivalent). For whatever reason, these users chose not to—or were unable to—exploit the energy-saving features of programmable thermostats.

Only certain kinds of information can be gleaned from the Mechanical Turk survey (and the associated photographs) but even this is valuable. The Workers appear to have a higher than average ability to work with computers and other electronic controls; nevertheless, even this skilled group chose not to program their thermostats.

One advantage of the Mechanical Turk on-line approach is simplification of the survey tool. Questions can be tailored to the season (or even day) of the survey and extraneous questions related to other seasons can be eliminated. In our study, we eliminated winter heating questions because we sent the survey in the summer. We plan to repeat the survey in other seasons to identify constancy of behavior.

3.5. A survey of homes participating in a weatherization program

Twenty low-income families in Wisconsin were surveyed during the late winter 2010. A local weatherization organization selected the homes, administered the survey, and photographed

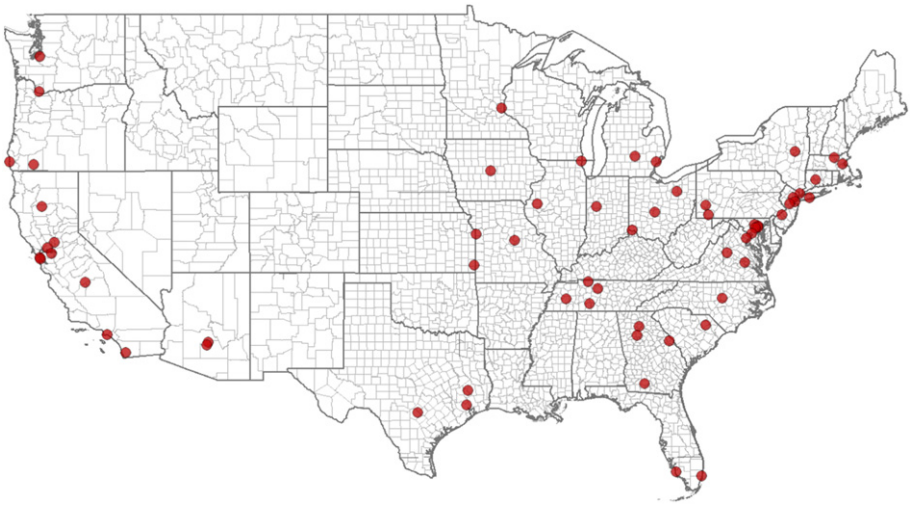

Fig. 1. Distribution of responses to Mechanical Turk survey.

the thermostats. No special selection criteria were used; these homes were simply the cohort being weatherized that week. The organization received $25 per home. The survey consisted of only ten multiple-choice questions because the crew had little time to interview the occupants and program requirements to maintain confidentiality.

The photos contained objective data, such as type and brand of thermostat, use of hold mode, accuracy of time settings, temperature set and any other characteristic noticeable from the display. We then compared these objective data with self-reported questionnaire responses, and noted the inconsistencies.

The survey of low-income homes identified three distinct thermostat usage patterns. About a third of the interviewees reported that they change settings every day, a third never change the thermostat, and a third weekly at most. Although most (85%) of the respondents declared that they use programming features to automatically raise or lower the temperature, the photos indicated a very different breakdown:

- 45% were in hold
- 30% were programmed
- 10% were manual thermostats (not programmable)
- 5% were off
- 10% operational status was not visible in the picture

This small survey (again unrepresentative) reveals that people either do not understand what "programmable" or "settings" mean, or they know very little about how to operate the thermostat. In light of these findings, the accuracy of self-reported behavior or settings should be considered carefully.

We plan to collect more information in collaboration with weatherization programs, covering different regions and seasons. These results will help the weatherization programs train installers, educate occupants, and select the thermostats most likely to save energy.

4. Laboratory tests to measure usability

4.1. Approach

Our review of the literature found no earlier attempts to quantitatively measure usability of thermostats or similar devices. We therefore investigated the feasibility of quantifying usability of thermostats in laboratory studies. Our long-term goal is to develop a method to establish a "usability score" so that manufacturers, consumers, and regulatory agencies can rank thermostats and establish minimum criteria for usability. The methodology will eventually involve four steps:

1. Define representative tasks to be accomplished with the device;
2. Measure people's ability to perform those tasks under controlled conditions using defined metrics.
3. Compute a "usability score" based on measurements for several tasks; and
4. Compare scores to a reference interface.

We report here results from only the first two steps; however, steps 3 and 4 are shown to indicate the direction and goals of future research.

The first step in measuring usability is defining the most common tasks associated with the thermostat. A "task" might be as simple as ascertaining the status of the thermostat; for example, "Identify the temperature the thermostat is set to reach". Alternatively, a task might involve changing the operation, such as, "Program the temperature to be 70 °F on Tuesday evenings at 7 PM." This approach is consistent with internationally accepted definitions of usability [31]. We assembled a list of tasks by studying the operating manuals and carefully observing and interviewing users. From a long list of tasks, we selected six that typified the range of tasks a typical user would need to understand in order to effectively operate the programmable thermostat. The list was further constrained by requiring that the tasks could be accomplished with most common programmable thermostats. The six tasks selected were:

Task 1: Turn the thermostat from "off" to "heat".
Task 2: Set the correct time on the thermostat's clock.
Task 3: Identify the temperature the device is set to reach.
Task 4: Identify what temperature the thermostat is set to reach on Thursday at 9:00 PM.
Task 5: Put the thermostat in "hold" or "vacation" to keep the same temperature while gone.
Task 6: Program a schedule and temperature preferences for Monday through Friday.

The above tasks are clearly defined and can be easily explained to test subjects. Successful operation of a programmable thermostat

requires proficiency in other tasks but these are representative; in other words, if users can perform these tasks, then they can use the most important programming features of the thermostat.

Five programmable thermostats were selected for testing, three were primarily touch screen controlled, one used buttons, and one used a web-based interface. The tests were conducted at a usability laboratory. A video camera recorded each test in the vicinity around the thermostat (so the subject's face was not captured). An image from a video is shown in Fig. 2.

Twenty nine subjects were recruited through on-line classified postings to sections for "creative gigs" and "labor gigs" in the San Francisco Bay area. Two were recruited from a similar posting to a university e-mail list. Their ages ranged from 18 to 65. Of the 31 participants, nine were female. All participants were given a small financial incentive for taking part in the study. Participants came from varied occupations and backgrounds, including construction workers, business managers, non-profit staff, maintenance workers and students. Participants were asked to rate their previous experience with programmable thermostats. Seventeen people reported their experience level with programmable thermostats as "low," eight as "moderate" and five reported having "no experience with programmable thermostats" (one participant gave no response).

Each subject was tested on two thermostats. Each test consisted of six tasks. Altogether 62 tests were performed, consisting of 372 tasks. The subjects did not receive any training prior to being tested; however, an operating manual was placed on a table next to the thermostat which they could consult if they wished.

The videos allowed us to observe in detail and record for each task the following:

- success or failure in accomplishing the task
- elapsed time to accomplish the task
- number of times buttons were pushed (or other actions)
- sequence of actions
- hesitations and comments of users

In this way we were able to convert a video record into several possible metrics of usability.

Our initial goal was to determine the viability of the task-based methodology and the identification of the best metric. We assumed that the subjects will vary in their abilities, but did the test procedure generate a significant range in the values of the metrics? Second, we assumed that the thermostats will vary in usability, but did the test procedure generate a significant range in the values of the metrics? Finally, was one metric superior to others?

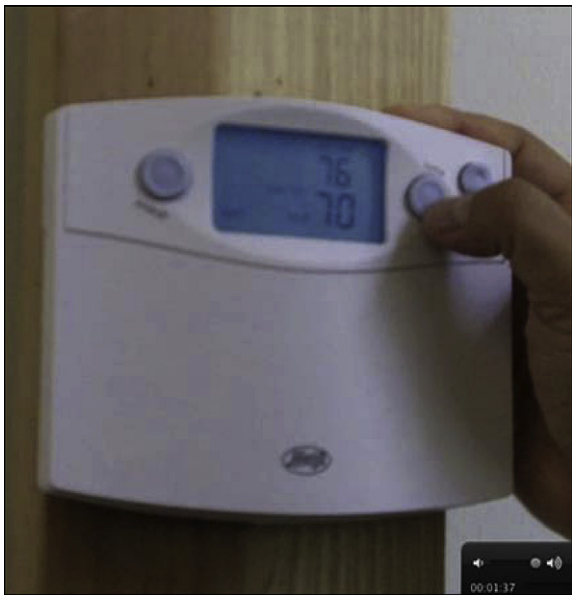

Fig. 2. Still image from video of subject performing a task.

4.2. Results

A wide range of usability was observed. For example, in Task 1,that is, switch the thermostat from off to set heat, 26% of the subjects were unable to accomplish the task at all. Fig. 3 displays the completion fraction for each thermostat. For thermostat A, all subjects successfully completed the task. In contrast, only 50% of the subjects using thermostat E completed the task. Similar results occurred for the other tasks (but are not presented here).

In Fig. 4 the elapsed time for each subject to accomplish (or fail to accomplish) Task 1 is plotted for each thermostat. Some subjects were able to accomplish Task 1 in less than 10 s. Most subjects were able to accomplish the task in less than 30 s; however, over 30% of the subjects required 31–120 s (Note that 2 min can feel like a very long time when standing in front of a thermostat trying to switch on the heat). These results indicate the range in the subjects' ability to perform this task. Similar results occurred for the other tasks (but are not presented here).

The time for subjects to accomplish the task varied widely for the same thermostat, too. For Thermostat C, one person successfully switched on the heat in 20 s, while another required 260 s. The remaining times were evenly distributed between the minimum and maximum times. Wide ranges in elapsed time occurred for Thermostats C and E but much less so for A, B, and D. Elapsed times for completion and success rates do not have a clear correlation. Compare thermostat A, which had 100% success rate and very low elapsed times, to thermostat E, which had a low success rate and a wide range in elapsed times.

Thermostats D and E had hinged covers which concealed some of the controls. Many subjects were unable to open the cover or did not recognize that a cover existed, resulting in more failures to complete. This illustrates how small design differences can have large impacts on successful operation of a device. Note that this task captures a subject's first encounter with the thermostat; the results could change once he or she becomes familiar with its operation. However, a continuing lack of familiarity—or "forgettability"—may be a reasonable assumption if occupants rarely interact with their units.

The results shown in Figs. 3 and 4 (and other results not shown here) demonstrated that the methodology produced a wide range of measured abilities of the subjects to perform the task.

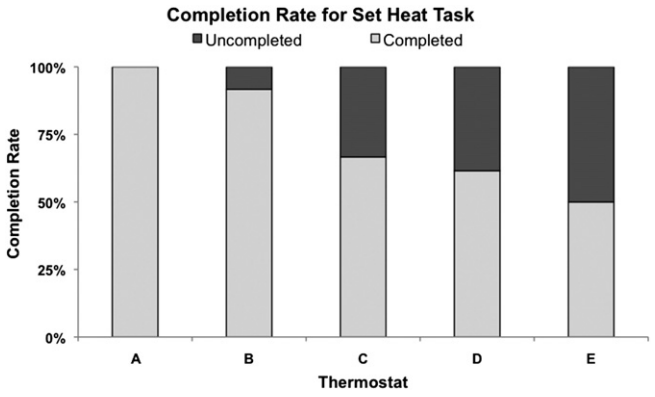

Fig. 3. Fraction of subjects that successfully completed the Set Heat Task.

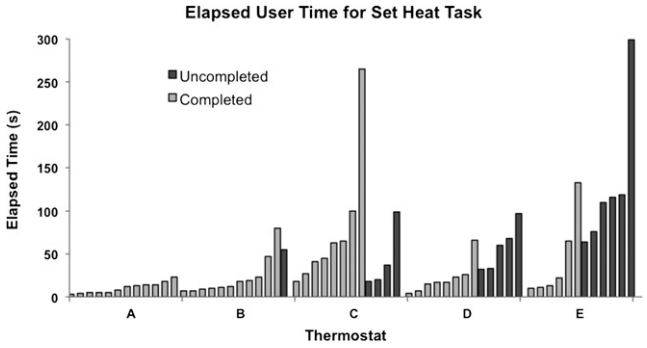

Fig. 4. Elapsed times for subjects to perform the Set Heat Task, including times for those who were not successful.

A second requirement for the task-based methodology is the ability to quantitatively differentiate levels of usability among thermostat interfaces. Fig. 4 displays the range in elapsed time to completion for accomplishing Task 1 with the five thermostats. The figure demonstrates that the task-based methodology and the metric achieved clear differentiation among the thermostats. The average time to accomplish Task 1 for Thermostat E was roughly eight times longer than for Thermostat A. Thermostats A and B were clearly superior (for this task) because the subjects were able to accomplish the task quickly and nearly all of the subjects successfully completed the task. In contrast, the subjects accomplished Task 1 on Thermostat D relatively slowly and a significant fraction was unable to complete it at all.

The results were similar for other tasks. Fig. 5 shows the average elapsed times for Tasks 1, 3, and 4. A wide range in average completion time was observed in all three tasks. The ranking of thermostats changed slightly depending on the task but, in general, a model with long average completion times for one task had long completion times for other tasks. Note that results for Tasks 2, 5, and 6 are still being evaluated but appear to be similar to the first three tasks presented here.

Figs. 4 and 5 used the metric of elapsed time for comparison of subjects and thermostats. Other metrics were investigated, including, the percentage of subjects that completed the task, the number of button pushes, and the ratio of observed button pushes divided by the minimum required. We found that all of the metrics produced sufficient ranges in results and all of the metrics generated the same ranking of usability for almost every task. The consistency of these results points to the robustness of the overall task-based approach to measuring usability.

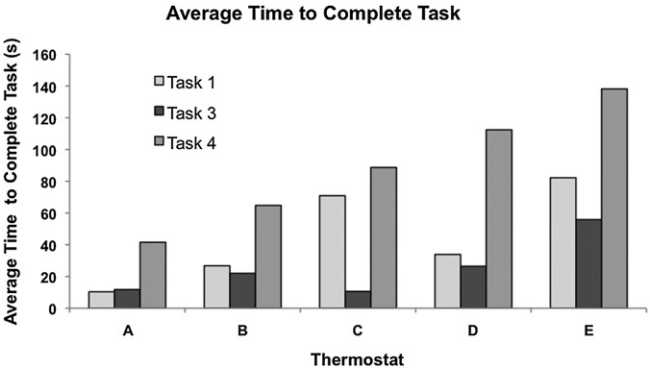

Fig. 5. Average elapsed times for subjects to complete Tasks 1, 3, and 4.

4.3. An improved usability metric

Average elapsed time for completion is an attractive metric because it is simple to understand and measure; however, elapsed time is misleading since the metric ignores those who fail to complete the task. We therefore explored a hybrid metric, combining both elapsed time to complete and successful completion of the task. We also sought to develop a metric that would be easier for manufacturers, regulators, and other stakeholders to interpret and compare. A common drawback of many usability metrics is that the value of the metric is unbounded and varies from task to task. This creates confusion; it is not obvious what value of a metric is "good" and the metric cannot be compared on an absolute scale from one task or device to another. For programmable thermostats, stakeholders need a single measure of usability to facilitate consumer understanding and to create an absolute scale of usability that is not dependent on arbitrary task length. In order to create such a metric, we chose a variant of the logistic function,

$$\frac{2}{1+e^{x}}$$

that maps $[0, \infty)$ to the interval $[1, 0)$. In other words, a shorter time for completion is mapped to a value close to 1, and a longer time is mapped to a value closer to 0.

We also needed to account for success rates on a per-trial basis (where a task "trial" is a single instance of a participant performing a task on a thermostat model, also sometimes called a "task observation"), rather than averaging overall trials of a given task. In order to accomplish this, we incorporated the task completion or success rate variable, s, directly into our primary equation, which we called the "M" statistic. The "M" statistic is calculated as follows on a per-trial basis:

$$Mi = \frac{2s}{1+e^{xi}}$$

where

$$x = t/k$$

$$s = \begin{cases} 0, \text{if subject failed to complete task} \\ 1, \text{if subject completed task} \end{cases}$$

t = time for subject to complete the task (seconds)
k = 50 (empirically determined constant)

Note that M will always be normalized between 0 and 1. The success rate variable, s, also always falls between 0 and 1. It can be a binary variable (where $s = 1$ if the task is completed and 0 otherwise), have multiple values for partial success (e.g. if the task has several subparts that can be completed successfully), or be a continuous variable that measures percentage of task completion.

The M-statistic combines time to complete the task with success of the trial in an intuitive manner: if the task is not completed so that $s = 0$, the value of the M-statistic is 0. Intuitively, this means that if the task was not completed, it should not matter how long the user spent attempting it; it is still a failure. If, on the other hand, the task is completed successfully, then the time on task weighs into the M-statistic. For example, a shorter task duration will yield a higher value of M, a longer task duration will yield a lower value of M, and an uncompleted task will set $M = 0$.

We found that a metric combining time to completion and success to complete was the most practical (which we called the "time and success metric"). The results for the three tasks, using the M-statistic, are shown in Fig. 6. An analysis of variance (ANOVA) showed that, for the time and success metric, the effect of thermostat model on usability was significant at $p < .01$.

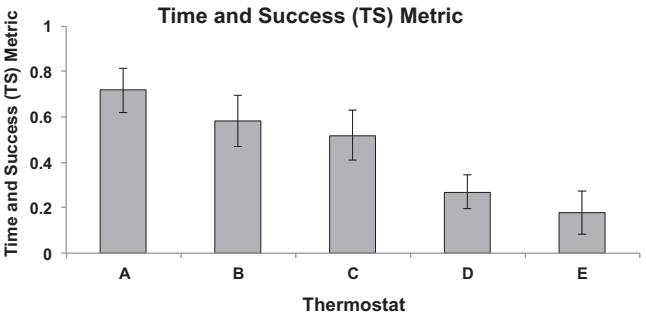

Fig. 6. Combined time and success metric for Tasks 1, 3, and 4.

Fig. 6 shows mean values, along with error bars at the 95% confidence level. Both of the concepts, time to completion and success to complete, are easy to understand. Furthermore, they are easy to measure in a laboratory with relatively simple equipment. These features make the time and success metric an attractive metric for quantifying usability.

5. Discussion

Each of these five investigations—though preliminary and limited in scope—provided insights into the ways people understood and operated programmable thermostats. All of these investigations found qualitative and quantitative evidence of usability problems. Furthermore, the problems spanned wide ranges in age of the users, income, and facility with technical devices.

Some people were intimidated by the thermostats and afraid to touch them for fear of not being able to restore the original settings. A large fraction of occupants left their thermostats permanently set to a single temperature, or used them as on/off switches, which bypassed the thermostat's advanced, energy-saving features. Some occupants mistakenly believed that they were using the programming features (but were not).

Programmable thermostats are a fundamentally different kind of energy-efficiency measure (compared to, say, adding insulation) because they require further consumer actions to save energy: the occupants must program the thermostats to shift to lower temperatures (in the winter). We had the impression, based on the responses to survey questions and supplementary information, that some consumers believed that purchasing and installing an "energy-saving programmable" thermostat would automatically result in lower energy use. Enhanced consumer education and assistance during installation will be necessary to increase energy savings from programmable thermostats. Still other policy measures may be needed to complement the usability improvements and education programs.

These investigations demonstrate the feasibility of collecting information about thermostat usage behavior through diverse methods, such as the Amazon Mechanical Turk, interviews, and observation. We believe that more complex investigations are possible, especially if the surveys can be supplemented with photographs and other forms of documentation. Surveys can be simplified; for example, questions about air conditioning can be eliminated when surveys are conducted in January. Designers and manufacturers could test new interfaces quickly and economically (but not confidentially).

Each of these investigations also has significant limitations. We kept on-line surveys, such as the Mechanical Turk, short to maximize the response rate. These surveys are attractive because they potentially can gather large samples quickly and economically. On the other hand, the investigations can realistically answer only a few, narrow, questions (such as if the occupants set the thermostat on long term hold or set the clock to the correct time). In future studies we would like to gather key demographic and economic information, as long as the response rates to surveys remain high. The value of these investigations might be increased by linking them to larger, representative, surveys, such as the national Residential Energy Consumption Survey (RECS) or regional surveys undertaken by utilities. (We hope to investigate this in future work.)

One goal of this research was to give researchers new tools to more effectively evaluate the impact of thermostats. We developed a procedure to measure usability of based on subjects performing a set of representative tasks needed to effectively operate the thermostats. We tested several metrics of usability and found that they all gave essentially the same rankings. One metric seemed best because one could easily calculate a usability score. It also captured two key aspects, time to complete and ability to complete. The methodology appears to be robust and will allow manufacturers—and regulators—to quantify a thermostat's usability. However, these results can only be considered "preliminary" because further research will be needed in several areas, including:

- How many tasks need to be created to adequately represent overall usability? Every test procedure is a trade-off between realism, cost, and repeatability.
- To what extent should the tests take into account subjects learning and becoming familiar with interfaces? The subjects' performance might change dramatically if the tests were immediately repeated.
- How many people should be on a user test panel and how should they be selected? These questions require guidance from both statisticians and policymakers. On the statistical side, panel size and make-up will determine confidence in the results. Policymakers need to decide to what extent elderly, disabled, color-blind, and non-English speakers should be included.
- Can repeatability be improved by testing subjects on a "reference" interface in addition to the product under test? This approach would lessen distortions caused by non-representative sampling.
- Does the test procedure stifle innovation? Thermostats are undergoing rapid changes in both technologies and requirements. For example, can this test accommodate voice commands or visual cues?

The U.S. Energy Star program intends to incorporate a minimum usability requirement based on this method in its next specification for Climate Control Devices (e.g., programmable thermostats). Many of these questions will be resolved in the process of finalizing the technical specification.

6. Conclusions

Residential thermostats control a significant fraction of energy use in North America and most developed countries. It is therefore important to ensure that the occupants correctly operate thermostats so as to attain thermal comfort with the greatest possible energy savings. Earlier studies, and considerable anecdotal evidence, suggested that users of encountered difficulty in correctly operating modern, programmable, thermostats. Our investigations confirm that major usability problems exist with residential thermostats. We undertook five separate investigations, each revealing different aspects of the usability problem.

All of these investigations involved small, targeted, groups so none can be considered truly representative of the entire population. Nevertheless, they all point to the same conclusion, that is,

many users have difficulty operating programmable thermostats and exploiting their energy-saving features. The laboratory investigation establishes a procedure to quantify the difficulty and measure progress in fixing usability problems. We report the results of these preliminary investigations because they are also proof of concepts for several promising avenues of research, notably the use of the Mechanical Turk crowd-sourcing survey approach and the test procedures to quantify usability. Based on these successes, we plan to greatly expand the scope of our research in these areas. The technique outlined to quantify usability appears to have applicability to other energy-consuming devices and their controls such as lighting controls, televisions, and home energy management systems.

Acknowledgments

The authors thank Gari Kloss for her enthusiastic assistance in all aspects of this research. This work was supported by the Office of Energy Efficiency and Renewable Energy, Building Technologies Program, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

[1] U.S. Department of Energy (DOE). 2008 Residential Energy End-Use Splits by Fuel Type (Quadrillion BTU). [Online]. Table 2.1.5 April 29, 2011. Available: http://buildingsdatabook.eren.doe.gov/ChapterIntro2.aspx?2#12011.
[2] Cross D., Judd D., Automatic setback thermostats: measure persistence and customer behavior. In: Proceedings of the international energy program evaluation conference; 1997.
[3] Haiad C, Peterson J, Reeves P, Hirsch J. Programmable thermostats installed into residential buildings: predicting energy savings using occupant behavior & simulation. Rosemead: Southern California Edison; 2004.
[4] Nevius M. and Pigg S., Programmable thermostats that go berserk: taking a social perspective on space heating in Wisconsin. In: Proceedings of the 2000 ACEEE summer study on energy efficiency in buildings; 2000. pp. 8.233–8.244.
[5] Shipworth M, Firth SK, Gentry MI, Wright AJ, Shipworth DT, Lomas KJ. Central heating thermostat settings and timing: building demographics. Build Res Inform Apr 2010;38(1):50–69.
[6] Sachs H. Programmable thermostats. Washington, D.C: American Council for an Energy Efficient Economy; 2004.
[7] Bouchelle M. P., Parker D. S., Anello M. T., Factors influencing space heat and heat pump efficiency from a large-scale residential monitoring study. In: Proceedings of the 2000 ACEEE summer study on energy efficiency in buildings, 2000.
[8] Meier A, Aragon C, Peffer T, Pritoni M. Thermostat interface and usability: a survey. Berkeley (California): Lawrence Berkeley National Laboratory; 2010.
[9] Boait PJ, Rylatt RM. A method for fully automatic operation of domestic heating. Energ Build 2010;42(1):11–6.
[10] Consumer Reports. Thermostats - Some make saving easier [Online]. April 29, 2011 Available: http://www.consumerreports.org/cro/appliances/heating-cooling-and-air/thermostats/thermostats-10-07/overview/therm-ov.htm; 2007.
[11] Critchleya R, Gilbertsona J, Grimsleya M, Greena G, Group WFS. Living in cold homes after heating improvements: evidence from warm-front, England's home energy efficiency scheme. Appl Energ 2007;84(2):147–58.
[12] Diamond R. In: Comfort and control: energy and housing for the elderly, 15. Environmental Design Research Association; 1984.
[13] Diamond R., Energy use among the low-income elderly: a closer look. Proceedings of the 1984 ACEEE summer study on energy efficiency in buildings; 1984.
[14] Freudenthal A, Mook HJ. The evaluation of an innovative intelligent thermostat interface: universal usability and age differences. Cognit Tech Work 2003;5(1):55–66.
[15] Fujii H, Lutzenhiser L. Japanese residential air-conditioning: natural cooling and intelligent systems. Energ Build 1992;18(3):221–33.
[16] Karjalainen S. The characteristic of usable room temperature control. Helsinki University of Technology; 2008.
[17] Linden A, Carlsson-Kanyama A, Eriksson B. Efficient and inefficient aspects of residential energy behaviour: what are the policy instruments for change? Energ Pol 2006;34(14):1918–27.
[18] Moore TG, Dartnall A. Human factors of a microelectronic product: the central heating timer/programmer. Appli Ergon 1982;13(1):15–23.
[19] Rathouse K, Young B. RPDH15: use of domestic heating controls. UK: Watford: Building Research Establishment; 2004.
[20] Dale HC, Crawshaw CM. Ergonomic aspects of heating controls. Build Serv Eng Tech 1983;4(1):22–5.
[21] Rathouse K, Young B. Market transformation programme - Domestic heating: use of controls [Online]. April 29, 2011 Available: http://efficient-products.defra.gov.uk/cms/library-publications/; 2004.
[22] Lutzenhiser L. A question of control: alternative patterns of room air-conditioner use. Energ Build 1992;18:192–200.
[23] Parker DS, Hoak D, Cummings J. Pilot evaluation of energy savings from residential energy demand feedback devices FSEC-CR-1742-08. Cocoa: Florida Solar Energy Center; 2008. p. 13.
[24] Kempton W. Two theories of home heat control. Cognit Sci 1986;10(1):75–90.
[25] Pritoni M, Meier A, Peffer T, Perry D, Aragon C. How Turkers use thermostats, in preparation.
[26] Karjalainen S. Thermal comfort and use of thermostats in Finnish homes and offices. Build Environ 2009;44(6):1237–45.
[27] Stiftung Warentest. Heizkörperthermostate - Ausgewählt, geprüft, bewertet. Test May, 2008. [Online]. April 29, 2011 Available: http://www.test.de/themen/umwelt-energie/test/Heizkoerperthermostate-Auf-Sparen-programmiert-1672635-1672639/.
[28] Råd och Rön. Test: Värmestyrsystem. Råd och Rön 2003;2:6.
[29] Amazon. Amazon Mechanical Turk - Welcome [Online]. April 29, 2011 Available: https://www.mturk.com/mturk/welcome; 13-Nov-2010.
[30] Paolacci Gabriele, Chandler Jesse, Ipeirotis Panagiotis G. Running experiments on Amazon mechanical turk. Judgment and Decision Making Aug 2010;5(5):411–9.
[31] ISO. Guidance on usability. Geneva: International Standardization Organization; 1998.