Measuring the Usability of Appliance Controls

Alan Meier*, Cecilia Aragon*, Daniel Perry**, Therese Peffer, Marco Pritoni****, Jessica Granderson****

**** Lawrence Berkeley National Laboratory, ** University of California, Berkeley, *** California Institute for Energy and Environment, **** University of California, Davis***

Abstract

A “smart” device will remain efficient only as long as the settings and other parameters allow it to be. Thus, the degree of usability is becoming an element of energy efficiency similar to other physical characteristics. We developed and tested a procedure to quantify the usability of thermostats. The procedure assumes that usability can be represented by a user’s ability to accomplish a set of tasks. Thirty one subjects were tested in their ability to accomplish six essential tasks on programmable thermostats. The tests revealed a wide variation in the subjects’ ability to accomplish the same task on different thermostats. Thus it was possible to discern thermostats that were more effective than others. We created a metric based on data that are easy to collect and unambiguous that appears to reflect the usability of a task. Metrics from different tasks can be added and an overall usability “score” calculated. This approach, as well as the metric, can be applied to other devices where poor usability may impede energy-saving behaviour.

Background

The controls of energy-using products are becoming increasingly sophisticated in order to provide both more features and increased energy efficiency. Most products covered by minimum energy performance standards (MEPS) now incorporate a microprocessor, a display, user input devices (e.g., keypads), sensors and other means of information input and output. The microprocessor takes these inputs, makes decisions, and determines the operating mode. Table 1 shown as Table 2 lists some modes for refrigerators, televisions, and heat pumps. Each mode results in a different level of energy consumption.

Table 1. Potential modes in three appliances.

Device	Modes (partial list)
Refrigerators	Compressor on (variable) Defrost on Fan on (variable) Ice making on Anti-sweat heater on External display (on/off/sleep) Data send/receive Microprocessor on
Televisions	Display on Brightness (variable) Sound level Timer Motion sensor control Resolution (variable)

	Automatic programming guide Fan (variable) Standby functions Screen saver
Heat pumps	Compressor on (variable) Fan (variable) Defrost heater on Crankcase heater on Display on Remote control active Timer active Off, but processor on

The enhanced controls can also lead to user confusion. A naïve user may inadvertently select settings resulting in higher energy consumption than necessary because the device's user interface employs:

- unfamiliar or inconsistent terms or symbols
- awkward procedures to change settings
- opaque procedures to make changes
- ergonomically difficult features

These problems appear in many types of appliances and energy-using equipment. The problem of inconsistent terms and symbols has been described in office equipment by Nordman [1]. However, similar cases are common in appliances, consumer electronics, and lighting controls, such as:

- Inconsistent symbols: in thermostats controlling heat pumps, the status light indicating operation of (high-cost) resistance heating use may be red or green (depending on model). Manufacturers also use at least three different terms for it.
- Awkward procedures: a thermostat requires over 10 keystrokes to lower temperature prior to leaving the building
- Opaque procedures: motion sensor is activated by rapidly flicking the light switch 4 times to enable or disable it.

When confronted with these situations, users—even those with the greenest intentions--will often select settings that are more convenient over energy saving. Many of these problems are related to the usability of the device, where usability is defined by ISO as "The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"[2].

Different approaches can be taken to improve usability and minimize user confusion. One approach is to harmonize the critical terms and symbols associated with the interface. In this way, users are more likely to be familiar with basic controls even when they confront an unfamiliar device. Nordman [3] describes how certain symbols were standardized for power controls in office equipment .

One obstacle to improving usability is the absence of recognized procedures to measure and quantify usability. Manufacturers have no way to compare prototype interfaces, consumer organizations have no means to rate interfaces, and governments have no metric for establishing minimum levels of usability. We describe below a methodology to quantify the usability of programmable thermostats and results of laboratory tests of five thermostats. This methodology appears suitable for measuring usability in other devices relying on complex controls.

Evaluating usability of products is commonplace; however, most evaluations address usability of one-off items, such as controls in airline cockpits or websites. The typical procedure is to compare one version against an improved version. To our knowledge, this is the first quantitative usability test developed for mass-produced products.

Measuring the Usability of Thermostats

In North American homes, a single thermostat typically controls the heating and cooling equipment. In the last fifteen years, digital, programmable, thermostats have been introduced. These thermostats allow the occupants to set a schedule for heating and cooling and save energy compared to constant temperatures. A schedule involving nighttime temperature setbacks (or set-ups in the cooling season) has been shown to reduce heating use 15% compared to constant temperature settings [4]. About 40% of U.S. homes now have programmable thermostats. Almost all new single-family homes have programmable thermostats and, in some regions, building codes require all new homes to be equipped with them.

Nevertheless, several evaluations in different parts of the United States found that homes with programmable thermostats actually used *more* energy than those operated with manual thermostats [5]. This is no surprise because a broad array of studies have found that the advanced features of programmable thermostats—and especially the programming capabilities—are rarely used and often bypassed by the occupants [6]. A major explanation for failure to exploit the programmable thermostats is the difficulty in setting and changing them; in other words, these thermostats suffer from poor usability.

A Usability Score Based on Tasks

Some thermostats are easier to operate than others. But how can the superiority of one interface be measured? Ideally, the test method should resemble an energy test procedure, that is, be clearly defined, and have repeatable, quantifiable, results. These measurements of usability could then be used to establish a "usability score" which would allow manufacturers, consumers, and regulatory agencies to rank thermostats and establish minimum criteria for usability. We therefore investigated the feasibility of quantifying usability of thermostats. The procedure is based on controlled interactions between people and thermostats.

The measurement method involves two steps:

1. Define representative tasks to be accomplished with the thermostat;
2. Measure people's ability to perform those tasks under controlled conditions using defined metrics.

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 22°C on Tuesday evenings at 7 PM." Assembling tasks involves studying the operating manuals and carefully observing and interviewing users. It is also necessary to consider if the user is expected to interact with the thermostat as a total novice (such as when one enters a hotel room), daily, or somewhere in between. 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 eventually selected were:

Task 1: Turn the thermostat from "off" to "heat."

Task 2: Set the correct time.

Task 3: Identify the temperature the device is set to reach.

Task 4: Identify the temperature that the thermostat is set to reach for 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 features of the thermostat. The same approach could be applied to other sorts of controls, such as for lights or heat pump water heaters.

We sought to observe in detail and record different aspects of usability with which we might offer indicators of usability. The following aspects were collected for each subject during each test:

- 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
- verbal comments.

We recorded the sessions with a video camera; this way we were able to convert the data collected on a video record and determine the aspects listed above.

Our initial goal was to determine the viability of the task-based methodology and the identification of the best metric. Did the test procedure generate a significant range in the metrics? Did the test procedure applied to different thermostats generate a significant range in a metric? Finally, was one metric superior to others?

Details of Experiment

Five programmable thermostats were selected for testing. Three were primarily controlled through a touchscreen and one was a web-based interface. The tests were conducted at a usability laboratory. The laboratory set-up was very simple (see Figure 1). A video camera recorded each test in the vicinity around the thermostat (so the subject's face was not captured). The camera captured images similar to that shown in Figure 2. Thirty-one participants were recruited (22 male, 9 female), with ages ranging from 18 – 65. The subjects had many different occupations and varying levels of previous experience with programmable thermostats. Each subject was tested on two thermostats. Each test consisted of six tasks. Altogether 62 tests were performed, consisting of 372 tasks.

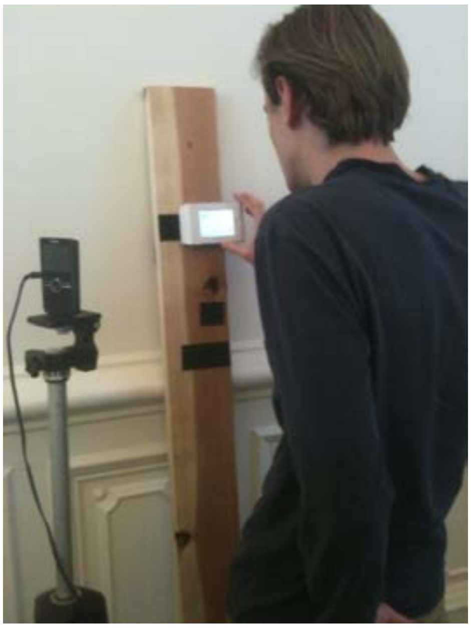

Figure 1. Laboratory set-up for measuring the usability of a thermostat.

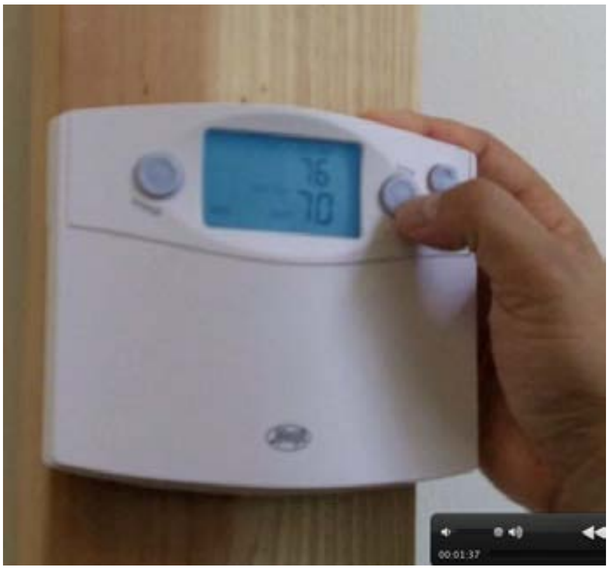

Figure 2. Still image from a video of a person performing Task 1.

Results: Metrics of Usability

A wide range of usability was observed. Figure 3 shows the pooled results when the subjects were asked to perform Task 1 (“turn the thermostat from off to heat”) on the five thermostats. Each subject performed Task 1 on two different thermostats. The metric was elapsed time to complete the task.

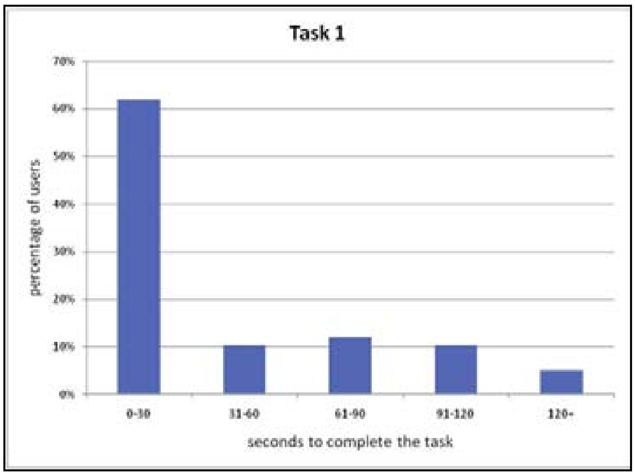

Figure 3. Distribution of times subjects required to complete Task 1 (excluding those who were unable to complete).

Most subjects were able to accomplish the task in less than 30 seconds; however, over 30% of the subjects required 31 – 120 seconds. (Note that two minutes can feel like a very long time when trying to switch on the heat.) About 26% of the subjects were unable to accomplish the task at all and are not displayed in this Figure. The results shown in Figure 3 (and other results not shown here) demonstrated that the methodology produced a wide range of measured abilities of the subjects to perform the task.

A second requirement of the task-based methodology is the ability to quantitatively differentiate levels of usability among thermostat interfaces. Figure 4 displays the range in elapsed time to completion for accomplishing Task 1 with the five thermostats. The times for *not* completion are shown in red; this is where the subject mistakenly believed that he or she had completed the task or gave up. The times are averages based on about 12 subjects tested on each thermostat. The Figure demonstrates that the task-based methodology and the metric permitted easy 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 were unable to complete it at all. Both Thermostats D and E had hinged covers concealing the controls, which many subjects either did not recognize or were unable to open. This illustrates how small design differences can have large impacts on successful operation of a device.

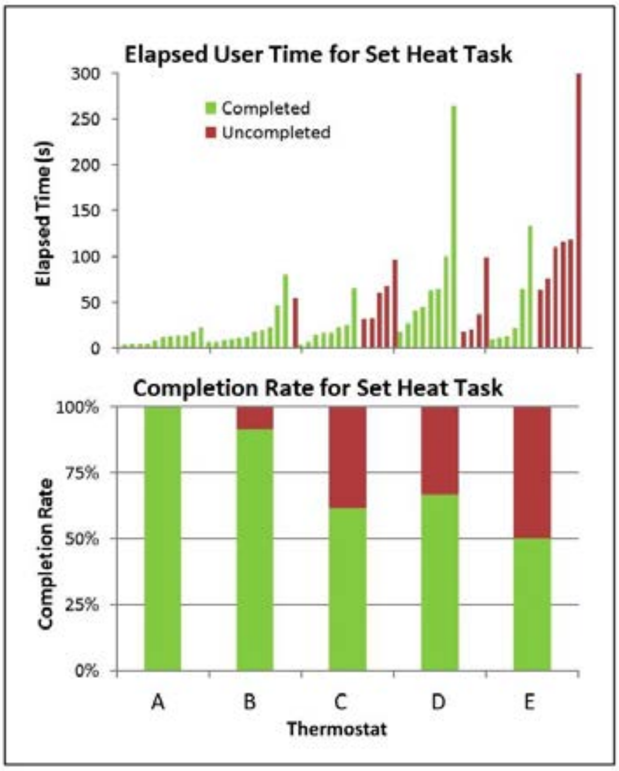

Figure 4. The subjects' completion times and completion rates for Task 1.

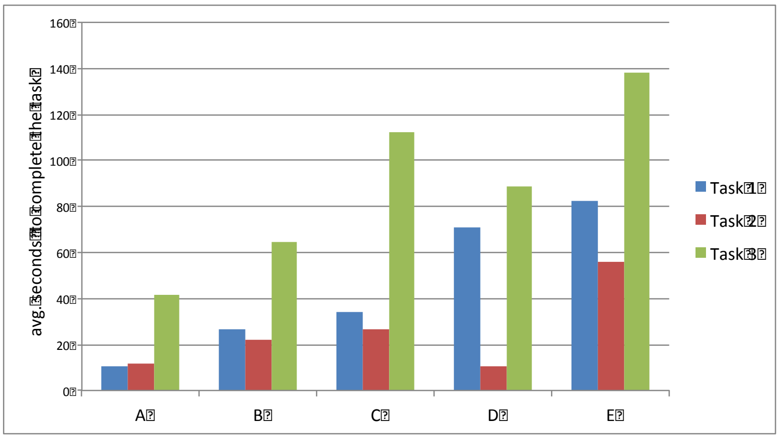

Figure 5. Average time to complete Tasks 1, 2, and 3 for the five thermostats.

The results were similar for other tasks. Figure 5 shows the average elapsed times for Tasks 1, 2, and 3. 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.

The average elapsed time for completion is an attractive metric, with robust results; however, it may be misleading when many subjects fail to complete the task because those measurements must be omitted in order to avoid a nonsensical solution (e.g., infinity). We therefore developed a hybrid metric, combining both elapsed time to complete and successful completion of the task. We also wanted the metric to be bounded, that is, from 0 to 1 so that results from different tasks were more easily comparable. These features make the metric simpler to interpret. The metric, “Time and Success Metric” is based on a logistic function to capture the features described above.

The time and success metric, “*M*”, is calculated as follows on a per-trial basis:

$$M_i = \frac{2s}{1 + e^{x_i}}$$

where

$$x_i = \text{distinguishing variable for each metric}$$

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

Note that M_i 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 metric combines time on task with success of the trial in an intuitive manner: if the task is not completed so that *s* = 0, the value of the metric 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 metric. 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.

The results for the three tasks combined, using the time and success metric (and k_1 = 50) are shown in Figure 6. The Figure displays mean values, along with error bars at the 95% confidence level.

Both of the concepts, time to completion and success to complete, are intuitively easy to understand.

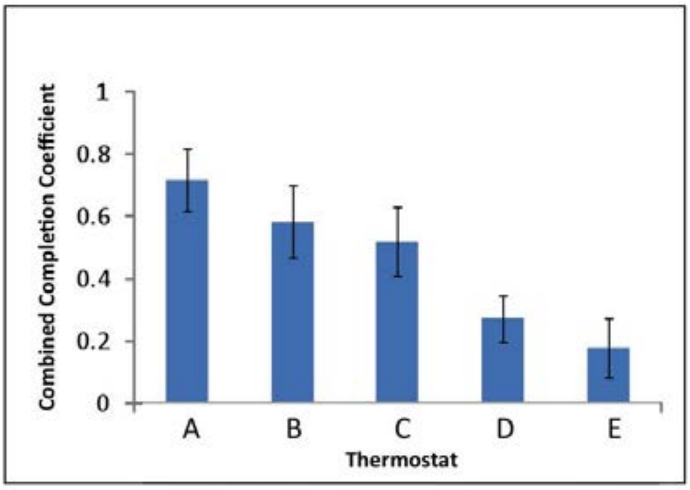

Figure 6. Time and success metrics for the five thermostats based on Tasks 1, 2 and 3. NOTE: figure will be redrawn with correct vertical axis.

Furthermore, they are easy to measure in a laboratory with relatively simple equipment. These features make the task and success metric an attractive metric for quantifying usability.

Discussion

These results suggest that it is possible to quantitatively evaluate the usability of thermostats. These results also suggest that a usability score, based on a combination of tasks, will be a meaningful indicator of overall usability. The results are promising but further research is still needed gain greater confidence in the approach. Some topics for further research include:

- 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, we need large enough test panels to attain satisfactory confidence in the results. Policymakers need to decide to what extent elderly, handicapped, colour-blind, and non-English speakers are included.
- Repeatability is a key requirement for any test procedure. We have not yet confirmed that the test results can be duplicated in other laboratories.
- Can repeatability be improved by testing subjects on a “reference” interface in addition to the product under test? A reference interface would make it possible to calibrate the panel of subjects and potentially 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?

Energy Star is addressing many of these issues [7] because it intends to include a usability criterion in its next specification for programmable thermostats (which it calls “climate control devices”). To our knowledge, this Energy Star specification is the first application of a quantitative usability requirement for the controls of a device.

Application of this Approach to Other Products

This approach, as well as the time and success metric, can be applied to other devices where poor usability may impede energy-saving behaviour. Modern lighting controls–especially those with several features–in commercial buildings suffer from usability problems. Occupants are often frustrated and unable to easily obtain the desired illumination conditions. It is easy to construct a list of representative tasks for lighting controls. This list would include:

- Determine status of lights
- Switch light on
- Switch light off
- Identify if light has dimming capability
- Dim light to about 50%
- Determine if light is connected to a sensor

These tasks might seem trivial yet Figure 7 illustrates the diversity of controls (and the complexity of actions needed to accomplish the tasks) that a user will confront.

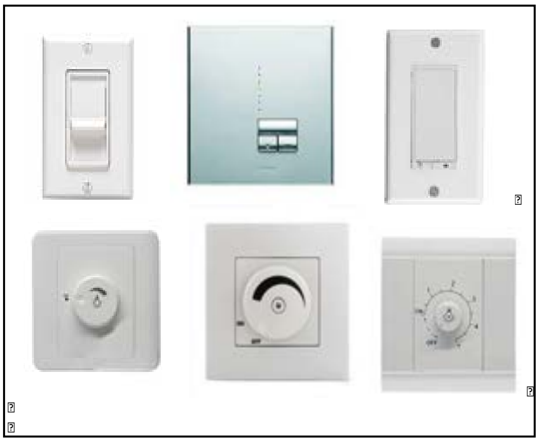

Figure 7. Six lighting controls found in commercial buildings.

For Task 3, dim lights to 50%, the procedure is different for almost every control (and not always obvious). The time and number of actions varies from a single rotating action to multiple button pushes. One must also take into consideration that there will be more first-time users than with residential thermostats.

Heat pump water heaters require sophisticated controls so as to ensure maximum efficiency while meeting hot water needs. Figure 8 shows the controls for three commercially-available heat pump water heaters. Incorrect settings of these controls can lead to significantly higher energy consumption without the consumer being aware. The likelihood of incorrect settings is high because controls are confusing and occupants are not familiar with this new device. On the other hand, the use situation is different from thermostats because users are likely to select their preferences once and leave them for long periods. This may encourage users to devote more time to initial settings. (Field research needs to verify actual operating patterns.)

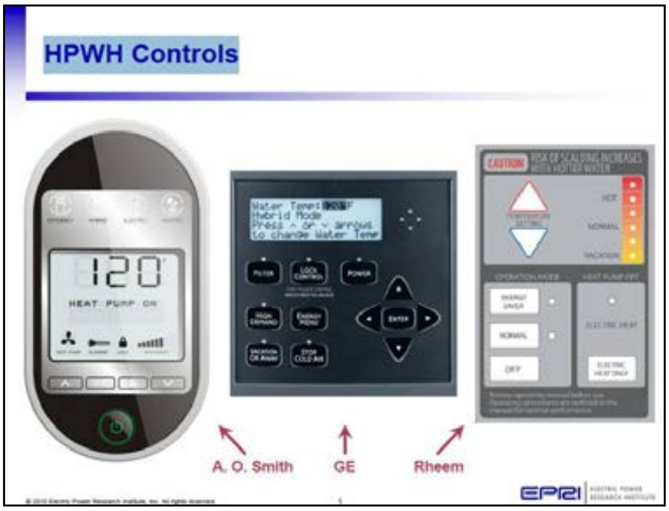
HPWH Controls

Figure 8. Controls for three residential heat pump water heaters (source EPRI).

Designing clear user interfaces for “smart” products will become a critically important requirement for ensuring energy-efficient operation.

Conclusions

The controls of energy-using products are becoming increasingly sophisticated in order to provide both more features and increased energy efficiency. Ironically, as the devices become “smarter”, the quality of the interface between the device and user rises in importance. A “smart” device will remain efficient only as long as the settings and other parameters allow it to be. Thus, the degree of “usability” is becoming an element of energy efficiency similar to other physical characteristics like insulation. To date, however, there has been no way to measure usability.

The digital programmable thermostat relies on user input to set operating parameters. Many users of programmable thermostats have been frustrated by the controls and, in some cases, have been unable to accomplish basic tasks necessary to effectively operate the devices. The results are thermostat settings that potentially lead to higher than necessary energy use, often without the knowledge of the user.

We developed and tested a procedure to quantify the usability of thermostats. The measurement of usability is based on the assumption that the essence of usability can be captured by a collection of representative tasks. We demonstrated that a relatively simple laboratory set-up and test procedure could collect adequate data for assessment. A range in human abilities in accomplishing a task was easily discerned. The same tests also revealed wide variation in the subjects’ ability to accomplish the same task on different thermostats. Thus it was possible to discern thermostats that were more effective than others.

We created the “time and success metric”, which appears to reflect the usability of a task. The data required to calculate the time and success metric are easy to collect and reasonably unambiguous. A second feature of the time and success metric is that metrics from different tasks can be combined through addition and an overall usability “score” calculated.

Many of the usability problems identified in thermostats appear in other products. We showed two examples, controls for lighting and heat pump water heaters. Other products, such as televisions, also deserve attention. Further research will still be needed to refine the approach and the metric; however, we believe that they are already suitable for quantitatively evaluating the usability of products. Manufacturers can use this procedure as a design tool and regulators can establish minimum usabilities for appropriate products.

References

[1] B. Nordman, A. Meier, and D. Auman, “Fixing the Power Management Controls Problem,” in *ACEEE Summer Study on Energy Efficiency in Buildings*, 2002.

[2] ISO, *Guidance on Usability*. Geneva: International Standardization Organization, 1998.

[3] B. Nordman, *Power Control User Interface Standard — Final Report*. Berkeley (Calif.): Lawrence Berkeley National Laboratory, 2002.

[4] A. Meier, C. Aragon, T. Peffer, and M. Pritoni, *Thermostat Interface and Usability: A Survey*. Berkeley (Calif.): Lawrence Berkeley National Laboratory, 2010.

[5] H. Sachs, *Programmable Thermostats*. Washington, D.C.: American Council for an Energy Efficient Economy, 2004.

[6] A. Meier, M. Pritoni, C. Aragon, D. Perry, and T. Peffer, “How People Actually Use Thermostats,” in *ACEEE Summer Study on Energy Efficiency in Buildings*, 2010.

[7] U.S. Environmental Protection Agency, “ENERGY STAR® Residential Climate Controls Draft Specification Framework –Performance-Based Usability Requirements,” 30-Nov-2010.