MEASURING THE USER EXPERIENCE

Collecting, Analyzing, and Presenting Usability Metrics

TOM TULLIS
BILL ALBERT

MORGAN KAUFMANN

CHAPTER

Behavioral and Physiological Metrics

7

During a typical usability test, most participants do much more than complete tasks and fill out questionnaires. They may laugh, groan, shout, grimace, smile, fidget in their chair, look aimlessly around the room, or drum their fingers on the table. These are all behaviors that are potentially measurable and offer insights into the usability of the product being tested. Most of this body language and verbalization can be observed and noted by an attentive test administrator, but some types of subtle or fleeting behavior are harder to observe. For example, facial expressions can change very rapidly, making a good-quality video recording of the participant's face very useful. And there are still other behaviors that most people aren't even conscious of, such as increased heart rate, pupil dilation, and slight increases in sweating, which require specialized equipment to monitor. All of these behaviors, those that are directly observable and those that require special instruments, are the subject of this chapter.

UNPLANNED RESPONSES IN THE USABILITY LAB

We once had a participant in our usability lab who was eight months pregnant. She was testing a prototype of a particularly challenging internal application. At one point during the session she said the interface was so bad that it was giving her labor pains! Thankfully, she admitted she was joking just before we called the paramedics.

7.1 OBSERVING AND CODING OVERT BEHAVIORS

There are probably as many different approaches to taking notes during a usability test as there are evaluators conducting usability tests. Each approach tends to be highly personalized. Some like a free-form "stream-of-consciousness" approach where they narrate the events of the session, others like to use forms where they note specific events and behaviors, and still others like to use sophisticated data-logging tools that

automatically time-stamp all their entries. All of these approaches are useful, depending on the purposes of the usability study, but to be useful as a metric, some type of structure must be applied to these observations. Although it's possible to apply some structure to free-form notes *after* a test session, it's more effective to identify some degree of structure *before* the test, while also allowing for the indication of behaviors that don't fit within the structure.

A participant's overt behaviors in a usability session can be divided into two general categories: verbal and nonverbal. Verbal behaviors include anything the participant actually *says*. Nonverbal behaviors include a range of other things that the participant might *do*. Both can be helpful in identifying parts of an interface or product that cause problems for users or, conversely, that delight users.

7.1.1 Verbal Behaviors

Verbal behaviors provide valuable insight into a participant's emotional and mental state while they are using a product. The participant will probably make many comments, some negative ("This is hard" or "I don't like this design") and some positive ("Wow, this is much easier than I expected" or "I really like the way this looks"). Some comments are neutral or just hard to interpret, such as "This is interesting."

The most meaningful metric related to verbal behaviors is the ratio of positive to negative comments. To do this type of analysis, you first need to catalog all verbal behaviors and then categorize each one as positive, negative, or neutral. Once this is complete, simply look at the ratio of positive to negative comments, as illustrated in Figure 7.1. Only knowing that positive behaviors outnumbered negative comments by a 2:1 ratio does not say a lot by itself. However, it's much more meaningful if the ratios are compared across different design iterations or between different products. For example, if the ratio of positive to negative comments has increased significantly with each new design iteration, this would be an indicator of an improved design.

It's also possible to get more granular by differentiating among different types of verbal behaviors, such as the following:

- Strongly positive comments (e.g., "This is terrific!")
- Other positive comments (e.g., "That was pretty good.")
- Strongly negative comments (e.g., "This website is terrible!")
- Other negative comments (e.g., "I don't much like the way that worked.")
- Suggestions for improvement (e.g., "It would have been better if . . .")
- Questions (e.g., "How does this work?")
- Variation from expectation (e.g., "This isn't what I was expecting to get.")
- Stated confusion or lack of understanding (e.g., "This page doesn't make any sense.")
- Stated frustration (e.g., "At this point I'd just shut it off!")

These types of data are analyzed by examining the frequency of comments within each category. Like the previous example, comparing across design iterations or

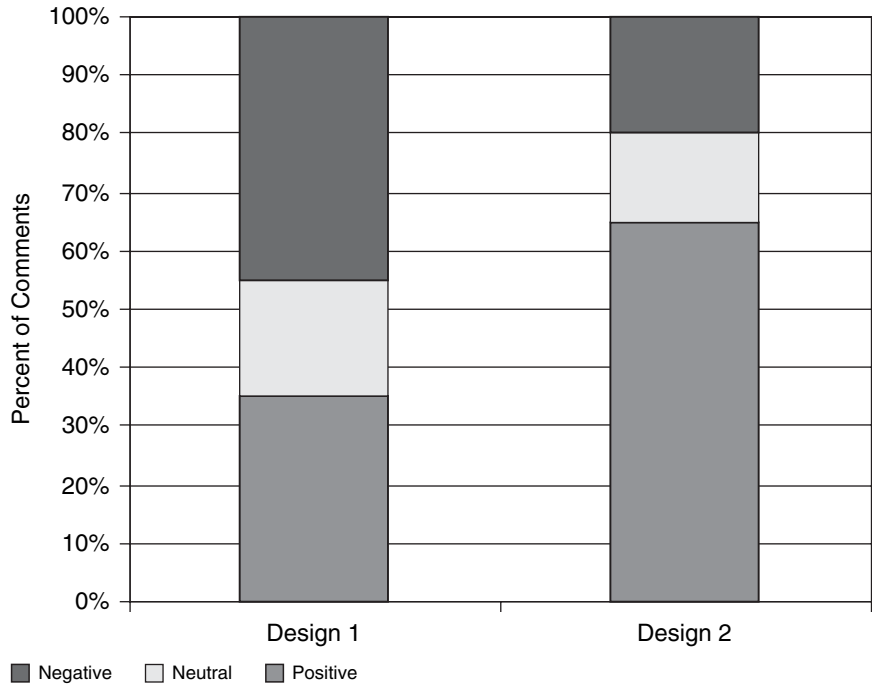


FIGURE 7.1

Example of coding the percentage of neutral, positive, and negative comments for two different designs.

products is the most useful. Categorizing verbal behaviors beyond just the positive, negative, or neutral can be somewhat challenging. It's helpful to work with another usability specialist to get to some level of agreement about categorizing each verbal behavior. Make good use of video recording. Even the best note takers can miss something important. Forms may be very helpful in documenting both verbal and nonverbal behaviors. Figure 7.2 shows an example of a form that could be used for coding observations and other data during a usability test.

7.1.2 Nonverbal Behaviors

Nonverbal behaviors can be very revealing about a participant's experience with a product. These might include facial expressions (frowning, smiling, looks of surprise, furrowing brow) or body language (fidgeting, leaning close to the screen, rubbing the head). Deriving any meaningful metrics from these nonverbal behaviors is somewhat challenging and most useful only for certain types of products. If you're evaluating websites, software, or other products that have very few physical demands, these metrics may have limited utility. But there are cases where some

Usability Test Observation Coding Form

Date: _____ Participant #: _____ Task #: _____

Start Time: _____ End Time: _____

Verbal Behaviors	**Notes**
☐ ☐ ☐ Strongly positive comment	
☐ ☐ ☐ Other positive comment	
☐ ☐ ☐ Strongly negative comment	
☐ ☐ ☐ Other negative comment	
☐ ☐ ☐ Suggestion for improvement	
☐ ☐ ☐ Question	
☐ ☐ ☐ Variation from expectation	
☐ ☐ ☐ Stated confusion	
☐ ☐ ☐ Stated frustration	
Other:	

Nonverbal Behaviors	**Notes**
☐ ☐ ☐ Frowning/Grimacing/Unhappy	
☐ ☐ ☐ Smiling/Laughing/Happy	
☐ ☐ ☐ Surprised/Unexpected	
☐ ☐ ☐ Furrowed brow/Concentration	
☐ ☐ ☐ Evidence of impatience	
☐ ☐ ☐ Leaning in close to screen	
☐ ☐ ☐ Variation from expectation	
☐ ☐ ☐ Fidgeting in chair	
☐ ☐ ☐ Random mouse movement	
☐ ☐ ☐ Groaning/Deep sigh	
☐ ☐ ☐ Rubbing head/eyes/neck	
Other:	

Task Completion Status Notes:

Incomplete:	Complete:
☐ Participant gave up	☐ Fully complete
☐ Task "called" by moderator	☐ Complete with assistance
☐ Thought complete, but not	☐ Partial completion

FIGURE 7.2

Form for coding observations during a usability test. This form is designed to be completed for each task that the participant attempts. It helps to have multiple observers complete the form independently and then reach a consensus later.

of these nonverbal behaviors may be indicative of frustration or impatience on the part of the user that may be very important. A number of years ago we were conducting a usability test of a web-based application that had unusually long response times to certain user requests. Although some participants in the test explicitly verbalized their frustration with the long delays, others did things like drum their fingers on the table, groan, or rub their head while waiting for the response.

Nonverbal behaviors may be particularly useful if the product has some physical, perceptual, or cognitive demands. For example, the setup procedure of a popular MP3 player requires the user to enter a serial number from the device when installing the software. The serial number is quite small and printed in white characters on a metallic surface, making it very difficult to read. Nonverbal behaviors such as squinting, turning the device, moving it into better light, or getting a teenager to read the number are good indicators that this task presents some difficulty for users. In this situation, it may be helpful to analyze the frequency of these nonverbal behaviors. As with verbal behaviors, it's also helpful to see how nonverbal behaviors change across different design iterations or when comparing various products.

7.2 BEHAVIORS REQUIRING EQUIPMENT TO CAPTURE

Whereas the previous section focused on overt behaviors that a skilled observer can reasonably detect in real time during a usability test, we're now going to turn to a finer-grained analysis that typically requires equipment to capture. These include a detailed analysis of facial expressions, eye-tracking, pupil diameter, skin conductance, and others.

7.2.1 Facial Expressions

Recognizing and interpreting facial expressions is a key part of human-to-human communication. The lack of that visual channel is one of the reasons that we sometimes miss subtleties in telephone conversations or e-mail. Many psychologists and others argue that facial expressions are a more accurate window into what people are actually feeling than what they say. Perhaps a more detailed analysis of the facial expressions that the participants in a usability test make could give us useful insight into their reactions to the product and how it could be improved.

One technique is for a trained observer to do a detailed analysis of the facial expressions from a good-quality video recording. This could be done using some method of classifying specific facial expressions. In the 1970s Paul Ekman and Wallace Friesen (1975) developed a taxonomy for characterizing every conceivable facial expression. They called it the Facial Action Coding System (FACS), which

included 46 specific actions involving the facial muscles. More recently, they developed the FACS Affect Interpretation Dictionary (FACSAID) to specify the complex linkage between facial actions and emotional response (Ekman, Friesen, & Hager, 2002).

Of course, the main drawback of this kind of analysis of facial expressions is that it's very labor-intensive. That has prompted some researchers to investigate automated ways of measuring facial expressions. Two general approaches have been studied: video-based systems that attempt to recognize specific facial expressions and electromyogram (EMG) sensors that measure activity of specific muscles of the face.

EARLY ANALYSES OF FACIAL EXPRESSIONS

The scientific analysis of facial expressions dates back at least to Charles Darwin, who published *The Expression of the Emotions in Man and Animals* in 1872. He described in great detail a variety of facial expressions for such feelings as anxiety, grief, despair, joy, anger, contempt, disgust, guilt, astonishment, and horror. Darwin argued that these expressions cut across cultures, noting that "the young and the old of widely different races, both with man and animals, express the same state of mind by the same movements." Interestingly, Darwin's book was one of the first to include photos.

Video-Based Systems

Although video-based systems are far less intrusive than EMG systems, the analyses involved are computationally challenging, largely due to the varying appearances of people and their facial expressions. There have, however, been some successes, such as the work of Essa and Pentland (1997), who developed a system capable of recognizing six static facial expressions: anger, disgust, happiness, surprise, eyebrow raise, and neutral. More recently, den Uyl and van Kuilenburg (2005) tested a system called FaceReader against a reference database of 980 images of facial expressions. They found that the system accurately classified 89 percent of the expressions it was shown into one of six emotions: happy, angry, sad, surprised, scared, disgusted, and neutral.

Consider the images in Figure 7.3, which are taken from a video recording of a usability test participant. These images were captured while the participant was performing a particularly challenging and frustrating task. It's easy to see that there is information in her facial expressions. The key is figuring out how to capture and characterize the information in an efficient way.

Electromyogram Sensors

An alternative to video analysis is the use of electromyogram (EMG) sensors to measure facial expressions, as illustrated in Figure 7.4. These sensors measure electrical activity in certain muscles of the face. They are most commonly used to measure activity of two muscle groups: the corrugator muscle of the forehead, which is associated with frowning, and the zygomatic muscle of the cheeks, which is associated with smiling (Benedek & Hazlett, 2005). But can any of this actually be of value in a usability test?

FIGURE 7.3

Captured images from a video recording of a test participant performing a particularly challenging and frustrating task. One of the things we discovered in capturing these images is how fleeting facial expressions can be, with many lasting less than a second.

FIGURE 7.4

Facial EMG sensors on a participant in a usability study. The sensors on the forehead measure electrical activity of the corrugator muscle, which is associated with frowning. The sensors on the cheek measure activity of the zygomatic muscle, which is associated with smiling. *Source:* From Benedek and Hazlett (2005). Used with permission.

Richard Hazlett of the Johns Hopkins University School of Medicine reported a study in which he used EMG sensors on the corrugator muscles of 28 participants while they performed five tasks on one of two websites: Jones of New York and Tylenol (Hazlett, 2003). He calculated a continuous Frustration Index

for each participant from the corrugator EMG by comparison to a baseline established prior to the tasks. Hazlett demonstrated that this Frustration Index correlated well with more conventional measures of task and site difficulty. He found that the mean Frustration Index was significantly greater for tasks answered incorrectly than for those answered correctly. The Frustration Index, which is a continuous measure, also helped to identify specific pages in the sites that were the most problematic.

More recently, Hazlett teamed with Joey Benedek of Microsoft to assess the value of facial EMG in measuring reactions to software (Benedek & Hazlett, 2005). In one study, participants wearing facial EMG sensors were shown a demonstration of possible new features of a desktop operating system. After the demonstration, the participants were asked to list the features of the system that they particularly liked. The researchers calculated a ''Desirability Rating'' from the EMG data while participants viewed each of the features of the system. This Desirability Rating was based on cases where the zygomatic EMG was at least one standard deviation above the baseline (i.e., a positive response) and there was no accompanying corrugator EMG (i.e., a negative response). They found a good correlation between this Desirability Rating for each of the features and the participants' recall of the features they liked.

In a second study, Benedek and Hazlett measured EMG responses while participants performed nine tasks (e.g., burning a CD, making a playlist) using one of two versions of a media player. The researchers defined a measure of ''elevated tension'' as the number of seconds that the participant had a corrugator EMG value greater than one standard deviation above that participant's overall corrugator mean. They found that four of the tasks yielded very similar levels of this elevated tension measure for the two media players. But for the other five tasks, Media Player B showed significantly lower levels of elevated tension. They also looked at the correlation between time-on-task and elevated tension, and they found only a moderate correlation ($r = 0.55$). Elevated tension was associated with longer task time for six of the nine tasks, but the effect was reversed for the other three tasks. Their explanation for these three tasks was that the participants were simply enjoying themselves while taking longer to perform the tasks (which included tasks like ''Play all classical music''). This shows that time-on-task by itself may not always be a good indicator of task difficulty and that another measure, such as elevated tension based on facial expressions, can add value.

Measuring Facial Expressions in Everyday Usability Testing

Unfortunately, the analysis of facial expressions isn't quite ready for prime-time usability testing, unless you have access to EMG equipment and don't mind submitting your participants to it, or you're willing to devote the time and effort to detailed analyses of video recordings. Therefore, we recommend that until the technology becomes more readily available and less intrusive, you should use informal observation of facial expressions as a way to help you identify situations where you might want to probe users about their thoughts or reactions.

7.2.2 Eye-Tracking

Eye-tracking in usability testing has become significantly more common over the past few years. Thankfully, these systems have also become more reliable and easier to use. Although still not a mainstay of most usability labs, perhaps they will be in the not-too-distant future.

Although a few different technologies are used, many eye-tracking systems, such as the one shown in Figure 7.5, use a combination of an infrared video camera and infrared light sources to track where the participant is looking. The infrared light sources create reflections on the surface of the participant's eye (called the corneal reflection), and analysis routines compare the location of that reflection to the location of the participant's pupil. The location of the corneal reflection relative to the pupil changes as the participant moves his eyes. You must first calibrate the system by asking the participant to look at a series of known points; then the system can subsequently interpolate where he is looking based on the location of the corneal reflection.

Some eye-tracking systems use a head-mounted apparatus to allow for movement of the head, whereas other systems use either an optical or a magnetic system to track the participant's head and keep up with his eyes remotely. The latest eye-tracking systems are very unobtrusive, using optical tracking of the participant's eyes and allowing for accurate gaze tracking with minimal setup and calibration.

The information provided by an eye-tracking system can be remarkably useful in a usability test. Simply enabling observers of the test to see where the participant is

FIGURE 7.5

An eye-tracking monitor from Tobii Technology. Infrared light sources and an infrared video camera are built into the bezel of the monitor. The system tracks the participant's eyes automatically, with no need for a head-tracking apparatus.

looking in real time is extremely valuable. Even if you do no further analyses of the eye-tracking data, just this real time display provides insight that would not be possible otherwise. For example, assume a participant is performing a task on a website and there's a link on the homepage that would take him directly to the page required to complete the task. The participant keeps exploring the website, going down "dead-ends," returning to the homepage, but never reaching the required page.

In a situation like this, you would like to know whether the participant ever saw the appropriate link on the homepage or whether he saw the link but dismissed it as not what he wanted (e.g., because of its wording). Although you could subsequently ask participants that question, their memory may not be completely accurate. With an eye-tracking system you can tell whether the participant at least fixated on the link long enough to read it.

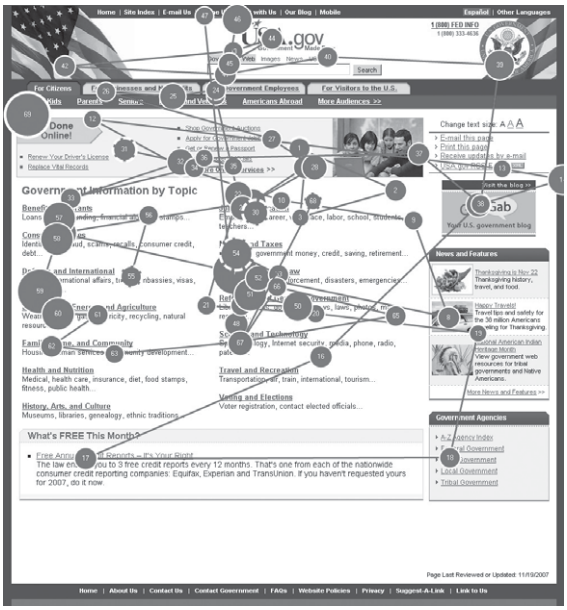
Figure 7.6(a) shows a plot of the series of fixations that an individual participant made on the USA.gov homepage. A fixation is defined by a pause in the eye's movement within a well-defined area. The fixations are numbered to indicate their sequence. The size of each circle is proportional to the length or duration of the fixation. The *saccades*, or movements between fixations, are shown by the lines. The series of fixations by multiple participants on the same page can be analyzed to create a "heat map," shown in Figure 7.6(b). In this visualization, the lightest areas represent greater density of fixations.

Proportion of Users Looking at a Specific Element or Region

One of the simplest analyses that can be done with an eye-tracking system is determining what percentage of participants in a usability test fixated on a specific element or region of interest. For example, we compared four different treatments for the same area on a web page. It was a small rectangular region of the page (a "bricklet") that was always in the same location and had the same content, but we varied the design of the area itself. Participants in the usability session performed four different tasks using the prototype. Only one of the tasks was directly related to the contents of this area.

Our goal was to see which designs resulted in more fixations on this bricklet. The results are shown in Figure 7.7. Analyzing this data is fairly straightforward, but you should keep a couple of things in mind:

1. Define the specific element of interest in terms of x, y coordinates on the page. Most eye-tracking analysis programs make this easy to do. These elements are usually called "areas of interest," "look-zones," or something similar.
2. Define a minimum total fixation time for the element of interest. For the data in Figure 7.7, we chose a minimum of 500 msec; we estimated that this would be the minimum time needed to get any useful information out of the element.



(a)

(b)

FIGURE 7.6

Gaze plot (a) showing a series of fixations one participant made while viewing the USA.gov homepage. The size of each circle corresponds to fixation's time. Heat map (b) reflecting density of fixations across a number of participants on different parts of the web page. The lightest areas indicate the most fixations.

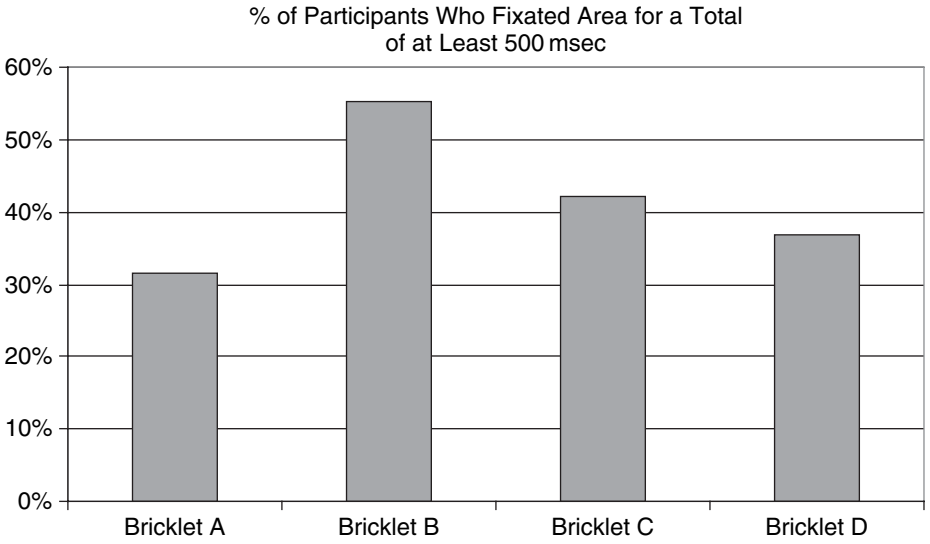


FIGURE 7.7

Percentage of participants who fixated for at least 500 msec total on each of the four bricklets. More of them looked at bricklet B than any of the others.

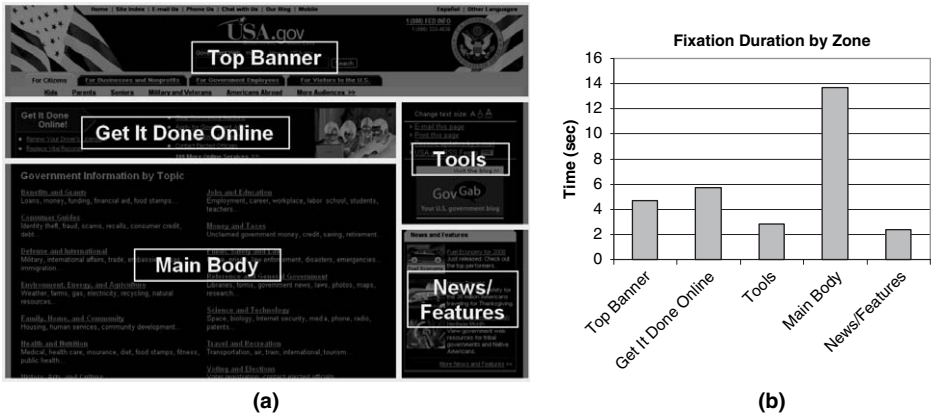


FIGURE 7.8

Fire regions defined on the USA.gov homepage (a), and the total fixation duration for each of those regions (b).

Time Spent Looking at a Specific Element or Region

Another way to analyze eye-tracking data is shown in Figure 7.8. In this example, regions of the page are defined (a), and the visualization represents the percentage of time that the participants spent looking in each of those regions (b).

Another way to use eye-tracking in usability studies is to compare the effectiveness of different locations on a web page for the same element. For example, Albert (2002) studied two locations—above or below the top branding area—for an ad on the results page of a web search engine. He found that participants spent about seven times longer looking at ads below the branding area than above it, as shown in Figure 7.9.

When analyzing the time spent looking at different regions, keep the following in mind:

- Clearly define each region. Do not leave any space undefined. Each region should be fairly homogeneous, such as navigation, content, ads, legal information, and so forth.
- Analyze the time as a percentage of total time spent on the page, not as an absolute amount of time, since the absolute amount of time can vary widely between participants.
- Only look at time data when the participant is engaged with the task. Do not include any time data when the participant is debriefing about her experience and still being tracked.
- When presenting data by look zones, the question about where participants actually looked within the zones typically comes up. Therefore, we recommend including a heat map, as in Figure 7.6, that shows the continuous distribution of fixations.

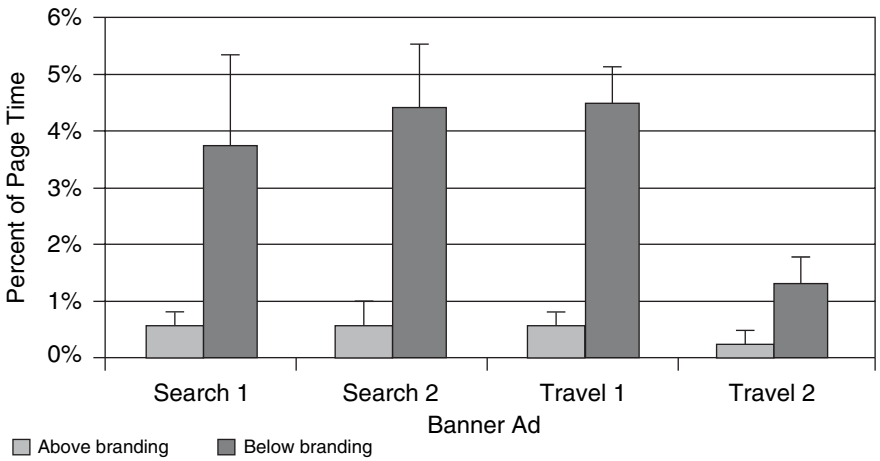


FIGURE 7.9

Data showing the percentage of total page viewing time that participants spent looking at ads. Four different ads (Search 1 & 2, Travel 1 & 2) and two different placements (above or below branding area) were tested. *Source:* Adapted from Albert (2002); used with permission.

Time to Notice a Specific Element

In some situations it's helpful to know how long it takes users to notice a particular element. For example, you may know that users spend only seven seconds on average on the page, but you want to make sure that a specific element, such as a "continue" or "sign up" button, is noticed within the first five seconds. It's helpful that most eye-tracking systems time-stamp each fixation (i.e., the exact time that each fixation occurred).

One way to analyze these data is to take an average of all the times at which the particular element was first fixated. The data should be treated as elapsed time, starting from the initial exposure. The average represents the amount of time taken to first notice the element, for all of those who *did* notice it. Of course, it's possible that some of the participants may not have noticed it all, let alone within the first five seconds. Therefore, you may come up with some misleading data showing an artificially quick time by not taking all the participants into account.

Perhaps a better way to analyze these data is to consider the proportion of participants who noticed the specific element within the specified time period. To do this, simply filter for those fixations that occurred within the specified time period. Then determine whether each participant had a fixation (or set of fixations) on the element during this window of time.

Scan Paths

Two other metrics from an eye-tracking system can be helpful in evaluating the effectiveness of an interface: length of eye movements and duration of fixations. For example, Fukuda and Bubb (2003) used these two metrics to compare the effectiveness of three different designs for web pages displaying subway timetables. They tested two groups of users: younger (17–29 years) and older (62–74 years). They found that two of the designs resulted in shorter eye movements than the other design for the first two tasks, but the opposite was true for the last two tasks. In general, a design that results in shorter eye movements can be considered a more efficient design because users have to move their eyes less to get the required information. They also found that the navigation elements in the designs that used smaller fonts (e.g., 10-point or smaller) tended to get longer fixations than those that used larger fonts. Longer fixation times generally indicate longer reading or processing times.

7.2.3 Pupillary Response

Closely related to the use of eye-tracking in usability studies is the use of information about the response of the pupil. Most eye-tracking systems must detect the location of the participant's pupil and calculate its diameter to determine where he or she is looking. Consequently, information about pupil diameter "comes for free" in most eye-tracking systems. The study of pupillary response, or the contractions and dilations of the pupil, is called pupillometry. Most people know that the pupil

PUPILLARY RESPONSE AND POKER

One of the most common "tells" in poker is a player's eyes, and especially his or her pupils. It's almost impossible for most players to keep their pupils from dilating when they get a hand that they're excited about. This is one reason most serious poker players wear sunglasses or a visor.

contracts and dilates in response to the level of ambient light, but many people don't know that it also responds to cognitive processing, arousal, and increased interest.

You can easily see this for yourself in a couple of ways. If you happen to have a cat, dangle an attractive target in front of it to encourage it to "pounce." As the cat prepares to pounce, watch how its eyes rapidly dilate. Although most of us don't pounce anymore, we have a similar pupillary response when focusing intensely on something. Another way to demonstrate this is to enlist a friend to do some mental arithmetic (e.g., 528 + 356) while you watch his pupils.

The psychological study of pupillary response began in earnest in the 1960s. One of the earliest studies to find evidence of pupil dilation in response to mental effort was by Hess and Polt (1964). It was further studied by Kahneman and Beatty (1966) and later described by Kahneman (1973) in his classic book *Attention and Effort*. But one of the challenges of using pupil dilation as a metric in usability studies is that it has been shown to correlate with a variety of different states of the user, including physical effort (Nunnally, Knott, Duchnowski, & Parker, 1967), mental effort (Hess & Polt, 1964), level of interest (Libby, Lacey, & Lacey, 1973), and emotional response (Steinhauer, Boller, Zubin, & Pearlman, 1983).

To confuse matters even more, there is also some evidence that in a memory overload condition, the pupil actually contracts (Granholm, Asarnow, Sarkin, & Dykes, 1996). In spite of these complications, some researchers have forged ahead to develop practical applications of pupillary response data. Sandra Marshall (2000) at San Diego State University developed an "Index of Cognitive Activity," or ICA, based on pupillary response. Her approach was sufficiently original that she was granted a U.S. patent for it.

Iqbal, Zheng, and Bailey (2004) studied pupillary response while participants performed easy or difficult versions of four different computer tasks: object manipulation (dragging and dropping e-mail messages), reading, mathematical reasoning, and searching for a product from a list of similar products. As shown in Figure 7.10, they found that when they focused on the cognitive components of the tasks (factoring out the motor components), the participants showed significantly larger pupil dilations in response to the difficult tasks than to the easier tasks.

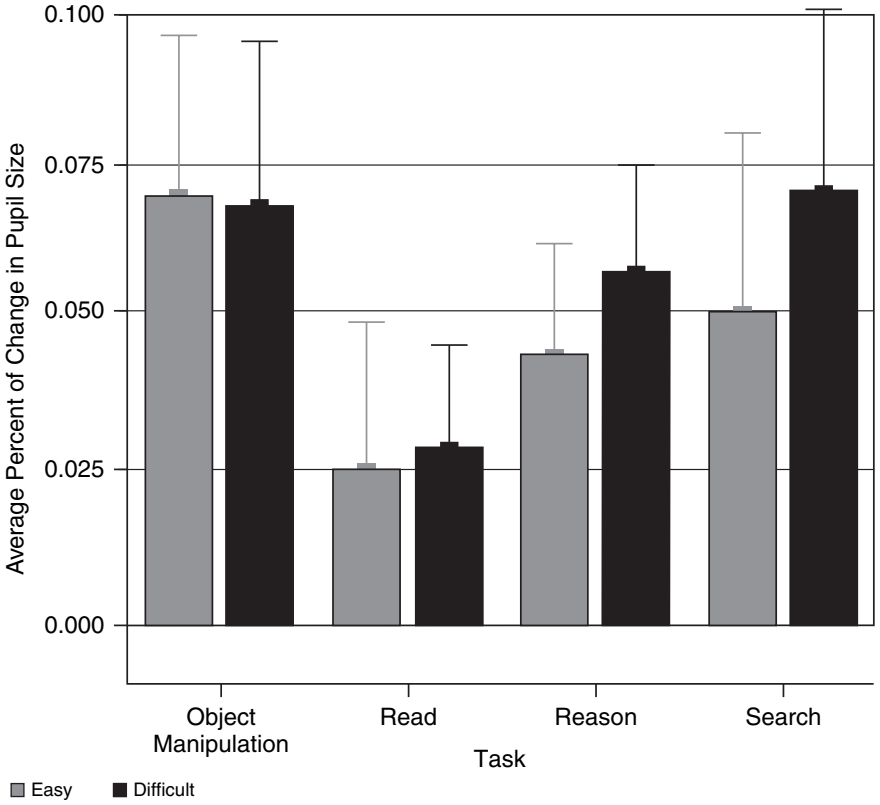


FIGURE 7.10

Data showing the change in pupil diameter (relative to baseline) for easy and difficult versions of four different tasks. Igbal et al. found that when they focused on the cognitive components of the tasks (Reason and Search), the more difficult versions yielded a significantly greater change in pupil diameter. *Source:* Adapted from Iqbal, Zheng, and Bailey (2004); used with permission.

Pupillometry in Everyday Usability Testing

Because pupil dilation is correlated with so many different mental and emotional states, it's difficult to say whether pupillary changes indicate successes or failures in everyday usability testing. However, measuring pupil diameter may be useful in certain situations where the focus is on the amount of mental concentration or emotional arousal. For example, if you are mainly interested in eliciting an emotional response to a new graphic on a website, then measuring changes in pupil diameter (from baseline) may be very useful. To do this, simply measure the percentage deviation away from a baseline for each participant and then average those deviations across the participants. Alternatively, you can measure the percentage of

participants who experienced dilated pupils (of a certain amount) while attending to a particular graphic or performing a specific function.

7.2.4 Skin Conductance and Heart Rate

Two physiological measures that have long been known to correlate with stress are skin conductivity and heart rate. Skin conductivity is typically measured using Galvanic Skin Response, or GSR, which measures electrical resistance. As we sweat even small amounts, the additional moisture in the skin increases its conductivity. Heart rate, of course, is also associated with stress: The heart beats faster under stress. Closely related is what's called Heart Rate Variability (HRV). As stress increases, HRV—the heart's ability to beat faster or slower in response to emotional or physical demands—tends to decrease. GSR, heart rate, and HRV are all used in various forms of biofeedback, where the participant uses feedback from devices monitoring these levels to learn to relax. In fact, a computer game called *The Journey to Wild Divine* includes devices for measuring GSR, heart rate, and HRV. In the game, the user explores a virtual world that includes soothing music and beautiful imagery. As a part of the exploration, various exercises are introduced to practice relaxation.

Several studies have sought to determine whether skin conductivity and heart rate could be used as indicators of stress or other adverse reactions in a usability setting. For example, Ward and Marsden (2003) used skin conductance and heart rate to measure user reactions to two versions of a website: a well-designed version and a poorly designed version. The poorly designed version included extensive use of drop-down lists on the homepage to ''hide'' much of the functionality, provided impoverished navigational cues, used gratuitous animation, and had occasional pop-up windows containing ads. The heart rate and skin conductance data were plotted as changes from the participant's baseline data established during the first minute of the session.

Both measures showed a decrease in heart rate and skin conductance for the well-designed website. For the poorly designed site, the skin conductance data showed an increase over the first five minutes of the session, followed by a return to baseline over the final five minutes. The heart rate data for the poorly designed version showed some variability, but the overall trend was to stay at the same level as the baseline, unlike the well-designed version, which showed a decrease relative to the baseline. Both measures appear to reflect greater stress in interacting with the poorly designed site.

In a study of participants playing a 3D video game (*Super Mario 64*), Lin, Hu, Omata, and Imamiya (2005) looked at the relationships among task performance, subjective ratings of stress, and skin conductance. The tasks involved playing three different parts of the game as quickly and accurately as possible. Participants played each part (task) for ten minutes, during which period they could potentially complete the goal (succeed) multiple times. As shown in Figure 7.11, there was a strong correlation between participants' ratings of how stressful

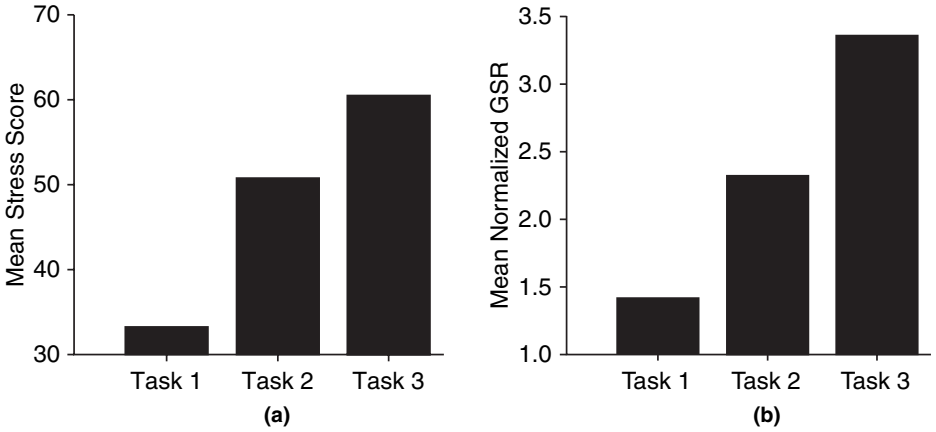


FIGURE 7.11

Data showing subjective ratings of stress (a) and normalized GSR (b) for three different tasks in a video game. Both show that Task 3 was the most stressful, followed by Task 2 and then Task 1. *Source:* Adapted from Lin et al. (2005): *Proceedings of OZCHI 2005,* the conference of the Computer–Human Interaction Special Interest Group (CHISIG) of the Human Factors and Ergonomics Society of Australia (HFESA), Todd Bentley and Sandrine Balbo (Eds). Canberra (Australia), 21–25 November 2005. (ACM International Conference Proceedings Series under the ACM ISBN 1–59593–222–4.)

each of the tasks was and their normalized GSR (change relative to the participant's baseline GSR) during performance of each task (a). In addition, the participants who had more successes during the performance of each task tended to have lower GSR levels, indicating that failure was associated with higher levels of stress (b).

Trimmel, Meixner-Pendleton, and Haring (2003) measured skin conductance and heart rate to assess the level of stress induced by the response times for web pages to load. They artificially manipulated page load times to be 2, 10, or 22 seconds. They found significant increases in heart rate as response time (page load time) increased, as shown in Figure 7.12. A similar pattern was found for skin conductance. This is evidence of physiological stress associated with the longer response times.

There have been some advances in the development of devices for measuring skin conductance and heart rate that are less obtrusive than traditional methods and potentially make the devices more suitable for usability testing. For example, Rosalind Picard and Jocelyn Scheirer (2001) of the MIT Media Laboratory invented a device called the Galvactivator glove (Figure 7.13) for measuring skin conductance using a glove that slips on the hand but leaves the fingers completely free. Likewise, Jukka Lekkala's research group at the Technical Research Centre of

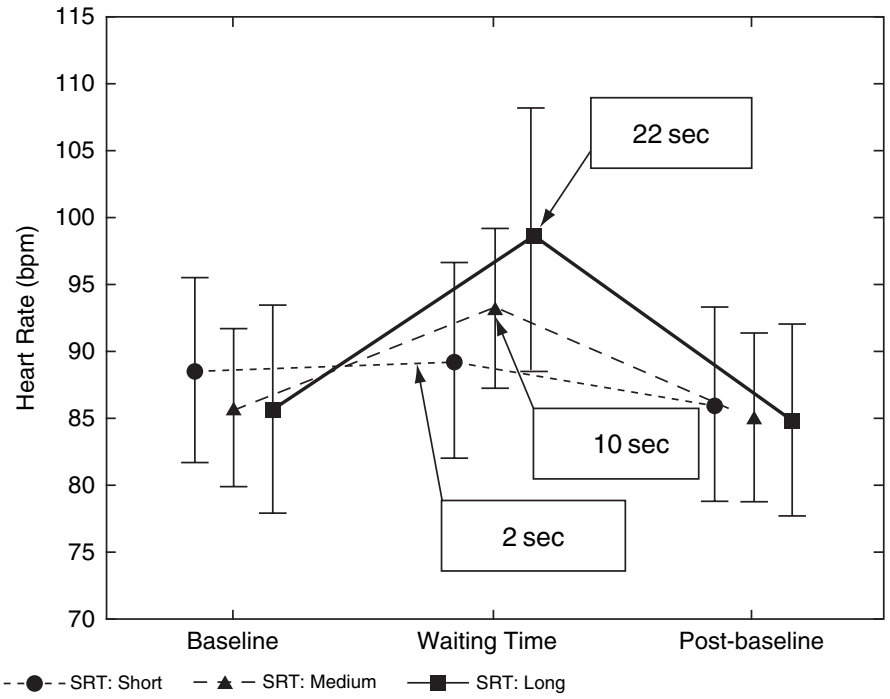


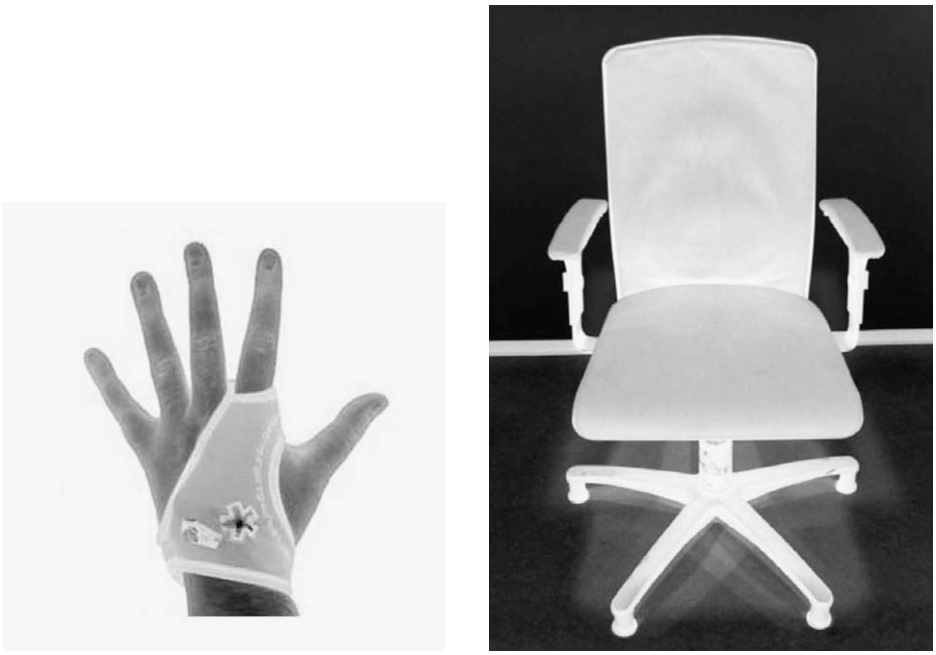
FIGURE 7.12

Data showing the heart rate of participants as they experienced different levels of response time waiting for web pages to load. Wait times of 10 and 22 seconds yielded progressively greater increases in heart rate relative to the baseline, indicating physiological stress. *Source:* Adapted from Trimmel et al. (2003); used with permission.

Finland invented a device called the EMFi chair (also shown in Figure 7.13) for unobtrusively measuring the heart rate of someone sitting in it (Anttonen & Surakka, 2005).

Measuring Stress in Everyday Usability Testing

Measuring stress as part of a typical usability study is rarely done, not because it wouldn't be valuable but because the instruments available today are simply too obtrusive for a typical usability lab. Participants are already under considerable pressure when they come into a lab with video cameras, possibly a one-way mirror, and even an eye-tracker. If they were asked to place clips on their fingers to measure stress levels, they might head straight for the door. Perhaps in the next few years new technology will become readily available, like that shown in Figure 7.13, that can measure stress levels on a continuous basis while not impeding the user experience.



(a) (b)

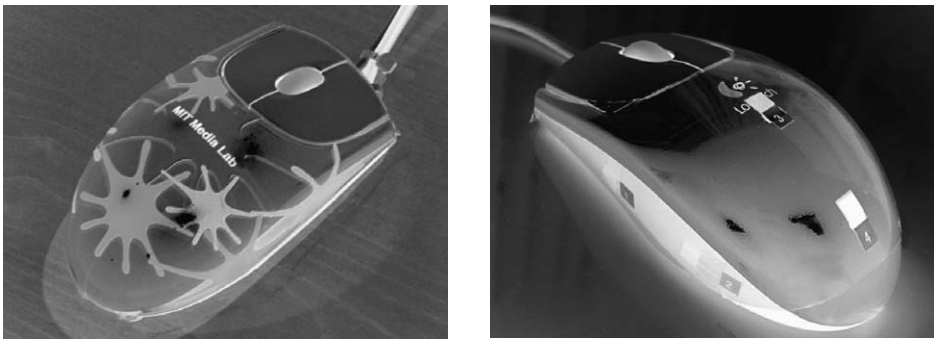
FIGURE 7.13

Advances in technology may allow for less obtrusive ways to measure skin conductance and heart rate. The Galvactivator glove (a) from the MIT Media Lab measures skin conductance while leaving the fingers free. The EMFi chair (b) from the Technical Research Center of Finland measures the heart rate of someone sitting in the chair. *Source:* (a) Courtesy of MIT Media Lab. (b) Courtesy of Affective Computing Research Group.

7.2.5 Other Measures

A few creative researchers have come up with some other techniques that might be appropriate for assessing the user's level of frustration or engagement while interacting with a computer. Most notably, Rosalind Picard and her team in the Affective Computing Research Group at the MIT Media Lab have investigated a variety of new techniques for assessing the user's emotional state during human–computer interaction. Two of these techniques that might have application to usability testing are the PressureMouse and the Posture Analysis Seat.

The PressureMouse (Reynolds, 2005), shown in Figure 7.14, is a computer mouse with six pressure sensors that detect how tightly the user is gripping the mouse. Researchers had users of the PressureMouse fill out a 5-page



(a) (b)

FIGURE 7.14

The PressureMouse is an experimental mouse that can detect how tightly the user is gripping it. The plastic overlay (a) transmits pressure to six sensors on the top and sides of the mouse (b). As users become frustrated with an interface, many of them subconsciously grip the mouse tighter. *Source:* The pressure-sensitive mouse was developed by Carson Reynolds and Rosalind Picard of the MIT Media Lab.

web-based survey (Dennerlein et al., 2003). After submitting one of the pages, participants were given an error message indicating that something was wrong with their entries on that page. After acknowledging the error message, the participants were then taken back to that page, but all the data they had entered had been deleted and they had to reenter it.

As illustrated in Figure 7.15, participants who had been categorized as members of a "high-response" group (based on their negative ratings in a usability questionnaire about the online survey) gripped the mouse significantly tighter for the 15 seconds *after* their loss of data than they did for the 15 seconds *before*.

The Posture Analysis Seat measures the pressure that the user is exerting on the seat and back of the chair. Kapoor, Mota, and Picard (2001) found that they could reliably detect changes in posture on the part of the participant, such as sitting upright, leaning forward, slumping backward, or leaning sideways. These may be used to infer different levels of engagement or interest on the part of the participant.

These new technologies have yet to be used in everyday usability testing, but they look promising. As these or other technologies for measuring engagement or frustration become both affordable and unobtrusive, they can be used in many situations in which they could provide valuable metrics, such as designing products for children who have limited attention spans, evaluating users' patience for download times or error messages, or measuring teenagers' level of engagement with new social networking applications.

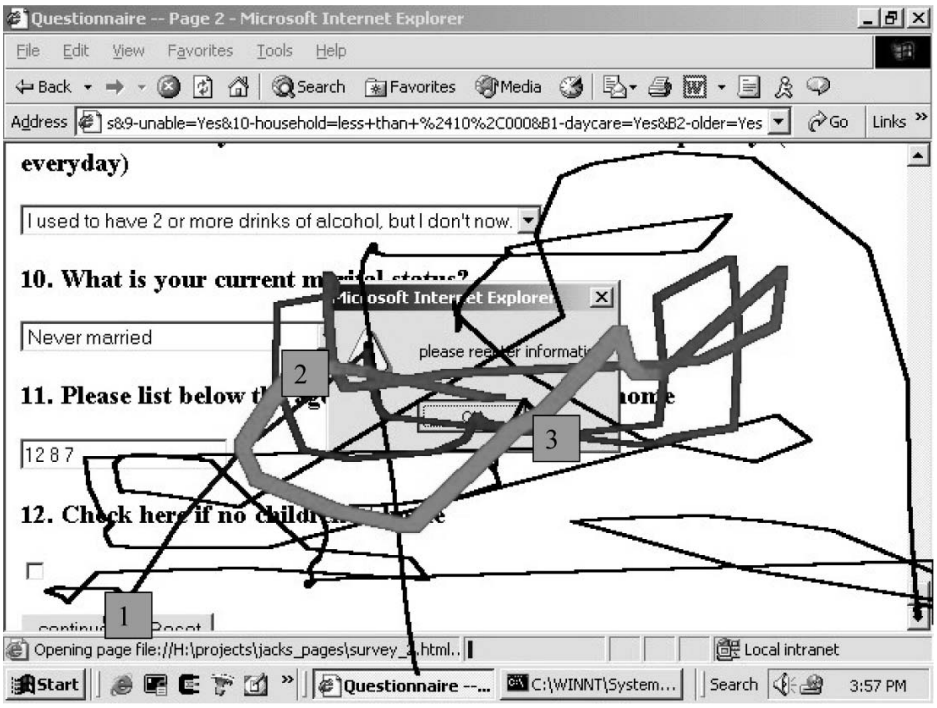


FIGURE 7.15

In this visualization of data from the PressureMouse, the mouse leaves a "trail" on the screen. The thickness of the trail indicates how tightly the participant is gripping the mouse. In this example, the participant is initially gripping with normal pressure while completing the online survey. When he clicked on the "Continue" button (#1), the pressure was still normal, until he started reading the error message, which caused him to grip the mouse tighter (#2). Finally, after dismissing the dialog box and seeing that the data he had entered was now gone, his grip on the mouse got even tighter (#3). *Source:* Adapted from Reynolds (2005); used with permission.

7.3 SUMMARY

In this chapter we covered a variety of behavioral and physiological measures that might be helpful in usability testing as additional ways of learning about the users' experiences with a product and their reactions to it. Some of these can be detected by careful observation, and some require specialized equipment. But the point about all of them is that you don't want to rely entirely on what the users say or how they perform tasks. These techniques are designed to give you additional insights that might help pinpoint parts of the interface that are working particularly well for the users or that are particularly frustrating. Here's a summary of some of the key points to remember.

1. A structured approach to collecting observational data (both verbal and nonverbal) during a usability test can be very helpful in subsequent analysis (e.g., tabulating the number of positive and negative comments made by participants during each of the tasks). A form with check boxes and other places to mark key events and behaviors on the part of the participant during each task can help facilitate this.
2. Facial expressions that participants make during a usability test may give you additional insight into what they are thinking and feeling beyond what they say. A trained observer can detect and categorize many of these expressions (e.g., frowns, smiles), but some are very fleeting and may require video analysis. Automated techniques for capturing this information using sensors on the face are more intrusive than desired for normal usability testing, but techniques using automated video analysis are being actively researched.
3. Eye-tracking can be a significant benefit in many kinds of usability tests. The technology continues to improve, becoming more accurate, easier to use, and less intrusive. Perhaps its key value can be in determining whether participants in a usability test even looked at a particular element of the interface. We've found that web designers and visual designers are keenly interested in the kinds of analyses you can get from eye-tracking data, such as the "heat maps" of where participants looked on the pages of a website.
4. If you're using eye-tracking, it might be worthwhile to also look at the pupil diameter data that the system captures. Participants' pupils tend to dilate with higher mental workload and with overall arousal.
5. Skin conductance and heart rate can be used to detect parts of an interface that participants find particularly frustrating. But the technology readily available today for measuring these is too intrusive for normal usability testing. Less intrusive technology may be available in the future.
6. Other techniques for capturing information about the participant's behavior, such as a mouse that registers how tightly it is being gripped, are on the horizon and may become useful additions to the battery of tools available for use in usability testing.