Available online at www.sciencedirect.com
ScienceDirect
Int. J. Human-Computer Studies 65 (2007) 361–375
International Journal of Human-Computer Studies
www.elsevier.com/locate/ijhcs

Relative subjective count and assessment of interruptive technologies applied to mobile monitoring of stress

Rosalind W. Picard[a,*], Karen K. Liu[b]

[a]MIT Media Laboratory, 20 Ames St. E15-020a, Cambridge, MA 02142, USA
[b]Microsoft Corporation, 1 Microsoft Way, Redmond, WA, USA

Abstract

A variety of technologies—from agents designed to assist or encourage you, to context-based messaging services—have the opportunity to interrupt you many times throughout the day. One of the challenges with designing new highly interruptive technologies is how to objectively assess their influence on human experience. This paper presents an assessment of a new mobile system that interrupts the wearer to support self-monitoring of stress. We utilize a diverse set of assessment techniques, including a newly proposed measure, relative subjective count, which compares the difference in perceived number of interruptions to actual number of interruptions. This measure, together with direct and indirect subjective reports, and a behavioral choice, is used to evaluate an empathetic version of the mobile system vs. a non-empathetic version. We found that post-experience direct questionnaire assessments such as "how stressful has using the system been?" do not significantly distinguish user experiences with the two systems; however, the new measure of relative subjective count, the behavioral choice, and another indirect questioning strategy, do point toward a preference for the empathetic system.

Keywords: Affective assessment; Relative subjective duration; Frustration measures; Relational computing; Wearable and mobile computing; Ubiquitous computing; Physiological monitoring; Stress monitoring

1. Introduction

Over the last decade, the use of computational technologies that people interact with continuously has exploded. Mobile phones, pagers, and other personal digital devices have become an essential part of ordinary daily interaction. Not only do these devices accompany us wherever our garments do, but increasingly they initiate interaction, interrupting our activities irrespective of whether we are speaking in a business meeting, enjoying an intimate encounter, or trying to sink a winning golf putt. The growing number of elderly adults and a push to improve the health of all adults is now spurring new developments of mobile health technologies, especially devices that monitor physiology and behavior while aiming to support people in adopting healthy choices. The very nature of such "helpful" devices is that they will interrupt you from time to time, and fail not because of any hardware or sensor shortcomings (although these exist) but rather because their interruptions are unpleasant. Users of desktop technologies are already familiar with the annoyances of interruptions by software that takes the initiative to ask you to install updates, upgrade various software packages, visit certain websites, or alert you to a number of other things it assumes you want to know. Technology users worldwide are increasingly bombarded with devices that interrupt, ostensibly to be helpful.

Inspired by the challenge of designing technologies that learn how to be better at interrupting people, we undertook the development of a new device aimed at helping people monitor their stress levels and interruptibility as they go about their daily activities in their usual environments. We chose to monitor stress since it is both a useful state (e.g., helping you win a race or make a paper deadline) and also a potentially harmful state (e.g., impeding immune system functioning and influencing the hormones that affect our

*Corresponding author. Tel.: +1 617 253 0611.
E-mail addresses: picard@media.mit.edu (R.W. Picard), karenliu@microsoft.com (K.K. Liu).

1071-5819/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijhcs.2006.11.019

susceptibility to and recovery from illnesses, provoking over-eating and other unhealthy behaviors, and in some cases facilitating depression (McEwen and Stellar, 1993; Sapolsky, 1998; Stress in College, 2003). The diseases that predominantly affect us now are ones of slow accumulation—heart disease, cancer, cerebrovascular disorders—diseases which are complexly intertwined with our emotions, physiology, immune system, personalities, and behaviors. It is clear that emotions and stress impact health; however, the influence of either has been hard to measure in any precise ongoing way. Few people keep track of their stress level through the day and how different activities affect it, nor is there any general understanding of how stress interacts with interruptions. We thus set out to build and evaluate a device to enable people to record their stress levels together with whether or not it was a good time to be interrupted. We designed the system to also let people easily label what it was they were doing when interrupted, and to record physiological information related to stress and activity. Our goal was to help people gather this information in a non-irritating way, so they would be willing to use the monitor over a long enough period of time that they could begin to better understand the way stress works in their life, and so we could begin to get better data about interruptibility and how it interacts with people's state and activities. Our biggest challenge was how to build a device that would be highly interruptive, help the wearer reflect on stress, and yet not increase their stress. In theory, we would also make it enjoyable to use: However, the nature of a device that aims to sample things about your behavior over the course of the day is that it will be interruptive, and frequent interruptions can cause a loss of sense of control. An ongoing lack of control is a well-known contributor to frustration and stress. Thus, we did not expect people to really be delighted when the device interrupted them.

Regularly interrupting people for the purpose of gathering data related to their thoughts, feelings, and activities is not a new endeavor, but is an established methodology in psychology. The original "Experience Sampling Method" (coined by Larson and Csikszentmihalyhi, 1983) referred to a particular technique whereby participants were interrupted by a device such as a pager and asked to fill out a log of experiences. The term *ecological momentary assessment*, or "EMA," is also used to refer to experience sampling as well as to procedures that sample aspects of a person's physical state (e.g., ambulatory blood pressure or heart rate) (Stone and Shiffman, 1994). Traditional EMA tools use random or fixed timing prompts to interrupt people many times during a day; however, mobile devices can also adapt their interaction by detecting changes in user state through body-worn sensors (Picard and Healey, 1997). The newest ESM tools use PDA's and sensors for detecting context shifts (Intille S.S. et al., 2003; Intille S. et al., 2003). Experience sampling has also been used to evaluate and assist in the development of ubiquitous computing applications (Walker and Consolvo, 2002), and has been useful in a diary study to examine task switching and interruptions with information workers (Czerwinski et al., 2004).

Lisa Feldman-Barrett, a psychologist and expert in experience sampling for emotion assessment, advised us early in the design of the user experiments that drop out rates are high and "you cannot pay people enough to be in these studies." Being interrupted more than a dozen times a day by a device asking you to stop what you are doing and give it information can become so irritating that it is a challenge to get people to stay in the experiments. She advised us on the importance of carefully motivating all the subjects up front to stick with the study.

While long-term experience sampling is a significant challenge because of subject drop-out rates, our challenge is greater still: we wish to craft interruptive technologies that are not simply tolerated for a paid experiment, but that people would actually want to use on their own, long-term. Is it possible for a device to interrupt you a dozen or more times a day, asking you to stop and give it significant information, without irritating you?

We find that a useful general approach to such a question is to try the exercise advocated in the media equation (Reeves and Nass, 1996) where you re-ask the question, substituting a person for the device. Thus, we ask: Is it possible for a *person* to interrupt you a dozen or more times a day, asking you to stop and give them significant information, without irritating you? The answer in this case is yes; however, we have to understand how people accomplish this, especially since it is not true of everybody. If we can figure out how people accomplish it, then it may be possible to imitate these features in the device, and get a similar result.

A key aspect of successful human–human interruption is the act of showing consideration for the feelings of the person being interrupted. This goes beyond merely issuing a social "Hello" or a polite "Thank you." Is the person you interrupted saying they are stressed right now? If they are, it can be appropriate to say something like "Sorry to hear" (and to not smile) before saying thank you for the data. If they say all is great, then it is better to say something like "great to hear" (and a smile is ok). Such an adaptive response, while short and subtle, can make a significant difference in influencing whether a person will be annoyed when you interrupt them again. We hypothesize that if an interruptive technology adapts its response in a way showing consideration toward the person's feelings, then it is likely to improve people's experience with that technology.

The challenge of designing these types of adaptive responses will also need to take into consideration a socio-cultural context. What may be considered an empathetic response in the United States may not be received in the same way to a user in Japan. The findings we report in this paper are limited to a group of subjects who live in the United States. However, the methodology we introduce here can be applied to any culture.

The act of acknowledging somebody's feelings, especially if you have caused them negative feelings, can have a calming effect, and this effect has been shown to be significant even when used by computers (Klein et al., 2002; Prendinger and Ishizuka, 2005). We thus decided to make a version of our device that empathetically acknowledged people's stress levels. We also decided to compare two strategies for generating the timing of the interruptions.

The rest of this paper describes the new wearable device for interactive stress monitoring and how we assessed it, comparing a version of it that used empathy and sensor-triggered timing ("empathetic system") with a version that ignored people's feelings and used random timing to generate interruptions but was otherwise identical to the other system ("non-empathetic system"). We examine a variety of assessment measures, including behavioral choice (which system did they choose to use again after having used both) and several kinds of self-report metrics. In particular, we propose a new measure, *relative subjective count (RSC),* which asks a subject, "How many times do you think the system interrupted you?" and divides this estimate by the actual number of interruptions. This measure aims to indirectly assess a user's overall frustration with a technology, in a way that avoids the possibly more threatening direct question of "how enjoyable was it?" Direct answers to such questions can vary enormously based on feelings toward the experiment, experimenter, and other factors that make self-reported affective evaluations notoriously fickle; hence, indirect techniques are of importance in the search for reliable affective assessment techniques (Picard and Daily, 2005). We found the new RSC measure varied significantly for the two systems, and also agreed with subjects' behavioral choices and several other direct and indirect self-report measures. The new measure may therefore provide a new way to assess frustration or irritation of an interruptive technology without having to ask directly about a user's feelings.

2. Wearable system for interactively monitoring stress levels: PMobile

We designed a custom system, named PMobile, to gather both passive sensor information: inter-beat intervals from the heart, pedometer and accelerometer activity, and location beacon information, as well as three pieces of interactive information: user's reported stress level, user's report of whether or not the timing of the interruption was good, and user's reported activity. This system is part of a growing collection of research efforts to use computer sensors in learning and reasoning about interruptibility and attention (Horvitz and Apacible, 2003; Hudson et al., 2003; Horvitz et al., 2003). The system PMobile was designed to enable greater understanding of how sensor and behavior data inform interruptibility of users in mobile situations.

Two versions of this system were created in order to support the experiment in the following section: an empathetic version (E-PMobile) and a non-empathetic version (N-PMobile). The two systems look physically the same and their interaction differed only in two subtle ways described below. To subjects in our experiments, each of whom used the two systems in counter-balanced order, the systems were simply described as "system one" and "system two."

Below we first describe the common features of the systems, followed by their differences.

2.1. PMobile custom hardware and software architecture

The hardware (see Fig. 1) consists of a Hewlett-Packard 5550 iPAQ together with a set of wireless sensors, some that are worn on the body, and some that the user places in his or her environment to label locations. The wireless sensing apparatus was developed in collaboration with Fitsense Technology (Fitsense) and includes a Pulser chest strap for electrocardiogram inter-beat interval (IBI) information, a Foot pod for accelerometer information, a Pacer for pedometer information, and Location Beacons for environment context.

As an alternative to global positioning technologies, the location beacons can be used to mark any selected physical context of interest to the user, including mobile locations such as the automobile or bicycle (for monitoring stress while commuting), simply by placing one of the tags in that location. The user was free to place the tags in any location of his or her choosing, and did not need to communicate

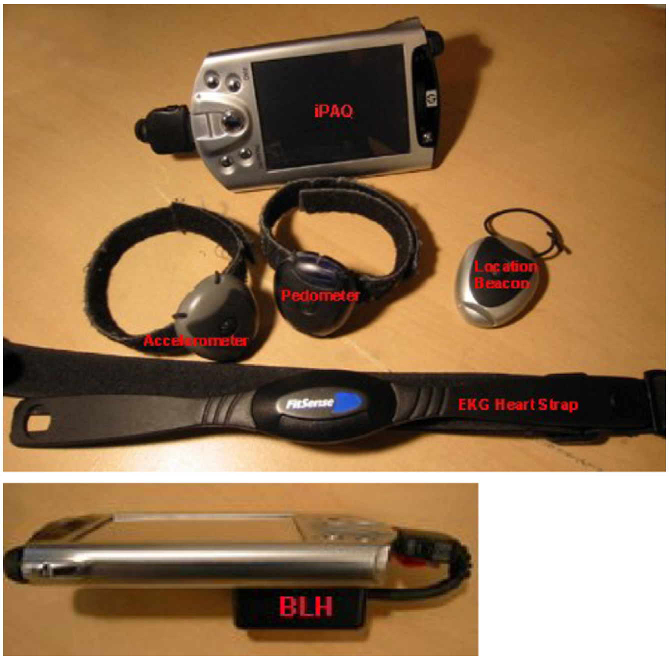


Fig. 1. Custom hardware for the PMobile System: Top: iPAQ with sensors. Bottom: iPAQ with attached Body Lan Hub (BLH), which communicates wirelessly to all the sensors.

this location to the experimenters, thus allowing for privacy in location data.

Since factors such as exertion, as well as emotional stress or arousal, contribute to increases in heart rate (The Surgeon General's Office, 1996), an accelerometer was combined with heart rate recordings to enable better decoupling of these elements in examining stress, as some studies have done to isolate exertion (Strath et al., 2002).

A BodyLAN Hub (BLH) connects the iPAQ to the different sensors on a BodyLAN wireless network. The BLH communicates with the iPAQ through a 4800-baud serial connection with two data lines: transmit and receive. The serial connection uses eight data bits, one stop bit, no parity bit, and no flow control. The BLH stores the latest message from each of the sensors and holds a settable reply message for each. Each BLH is assigned its own 32-bit address to ensure that there is no interference between two sensors on different network IDs (such as two Pulsers on different systems). The BLH has two high level modes: normal and learn. When the BLH and the sensor are both in learn mode, the BLH will automatically acquire the sensor, add it to its registry, assign the sensor an index, and switch the sensor from learn to normal mode. In normal mode, the BLH will only accept data from sensors with its own network ID and public sensors (such as the location beacons).

The Pulser chest strap is used to measure electrocardiogram information. It is set to transmit information every 2 s using a Data Variant 3 that was developed for this research. Data Variant 3 transmits a data message containing the beat count and the last 16 inter-beat intervals (IBI's) in milliseconds. The major features of the electrocardiogram are the P, QRS, and T waves which are caused by the corresponding electrical impulses in the heart of atrial depolarization, ventricular depolarization and ventricular repolarization (Mohrman and Heller, 1991; Dubin, 1996). IBIs are derived from the R-waves of the ECG by taking the time interval from the top of the QRS complex to the top of the next QRS complex.

The software and overall architecture for the system was developed by Karen Liu; details can be found in her thesis (Liu, 2004). The overall architecture is shown in Fig. 2. The continuous annotation system was written in Embedded Visual C++ 4.0 and runs on an HP 5550 iPAQ running Pocket PC 2003. The system consists of a sensor layer that communicates to a *Data Collection Platform*. The *Interaction Engine*, based off of the CAES Engine (Intille S.S. et al., 2003; Intille S. et al. 2003), schedules interactions to interrupt the user for annotations either through a timer or through triggers from the sensor information. The Interaction Engine uses a *Dialog Manager* in the QuestionDataFile format (Rondoni, 2003) of CAES to choose from a set of different interaction scripts and possible responses to user input. A GUI layer receives input from scheduled and triggered interactions, as well as the possible question/response scripts, and interacts with the user through different GUI screens.

Examples of the GUI screens are shown in Fig. 3. Users of both PMobile versions see the same front screen. Both have the same controls over muting the system, initiating annotations, and customizing the activities that they wish to use as labels. Users can initiate an interaction, or wait for the system to initiate (interrupt them). Users of both versions also have the option of simply recording a voice annotation by pushing a single button (for rapid annotation of where they are or what just happened, without initiating any dialog with the system.) This is a valuable feature when the user wants to record an event without looking at the device, e.g. "Just cut off by a big truck," spoken into the device without taking eyes off the road (while driving). Use of these optional audio annotations was not analyzed in this research.

After the data collection phase of the experiment was over (eight days, split into two 4-day sessions) subjects were presented with their personal data in the form of radial graphs such as that shown in Fig. 4. A subset of the data, heart rate, accelerometer and pedometer information could also be examined in real-time on the iPAQ (see Fig. 5).

2.2. Two versions of PMobile: empathetic and non-empathetic

We developed two nearly identical versions of the PMobile system so that one version could serve as a control for testing our ideas. When the user initiated interaction with the system, there were no perceivable differences in the two versions. However, when the system interrupted the user, there were two kinds of differences: (1) an empathetic line of text responding to the user's stated level of stress and (2) different timing of the interruptions, based on sensor input (see Table 1).

The GUI in Fig. 3 asked whether or not this was a good time to interrupt (with a binary choice), and asked the user to choose one of five levels of stress. This GUI is what users saw when *they initiated interaction* with either version of the system. A good and bad time annotation was offered here, since the user may want to record a bad time (i.e. taking an exam) in order to record how their heart and stress levels appeared at this time, or in order to teach the system bad examples. However, when *PMobile initiated interaction*, it did not show this GUI, but rather showed a text dialog that collected the same timing and stress level information in a conversational format. The dialog varied in wording from time to time, pulling text from a set of scripts, to reduce monotony. However, the dialog was always structured the same as the GUI: it first asked whether or not the timing of the interruption was ok, and then it asked you to select one of five (text described) levels of stress. The structure and scripting of the dialog was identical across the empathetic and non-empathetic conditions with only one exception: In the empathetic case, the system included an additional line of text, selected to respond to the user's stress level. This difference is illustrated in Fig. 6.

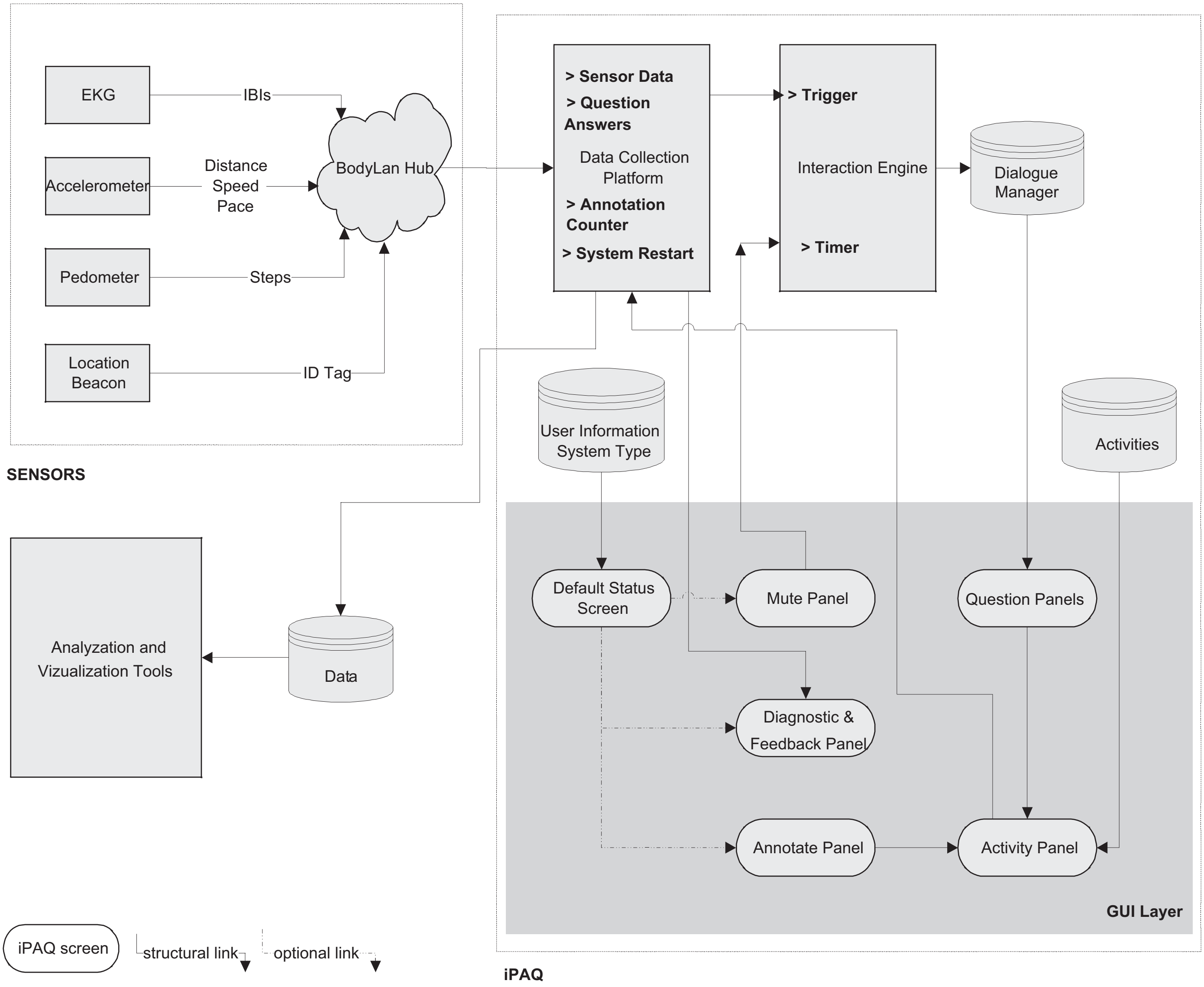
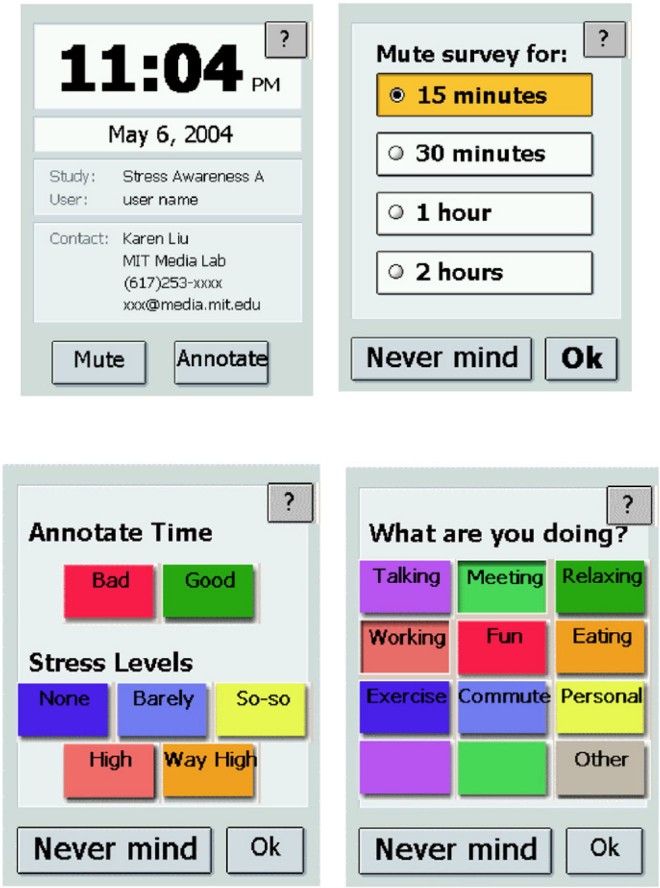


Fig. 2. Architecture of the PMobile system. Sensor information is sent to the BodyLan Hub which communicates to a Data Collection Platform. The Data Collection Platform stores sensor data, question answers, annotated activities, number of annotations, and whether the system has been restarted. The stored data is later used for the analysis and visualization tools. The Interaction Engine schedules interruptions to the user for annotations either through a timer or through triggers from the sensor information. The Interaction Engine uses a Dialog Manager to choose from a set of different interaction scripts and possible responses to user input. A GUI layer receives input from both system-triggered and user-triggered interactions, as well as the possible question/response scripts, and interacts with the user through different GUI screens.

It is important to underscore that the dialog differences in the two versions of the system are not because one was "friendly" or "social." Both systems used social and polite dialog. Both systems always greeted and thanked the user. The only difference in the dialogs was that the empathetic condition included one additional line, an empathetic response to the user's stated level of stress. Several different wordings of these responses were scripted for the five levels of stress, for a total of 26 varied responses (Table 2) The adaptive nature of the scripting was inspired by scripts for agents designed to facilitate a long-term social–emotional relationship with a person Bickmore and Picard (2005) and in this case followed a very simple algorithm (see Fig. 7).

The other difference in the two versions related to the timing of the interruptions. In the non-empathetic version, interruptions were set to occur at a randomly selected interval between a specified minimum and maximum number of minutes from the last session to ensure that questions are asked throughout the entire day, but that there is some degree of randomness in the interruptions. For the empathetic version, algorithms triggered the interruptions based on data from sensors. The algorithms triggered interruptions immediately after there was a change in a context beacon location or when there was a significant "heart-rate change event" as defined mathematically Liu (2004). Note that a similar approach with a different heart-rate trigger algorithm appeared about the



11:04 PM

May 6, 2004

Study: Stress Awareness A
User: user name

Contact: Karen Liu
MIT Media Lab
(617)253-xxxx
xxx@media.mit.edu

Mute survey for:
● 15 minutes
○ 30 minutes
○ 1 hour
○ 2 hours

Fig. 3. Both versions of the system use these four GUI screens. The upper left GUI gives contact info, date, and time. The upper right one allows muting the system (with its audible tone) for up to 2 h. The GUI at lower left is seen by the user only when he or she initiates interaction with the system; otherwise, this part of the interaction is replaced by either an empathetic or non-empathetic text dialog that collects the same timing and stress level information. The GUI at lower right contains activity labels customizable by the user for one-button labeling of events the user wants to record.

same time by John Rondoni (2003). The idea was that such context-based triggers would be less disruptive for gathering information from users.[1]

3. Experiment: assessment with multiple techniques

This section presents a user study for evaluation of the systems above. We employed multiple assessments, including a questionnaire with self-report measures, a new index we call relative subjective count, and a behavioral choice test. These three techniques were used to examine the following hypotheses:

H1. Subjects will find the empathetic system to be less disruptive and frustrating to use and will have a better experience while using the empathetic system.

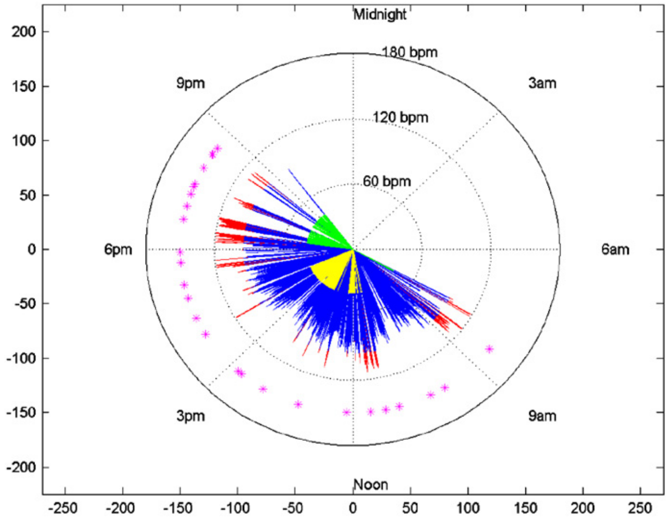
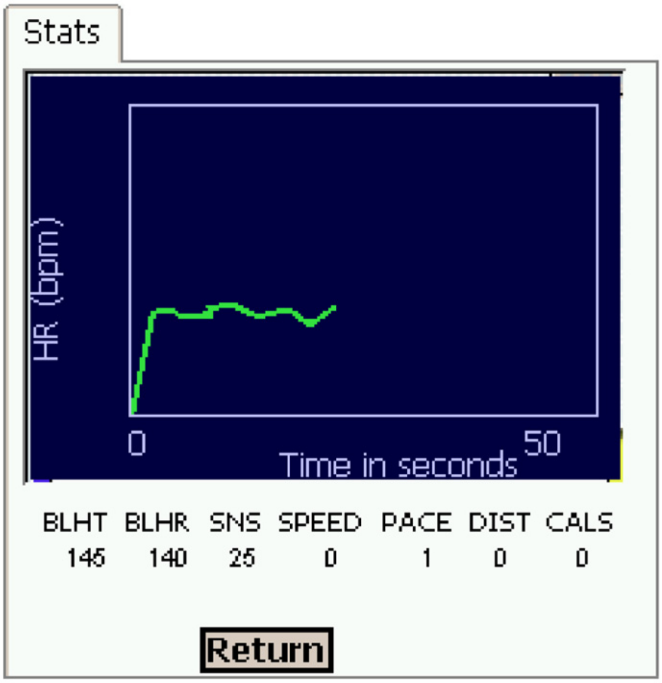


Fig. 4. Heart-rate data viewed on a 24-h radial plot. The green and yellow colors near the center indicate that the user was in the vicinity of those two context beacons. The purple asterisks near the perimeter indicate times when the user labeled an activity either via the GUI or via a voice message. The user could click on the asterisk to retrieve the label, e.g., after 5:30 p.m. where heart rate is 120 beats/min the user might have recorded, “ran to train.”



BLHT	BLHR	SNS	SPEED	PACE	DIST	CALS
145	140	25	0	1	0	0

Fig. 5. Here is a screen from the HP iPAQ showing heart rate, which could be viewed at any time by the user.

[1] Note that it remains to be determined how to ideally time interruptions to get accurate labels without increasing the wearer’s stress. This work took initial steps in that direction, and there is still much to learn. One might argue that times of changing activity are also likely to be times of greatest stress; hence, it might be advantageous to wait longer after an activity change for the first physiological signs of decreased stress before interrupting. However, that approach could also potentially interrupt a “flow” state, destroying the pleasurable focus of that state. More research is needed to better understand how user state interacts with different modes of interruption.

H2. Subjects will perceive that the empathetic system interrupted them less than the non-empathetic system (compared to the actual number of interruptions).

H3. Subjects will choose to continue working with the empathetic system.

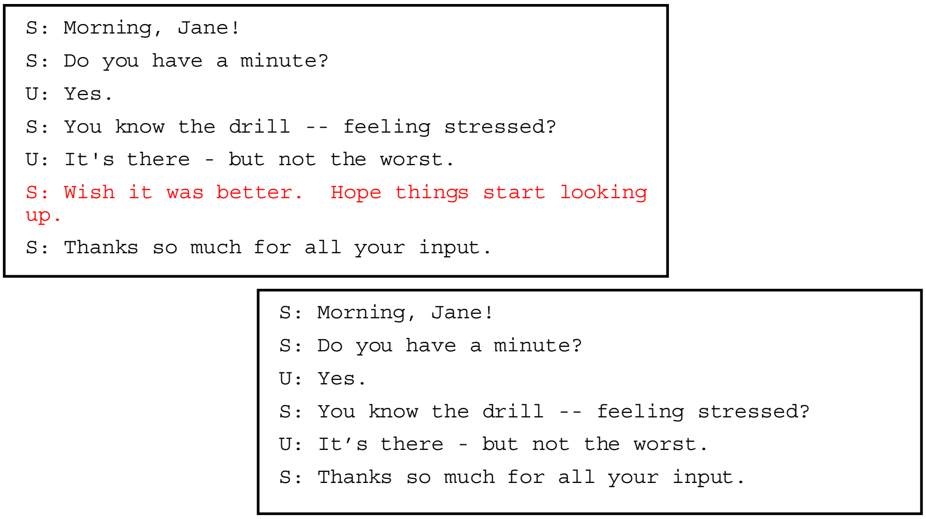
Table 1
The two versions of PMobile differed both in empathetic dialogue and in sensor-based interruption timing

PMobile two versions	Social and polite dialog	Empathetic response to stress level	Sensor-based interruption timing
Empathetic (E)	Yes	Yes	Yes
Non-empathetic (N)	Yes	No	No

We additionally measured physiological and sensor data; we comment on those at the end of this paper.

3.1. Subjects

Subjects were recruited via email solicitation and postings on public message forums. Each subject committed to wear a heart strap, accelerometer, and pedometer, to place two location beacons in different locations such as home and office, and to carry around the iPAQ from 9:30 a.m. to 9:30 p.m. each day "for 2 weeks." Because of the nature of the custom hardware, limited budgets for building duplicate systems, and the difficulty of getting all the system elements to work reliably at the same time, we were only able to have from 3 to 4 fully operating systems for subject use. Within time limits of the project, we signed up three separate batches of subjects, each batch



```
S: Morning, Jane!
S: Do you have a minute?
U: Yes.
S: You know the drill -- feeling stressed?
U: It's there - but not the worst.
S: Wish it was better.  Hope things start looking
up.
S: Thanks so much for all your input.
```

```
S: Morning, Jane!
S: Do you have a minute?
U: Yes.
S: You know the drill -- feeling stressed?
U: It's there - but not the worst.
S: Thanks so much for all your input.
```

Fig. 6. Sample dialog for the empathetic version (top) and non-empathetic version (bottom). The only difference is a short response to the level of stress stated by the user.

Table 2
Scripted responses to five levels of stress (used only in empathetic version)

Very stressed	Stressed	Neutral	Low stress	Very low stress
I'm sorry to hear that. I hope you feel better soon.	Really sorry to hear.	Seems things are going pretty neutral.	Seems like you're feeling good. Good to hear.	Great to hear!
Sounds really bad. I'm sorry that you're feeling that way. :(	Wish it was better. Hope things start looking up.	Seems like things are going ok.	Glad to hear.	Awesome. Have a good day!
Wow. You sound pretty stressed out. Hope things start looking up.	Sorry to hear—hope things calm down.	Does not sound too bad. Hope your day picks up.	Looks good. Happy to hear.	Sounds great!
Doesn't sound too good. I'm sorry to hear. :(	Sounds like you're pretty stressed. Sorry to hear.	Sounds like it's one of those so-so times.	Seems like things are going well. Nice.:)	These are the best! :)
	Sounds pretty bad. Hope things get better.		Sounds pretty good.	Happy to hear
			:)	Wonderful.
				Nice! :)

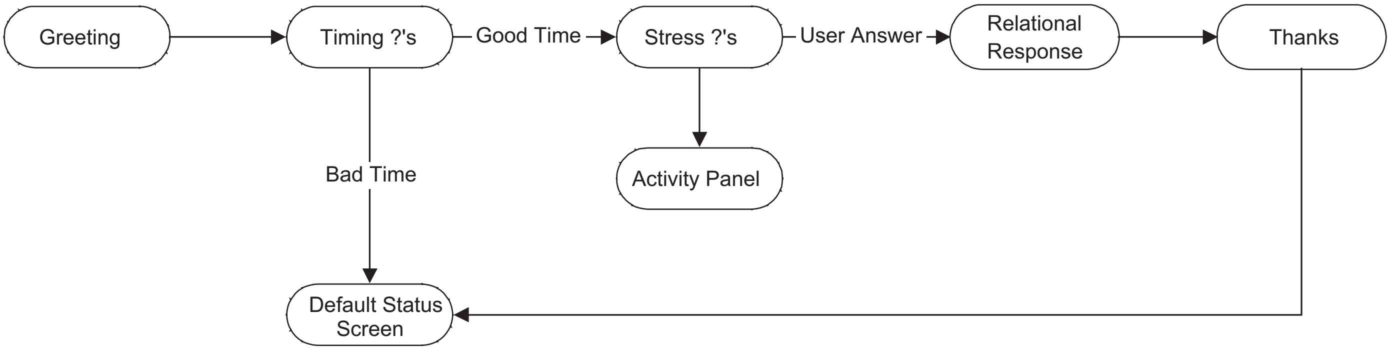


Fig. 7. Interaction for empathetic condition follows the transitions in this diagram.

for four days with one system, and for four additional days with the other system (in counter-balanced order). We were able to collect complete sets of data from a total of ten subjects (5 males and 5 females, ages 22–33, 6 students and 4 professionals), resulting in a total of 79 days of data (three subjects only finished 7 days, while one subject completed two extra days). Each day was designed to allow up to 12 h of use, resulting in around 600 h of tested system interaction data. Subjects received either a movie ticket or a gift certificate for a local coffee shop for each laboratory visit and $75 cash upon completion of all tasks in the study.

3.2. Apparatus

The experiment uses the PMobile software and hardware described above. Each subject was given a sensor system consisting of one heart strap, one accelerometer, one pedometer, one BLH, one iPAQ, and two location beacons.

3.3. Procedure

During the first laboratory meeting, subjects were told that the overall purpose of the study was to investigate people's stress patterns in natural activities and collect stress and activity information for developing computer algorithms to recognize patterns from sensors. They were asked to sign a consent form and the renumeration procedure was explained. Subjects were then shown how to put on the sensors, were given take-home instructions, and were asked to fill out end-of-day logs each day. Subjects were told that they would be asked to use two different systems—"System 1" for one 4-day session, "System 2" for a second 4-day session, and the system of their choice for a third and final 4-day session. Subjects did not know which system, empathetic or non-empathetic, they were using, nor were differences described for the systems during the experiment. Finally, subjects were given a questionnaire for obtaining base-line stress levels, demographic information, and personality information.

Each morning, subjects would put the heart strap around their chest and use a Velcro band to put the accelerometer and pedometer on each ankle. Subjects used either the empathetic or non-empathetic PMobile system for 4 days (session one) and filled out the end-of-day logs online or on paper each day. Paper versions of the end-of-day logs were given to subjects in case they did not have access to a computer at night.

At the end of the 4 days, subjects came into the laboratory to download their data and replace the batteries in their sensors. All subjects had the version of their system switched at this point (empathetic to non-empathetic, and vice-versa.) Subjects then used this other version of the system for four more days (session two), filling out end-of-day logs as before.

At the end of session two, subjects came into the laboratory to download their data and to begin what they believed to be the third and final session. They were first given time to explore their data with clickable annotations on radial plots. When finished, they were given an online evaluation questionnaire for the first two systems that they had used. (Questionnaires are viewable in Liu, 2004). After these questions, each subject was asked to select "System 1" or "System 2" for use in a third 4-day session and to explain why they made their choice. After this step, the next page of the form told subjects that the study was now completed, without the need for the third session, and asked how they felt about the study being completed. Subjects were then asked to meet with the experimenter and were told the complete goals and design of the study, why they were misled about session three, and compensated for their full participation for the time they were originally asked to commit to the study.

The protocol, as for all studies involving people at MIT, was pre-approved by the MIT Committee on the Use of Humans as Experimental Subjects. Subjects were openly debriefed as to the reason for deception about the third session and were given the right to withdraw their data. No subjects exercised this option.

3.4. Measures: self-report questionnaire

Questions were designed that both directly and indirectly addressed feelings about each system. All questions were asked after Session 2 was completed. Three questions were asked separately about both the "first system they had used" and the "second system they had used": "In general, how disruptive do you feel the timing of the interruptions were?" "How stressful has using the system been?" and "How responsive did you feel the system was to your stress?" We predicted that the empathetic system would score better than the non-empathetic on all three questions. Another set of five questions was asked after these about their experience with the stress awareness study. The exact five are given below, with "How do you feel about the study being completed?" being one example. The answers to these questions were compared across the subjects, pooling them into two groups based on which version of the system they had used most recently. The hypothesis here is that the subjects who ended with the empathetic system would *remember* having a better experience overall even if in actuality the experience was only better during session 2 (due to a strong influence on overall impressions believed to occur when an overall unpleasant experience ends relatively pleasantly (Redelmeier and Kahneman, 1996). Additionally, we reasoned that if one of the systems was significantly more frustrating than the other, then subjects who had just been using that system should be happier about quitting than those who had been using a less frustrating system. Thus, measuring these between-group differences, based on whether the group had used the empathetic or non-empathetic system in session two, should reveal indirect information about subjects' affective experience. Finally, we found in earlier work that current mood can significantly influence perception of technology quality such as audio-visual quality as well (Mueller et al., 2002). Thus, we might even find that the sensors and graphs are viewed more favorably if the user is currently happier with the system they are using. All of these questions were asked on a 7-point Likert scale. The specific items are listed together with results below.

3.5. *Measures: relative subjective count (RSC)*

A new assessment tool that we propose and evaluated in this work is the *RSC*. We were inspired to create this new measure from a similar measure known as relative subjective duration, which is a value that takes the user's estimated time to complete a task divided by total time to complete the task as an implicit probe for measuring user frustration or satisfaction (Czerwinski et al., 2001). In our work, the user's estimated number of interruptions (collected during end-of-day logs) divided by the actual number of interruptions was used as an index for probing user frustration. We hypothesized that the tendency to underestimate the number of interruptions would correspond to lower frustration with the device. Thus, without asking them directly how frustrating they thought the technology was, we tried to assess how it might make them feel.

3.6. *Measures: behavioral*

One of the strongest measures for evaluating preference or determining true motivations can be seen in the difference between what people say they want to do and what they actually do. When on a diet, people say they will choose a healthy snack, but later when they actually choose a snack, they may eat candy bars. People may say they buy cable television so they will watch the History Channel, but then, instead, they watch a channel they do not wish to admit watching. Acknowledging these well-established differences between what people say they will choose and what they actually do choose, we led subjects to believe that the experiment would last an additional 4-day session, and asked them to choose which system they wanted to use for the third session (of the two they had already interacted with.) Thus, we can expect that subjects selected the system that they most wanted to have interrupting them for an additional 4 days. Since subjects were led to believe that there were three sessions in the study, the strongest behavior measure for evaluating user system preference would be which system the user chooses to continue to work with.

Only after they had finished their evaluations of both systems, finished the comparison evaluation, and selected their system for the next session, did we tell them that they had completed the experiment, and that session three would not actually occur.

4. Results

All ten subjects completed the data collection as well as the laboratory visits and questionnaires. However, there was a problem with the sensor-based interruption triggers for the empathetic condition for three of the subjects. In all three cases the effect was the same: these three subjects, all using the empathetic system, were interrupted significantly more than were the other seven subjects. One of them had 3 days with over 50 interruptions each day, while the average across the three subjects was 28 interruptions per day using the empathetic system, and 10.4 using the non-empathetic. In contrast, the other seven subjects averaged 11.7 interruptions a day using the empathetic system, and 11.3 using the non-empathetic system. For Subjects 1 and 3, the problem was traced to the interruptions being triggered accidentally by a power level change, instead of by the context beacon change. These subjects were both using the empathetic system for session one. The bug was fixed and did not affect the non-empathetic data for these subjects. A different problem occurred with Subject 4 who used the empathetic system during session two. The sensor-based interruptions were designed to trigger with significant heart-rate changes, typically associated with a change in activity, and thus a likely change in context. While we had pilot-tested the sensor-based interruption algorithm on a number of people, our sample did not adequately include people who were overweight. Because heart rate increases when a person stands up, and can increase dramatically more for an overweight person, the heart-rate-based algorithm can trigger an interrupt every time they stand up. This appeared to be the problem with Subject 4, who had just begun a new exercise program to lose weight, and who received significantly more interruptions than we intended.

Given the excessive number of interruptions for these three subjects, we analyzed the data both with and without including them. With all ten subjects, the average number of daily interrupts per subject was 17 for the empathetic condition and 11.1 for the non-empathetic condition. Omitting data from the three subjects brings the daily average to 11.7 for the empathetic and 11.3 for the non-empathetic versions, thus nearly equalizing the number of interruptions. Note that it would be logical to expect that a system that interrupts more would be more disruptive, more stressful, and less preferred than one that interrupts less. EMA studies usually do not interrupt more than ten times a day, and they are known to be irritating, so the number of interruptions provided by our systems may be viewed as potentially really annoying according to EMA expert Lisa Feldman-Barrett. Thus, the problems with the interruption strategy that resulted in many extra interruptions are expected to work against our hypotheses, and findings that support the hypotheses while using all ten subjects may be viewed as stronger evidence for the value of the empathetic system. Below we report the results organized according to the three categories of measures as described above. These are followed by a discussion of the joint findings.

4.1. Self-report questionnaire

After subjects had used both systems (completed session two), they were asked to evaluate both the system used in the first session and the system used in the second system. Fig. 8 shows the system evaluation for both the empathetic system and non-empathetic system with the bar graph illustrating the mean values and standard deviations for Likert scale responses on a 1:7 scale. We present the results in two ways: (1) for all ten subjects, and (2) omitting the three subjects who encountered problems with the sensor-based interruptions.

For the question, "In general, how disruptive do you feel the timing of the interruptions were?" (7 = "disruptive", 1 = "not disruptive") there was a trend toward viewing the timing as more disruptive in the empathetic system when all ten subjects are considered, (5, SD = 1.3 for empathetic; 4, SD = 1.3 for non-empathetic) and this trend lessened when we examined only the 7 subjects with a more balanced average number of daily interruptions across the conditions (4.6, SD = 1.1 for empathetic; 4.3, SD = 1.1 for non-empathetic). Thus, the problems with the empathetic system interrupting more may have contributed to the perception that the timing of its interruptions was more disruptive. Note that if the adaptive sensor-based timing of the empathetic system's interruptions was perceived as better than the random timing of the non-empathetic system, we would expect to see the difference in this measure go the other way, at least for the 7 subjects where the average number of interruptions was relatively balanced across the systems; however, no such effect was apparent.

The ten subjects' self-reported answers to the direct question, "How stressful has using the system been?" (7 = "very stressful", 1 = "reduced stress") were 3.7 (SD = 1.3) for empathetic and 3.7 (SD = 1.6) for non-empathetic. The seven subjects reported a mean rating of 3.3 (SD = 1.1) for empathetic and 3.6 (SD = 1.3) for non-empathetic. Thus, asking people specifically about each system at the end of the experiment (after each person had used each system) did not reveal a large distinction.

For the question, "How responsive did you feel the system was to your stress," (1 = "Not responsive, 7 = "Very responsive") the subjects showed a trend toward saying the empathetic system was more responsive, 4.1 (SD = 1.2) for empathetic and 3.2 (SD = 1.0) for non-empathetic (ten subjects) and 4.0 (SD = 1.3) for empathetic and 3.1 (SD = 1.2) for non-empathetic (seven subjects).

Now we consider the data gathered from the five questions about experience with the study as a whole. These questions are more indirect in their assessment, as we are looking for a general bias in judgment given a more

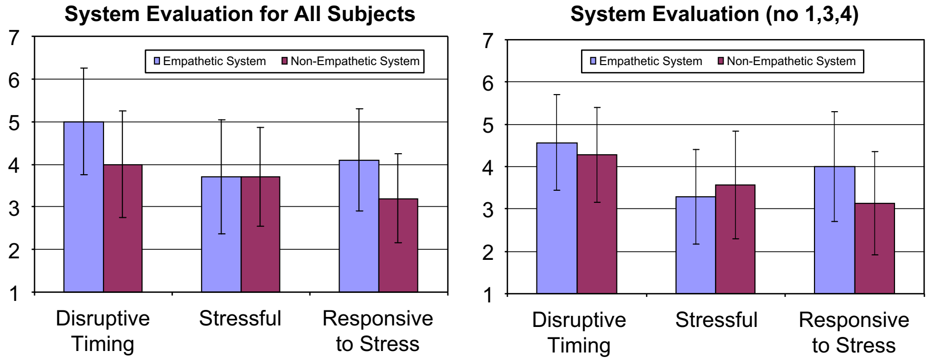


Fig. 8. Self-reported questionnaire data for all ten subjects (left) and for all but the three who experienced problems with the sensor-based trigger interruptions (right).

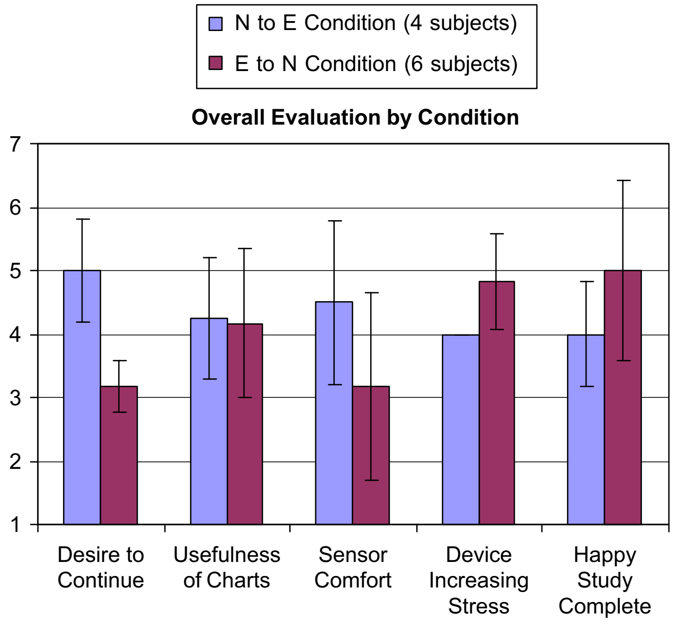


Fig. 9. Subjects who had just finished using the empathetic version (E) evaluated the experience overall more positively than those who had just finished using the non-empathetic version (*N*).

positive state at the time of the assessment. These data (see Fig. 9) were analyzed across the two groups based on which version they had just used in session two. The group who ended session two with the empathetic system contained 2 males and 2 females, while the group who ended with the non-empathetic system contained 3 males and 3 females.

Subjects who had most recently been using the empathetic system (N-E group) expressed significantly higher ratings when asked after session two, "To what extent would you like to continue working with the Stress Awareness system" (7 = "yes, very much so", 1 = "No way, I'd be happy to be free"), 5.0 (SD = 0.82) for empathetic and 3.2 (SD = 0.41) for non-empathetic. Given that this question applied to the experiment overall, and the N-E group felt positively about it (mean >4), while the E-N group was negative (mean<4), we interpret this as indirect evidence that those who had just been using the empathetic system remembered a better overall experience.

In response to the question "How useful were the charts of your heart activity with your annotations?" the mean response for both groups landed halfway between "7 = very useful" and "1 = not very useful", 4.3 (SD = 0.96) for empathetic and 4.2 (SD = 1.2) for non-empathetic, indicating that there is room for improvement in this aspect of the system (which took place on a computer in our lab, and not on the mobile devices). One subject commented that the radial plots were more difficult to read than a linear plot, since the radial plot used a 24-h time window, instead of the typical 12-h window typical in the United States.

Subjects who had most recently used the empathetic system showed a trend toward rating the sensors as more comfortable (7 = "very comfortable"; 1 = "very uncomfortable"), 4.5 (SD = 1.3) for empathetic and 3.2 (SD = 1.5) for non-empathetic. In general the females rated the chest strap more comfortable than the males, 4.2 (SD = 1.6) for females and 3.2 (SD = 1.3) for males, although the trend found here for improved sensor comfort in the N-E group is not due to gender since females were balanced in the N-E and E-N groups.

Answering the question, "How did you see the device interacting with your stress levels? (7 = "Increasing it", 1 = "Reducing it"), subjects reported no change in stress if they had just been using the empathetic device, 4.0 (SD = 0.0) and an average increase in stress, 4.8 (SD = 0.75) if they had ended the experiment using the non-empathetic device. Note that these questions were all asked about the "device" in a general way, referring to subjects' overall experience after having used both systems.

Finally, after being told that the study was complete and that there would not be a session three, each user was asked, "How do you feel about the study being completed?" (7 = "Yay! This totally made my day", 1 = "Super bummed. I'll miss my little buddy"). The difference was again in the predicted direction, 4 (SD = 0.82) for those who had just finished using the empathetic version and 5 (SD = 1.4) for those who has just finished using the non-empathetic version.

4.2. *Relative subjective count*

Each user was asked to fill out their interrupt estimate in the nightly log. However, it was not surprising that subjects often forgot to fill out the log or lost their paper logs. Since the survey ended at 9:30 p.m. each night, subjects may not have been around a computer or their paper logs when the system told them that the survey was over for the day, and they may have forgotten to by the time they got home to fill out the logs. However, we did not find any differences among the conditions in whether or not people filled out logs. Of the 79 days of data, 47 days contained log reports: 25 for empathetic and 22 for non-empathetic. The average number of days each person filled out logs was 4.7, regardless whether they were in the group of seven for whom the system worked as expected or in the group of three for whom the empathetic system interruptions were misfiring. Thus, the number of logs filled out by subjects did not appear to be related to which system they were using, or to the number of interruptions received.

From the 47 logs, we computed the total number of perceived interruptions for each condition, 356 for empathetic and 309 for non-empathetic, and divided this by the actual number of interruptions for each condition, 514 for empathetic and 288 for non-empathetic. The resulting relative subjective counts were RSC = 0.69 for empathetic and RSC = 1.07 for non-empathetic. The difference between these is statistically significant, t (45), $p = 0.0007$, indicating that subjects underestimated the number of interruptions when they were using the

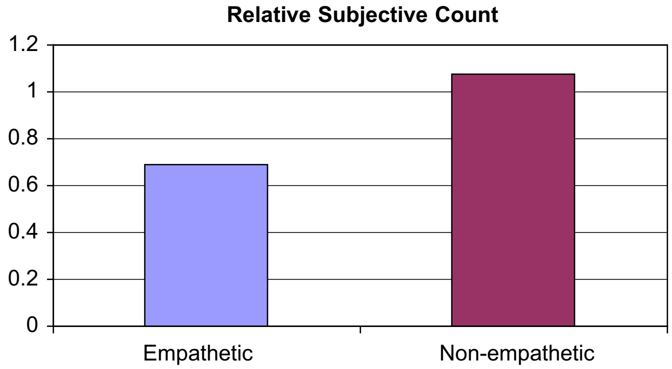


Fig. 10. Subjects ($N = 10$) underestimated the number of interruptions for the empathetic system, while overestimating those for the non-empathetic.

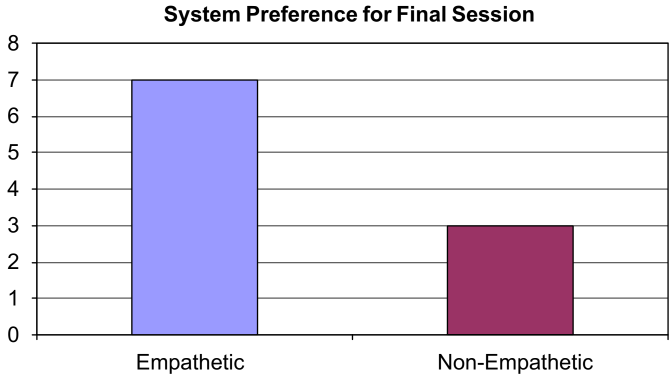


Fig. 11. Seven out of ten subjects chose to use the empathetic system for the third session.

empathetic version of the system (Fig. 10). It is possible, however, that RSCs have different meanings when there are around ten interruptions than when there are around 50. Thus, we repeated this comparison using only the seven subjects with the lower and relatively balanced number of interruptions. In this case, we find both RSCs are slightly reduced (RSC = 0.68 for empathetic and RSC = 0.96 for non-empathetic). The difference between RSCs for this set of seven subjects is again statistically significant t (31), $p = 0.014$, indicating that the difference in RSC continues to hold when the average number of interruptions is constant across conditions (Fig. 11).

4.3. Behavioral results

When asked, "Which system would you like to use for session 3?" ("System 1" or "System 2"), the empathetic system was selected over the non-empathetic by seven of the ten subjects, supporting Hypothesis 3. Of the three subjects who encountered the problems described earlier with the interruption timing, only one of them chose the non-empathetic system, while the two who actually had the highest number of interruptions chose the empathetic system despite that it interrupted them significantly more. Of the other two people who chose the non-empathetic system, one said, "Its interruptions were more predictable," despite that the interruptions were randomly triggered. The second said the empathetic version of the system "kept crashing." While it is tempting to blame their decision on that, we do not know if other subjects experienced crashing and just did not report it, or if they did not experience much crashing. Subjects were not asked directly about this, but all subjects were given daily opportunities to provide comments on the system with their logs, and no others reported problems with crashing. However, we are aware that some other users experienced technical difficulties with their systems (needing to restart typically because the heart rate sensor seemed to stop). We checked these reports and did not find any biases toward either system. However, we recommend that future investigations ask about performance to ensure there is no version-dependent problem. We also looked for a demographic or personality pattern among the three users who preferred the non-empathetic system and did not find any. Of these three, two were males and one female; two were graduate students—one from MIT and one from another Boston area college—and one was a non-student programmer.

5. Discussion and future work

Hypothesis H1 was that subjects would find the empathetic system to be less disruptive and frustrating to use and have a better user experience with it. While all our findings are with a small group, and thus any strong conclusions must await replication with a larger group, we did find all but one of the measures to be consistent with this hypothesis, with the exception being the perceived disruptiveness of the timing: the adaptive timing of the empathetic system was rated as more disruptive than the random timing of the non-empathetic system. Future work should be performed to decouple the interruption scheme from the empathetic response and assess both separately, and also make sure that the system fires properly for subjects having a range of body-mass index. One of the users who chose the non-empathetic system over the empathetic one commented that the latter one "seems to interrupt me more when I'm stressed." Since the heart-rate algorithm used to trigger interrupts could sometimes be set off by stress, this could be a general problem with the current algorithm.

We obtained converging evidence from the five questions we asked that compared the N-E group's responses to those of the E-N group, investigating if those who ended by using the empathetic system ended "on a better note." All the differences were in the predicted direction, with differences being strongest for the people who had just used the empathetic system indicating significantly greater interest in continuing to work with the stress awareness system. This kind of indirect evidence for the empathetic system fits with the pattern identified by Redelmeier and Kahneman in their medical experiments, whereby subjects who underwent an otherwise unpleasant medical experience remembered it more favorably if it ended better, even if overall there was more pain. Since experience-sampling methods are notoriously unpleasant, and empathy is supposed to help a person manage unpleasant feelings, we hypothesized that ending the experiment with use of the empathic system would be analogous to ending with less pain. Findings of the subject's reported experiences are consistent with this interpretation.

Our original plan was to analyze the physiological heart rate data jointly with the other collected data, to develop predictors for stress level based on a mixture of heart-rate variability and other activity and context variables. Problems with the hardware buffering, however, meant that critical timing information in the inter-beat interval data was not sufficiently precise to run a robust heart-rate variability estimator. We could and did confirm that average heart, as expected, was not overall related to reported stress level, even when trying to remove affects due to physical activity. Heart rate varies with each breath you take, and with many emotions and their associated breathing patterns, and is not specific enough to predict the five levels of stress that subjects reported.

With respect to H2—perceiving that the empathetic system interrupted them less than the non-empathetic system—we found confirmation from the newly proposed measure of relative subjective count, or RSC. The RSC was significantly underestimated for the empathetic version. On the one hand, we might argue that empathy does harm here because it makes a person's estimations less accurate. On the other hand, this oddity is to be expected given human experience that we are more likely to overlook the interruptions and annoyances of somebody who makes us feel less stressed, and to not overlook these problems when the individual leaves us more stressed. A system that imitates emotionally intelligent conversational moves, responding empathetically, is thus more likely to engender a willingness to overlook the annoyances it has brought about. Thus, we predicted that the more empathetic system would lead to fewer perceived interruptions, and this was the case. However, we should caution that this finding is an average effect, the number of subjects is small, and we only had 47 days of logs, so it remains to be confirmed in additional contexts with more people and over longer spans of time.

With respect to H3—subjects choosing to continue working with the empathetic system—we found behavioral confirmation via their system choice for session three. Behavioral choice is generally considered to be a stronger assessment than self-reported opinions when it comes to reflecting real human preference. While there were three dissenters, the convergence of the behavioral data with the direct and indirect subjective perceptions supports the conclusion that the empathetic system was generally preferred over the non-empathetic one. However, again we must caution that the 79 days of use of the two systems were only from 10 subjects, so it is important that such findings be investigated again for larger groups before strong general conclusions about the systems' merits are drawn.

Since the empathetic and non-empathetic systems differed in two ways—use of empathy, and use of sensor-based interruption timing—it is hard to know for certain which of these two influences gave rise to the preferences we saw. Given that the direct question about the disruptive nature of the interruption timing showed a leaning away from the empathetic system, while multiple other factors gave preference to the empathetic system as a whole, the improvement in user experience is more likely to be due to the only other difference between the two systems: the line of empathetic dialog, responding to the user's stress level. However, a future study that completely separates these two components would be important to confirm this.

5.1. Conclusions

This research has developed and assessed the world's first mobile stress-monitoring system that responds empathetically to the wearer's momentary report of stress level. The system allows for continuous, real-time user annotation of stress, activity and timing information through text and audio input on a mobile platform, interrupting the user an average of 11.5 times a day to sample activity, stress, and interruptibility. The platform supports continuous, wireless, and non-intrusive collection of heart signal data, accelerometer, and pedometer information, as well as automatic labeling of location information from context beacons. Future adaptations could also include camera and other sensor data, if desired. This system is the first of its kind to be affect and interruption-sensitive: it uses sensor data to adjust the timing of interruptions, and it adaptively responds with empathetic dialog tailored to specifically address the user's stress levels and the disruption the device may be incurring upon the user.

This paper has examined several kinds of assessment measures applied to the evaluation of two versions of the new interruptive stress-monitoring system. The assessments included both direct and indirect self-report measures and behavioral measures. The direct self-report measures included separate questions about each of the two systems such as "How responsive did you feel the system (A or B) was to your stress?" while the indirect self-report measures looked at influences of the most-recently used system on overall perceptions such as, "To what extent would you like to continue working with the Stress Awareness system?" We also measured behavior: which system a person chose to keep using after having used both. One of the new measures, relative subjective count, is proposed as a tool for indirectly examining frustration level related to technologies that involve frequent interruptions. This method is similar in spirit to the method of relative subjective duration, and thus may form part of a new category of tools for indirectly assessing frustration through the human tendency to under-estimate or over-estimate quantitative aspects of experience based on affect. However, so far the new measure has only been evaluated in one context—mobile interruptions of people in the United States—and over short periods of time (assessments spanning an average of eight 12 h days/subject.) Future work should investigate more contexts, diverse cultures, larger groups of people, and longer periods of time, to help determine if this is a truly general and useful new measure.

This work has shown converging findings from the indirect and direct self-reports, RSC, and behavioral measures, supporting a preference for a system that responds empathetically to one that does not, even when that system interrupted you an average of 17 times a day while the alternative interrupted you only an average of 11 times a day. Nonetheless, the sample size is small (79 days of data, 10 subjects) so there should be additional investigations with many more subjects, cultures, and contexts, before strong conclusions are formed. It would also be nice to repeat the experiments isolating the empathetic response (using identical interruption schemes for both conditions) as well as isolating the timing of interruption mechanism. While subjects did not seem to notice that one interruption process was random and the other was sensor-triggered (in fact, one liked the random better saying, "it was more predictable") it is still possible that the interruption timing strategies influenced the findings.

Finally, while the new assessment techniques presented in this work were only examined for this one new kind of affect-monitoring system, and thus the findings must be interpreted with caution, the applicability of the techniques we use is not limited to the system presented here. In particular, the strategy of examining RSC might be applied to any highly interruptive system. We can foresee this measure being applied in a variety of innovative technologies—especially in relational agents that interact with people regularly to motivate behavior change, e.g., at-the-moment encouragement from an exercise trainer, or reminding you that it is time to take medicine and confirming from you if you took it. Many systems that do not try to be social or polite, but simply bring regular interruptions for information (sales call logging, and more) would provide appropriate contexts for further examination of this measure. We thus offer this measure, and the other indirect assessments in this paper as possible new tools for indirect assessment of affect, allowing designers to avoid many of the problems associated with asking people directly how they feel. We encourage further examination of these techniques in larger groups and disparate contexts to illuminate how they relate to other measurements of affective experience and to behavioral preferences.

Acknowledgments

Thank you to Stephen Intille, Eric Horvitz, and Tim Bickmore for their insight and feedback for this work. We thank Tom Blackadar, Joe Wronski, and Bala Biyer at Fitsense Technology for their help customizing the hardware, and Lisa Feldman-Barrett and her group at Boston College for helpful insights into gathering emotional data through momentary assessment techniques. Thank you to Hyungil Ahn for performing the analysis of the physiological data. This research was supported in part by the Media Lab's Things That Think consortium.

References

Bickmore, T., Picard, R., 2005. Establishing and maintaining long-term human–computer relationships. ACM Transactions on Computer Human Interaction 12 (2), 293–327.

Czerwinski, M., Horvitz, E., Cutrell, E., 2001. Subjective duration assessment: an implicit probe for software usability. In: Proceedings of IHM-HCI, Lille, France.

Czerwinski, M., Horvitz, E., Wilhite, S., 2004. A diary study of task switching and interruptions. In: Proceedings of CHI, ACM Conference on Human Factors in Computing Systems, Vienna, Austria.

Dubin, D., 1996. Rapid Interpretation of EKG's, fifth ed. Cover Publishing Company, Florida.

Fitsense ⟨www.fitsense.com⟩, 2004.

The Surgeon General's Office, 1996. Physical Activity and Health: A Report of the Surgeon General. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Atlanta, GA.

Horvitz, E., Apacible, J., 2003. Learning and reasoning about interruption. In: International Conference on Multimodal Interfaces. ACM, Vancouver, BC.

Horvitz, E., et al., 2003. Models of attention in computing and communications: from principles to applications. Communications of the ACM 46 (3), 52–59.

Hudson, S.E., et al., 2003. Predicting human interruptibility with sensors: A wizard of Oz feasibility study. In: Proceedings of CHI.

Intille, S., et al., 2003. A context-aware experience sampling tool. In: Proceedings of the Conference on Human Factors and Computing Systems.

Intille, S.S., et al., 2003. Tools for studying behavior and technology in natural settings. UbiComp 2003: Ubiquitous Computing. Springer, Berlin.

Klein, J., Moon, Y., Picard, R.W., 2002. This computer responds to user frustration: theory, design, results, and implications. Interacting with Computers 14, 119–140.

Larson, R.W., Csikszentmihalyhi, M., 1983. The experience sampling method. New Directions for Methodology of Social and Behavioral Science 15, 41–56.

Liu, K.K., 2004. A Personal, Mobile System for Understanding Stress and Interruptions, in Media Arts and Sciences. Massachusetts Institute of Technology, Cambridge.

McEwen, B.S., Stellar, E., 1993. Stress and the individual: mechanisms leading to disease. Archives of Internal Medicine 153, 2093–2101.

Mohrman, D.E., Heller, L.J., 1991. Cardiovascular Physiology, third ed. McGraw-Hill, Inc., New York.

Mueller, F., Agamanolis, S., Picard, R.W., 2002. Exertion interfaces for sports over a distance. In: Proceedings of UIST, Paris, France.

Picard, R.W., Daily, S.B., 2005. Evaluating affective interactions: alternatives to asking what users feel. CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches, Portland, OR.

Picard, R.W., Healey, J., 1997. Affective wearables. Personal Technologies 1 (4), 231–240.

Prendinger, H., Ishizuka, M., 2005. The empathic companion: a character-based interface that addresses user's affective states. International Journal of Applied Artificial Intelligence 19 (3–4), 267–285.

Redelmeier, D.A., Kahneman, D., 1996. Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. Pain 66 (1), 3–8.

Reeves, B., Nass, C., 1996. The Media Equation. Cambridge University Press, New York.

Rondoni, J., 2003. A Context-Aware Application for Experience Sampling and Machine Learning. Massachusetts Institute of Technology.

Sapolsky, R., 1998. Why Zebras Don't Get Ulcers: An Updated Guide to Stress, Stress-Related Disease, and Coping. W.H. Freeman and Company, New York.

Stone, A.A., Shiffman, S., 1994. Ecological momentary assessment (EMA) in behavioral medicine. Annals of Behavioral Medicine 16, 199–202.

Strath, S.J., et al., 2002. Validity of the simultaneous heart rate-motion sensor technique for measuring energy expenditure. Medicine and Science in Sports and Exercise 34 (5), 888–894.

Stress in College: What Everyone Should Know, 2003. American College Health Association.

Walker, M., Consolvo, S., 2002. Experience sampling method for ubiquitous computing. Workshop on User-Centered Evaluation of Ubiquitous Computing Application, Ubicomp.