Towards *Personal Stress Informatics*: Comparing Minimally Invasive Techniques for Measuring Daily Stress in the Wild

Phil Adams[1], Mashfiqui Rabbi[1], Tauhidur Rahman[1], Mark Matthews[1], Amy Voida[2], Geri Gay[3,1], Tanzeem Choudhury[1], Stephen Voida[2]
[1]Information Science Department, Cornell University / [2]School of Informatics and Computing, Indiana University at IUPUI / [3]Communication Department, Cornell University
{pja22, ms2749, tr266, mjm672, gkg1, tkc28}@cornell.edu, {amyvoida, svoida}@iupui.edu

ABSTRACT
Identifying episodes of significant stress is a challenging problem with implications for personal health and interface adaptation. We present the results of a study comparing multiple modalities of minimally intrusive stress sensing in real-world environments, collected from seven participants as they carried out their everyday activities over a ten-day period. We compare the data streams produced by sensors and self-report measures, in addition to asking the participants, themselves, to reflect on the accuracy and completeness of the data that had been collected. Finally, we describe the range of participant experiences—both positive and negative—as they reported their everyday stress levels. As a result of this study, we demonstrate that voice-based stress sensing tracks with electrodermal activity and self-reported stress measures in real-world environments and we identify limitations of various sensing approaches.

Categories and Subject Descriptors
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous; J.3. [Computer Applications]: Life and Medical Sciences—Health.

General Terms
Measurement, Design, Reliability, Human Factors.

Keywords
Ubiquitous computing; stress; sensing; voice; electro-dermal activity; experience sampling; self-report; user modeling

1. INTRODUCTION
In the 2011 *Stress in America* survey, the American Psychological Association warned that stress is becoming a public health crisis [2]. Most Americans are suffering from moderate to high levels of stress, with nearly half reporting an increase in stress over the preceding five-year span. According to the APA, "job stress is estimated to cost U.S. industry $300 billion a year in absenteeism, diminished productivity, employee turnover and direct medical, legal and insurance fees" [3].

In general terms, stress is the reaction of an organism to a change in its equilibrium. In more practical terms, stress is the tension that a person experiences in response to an external stimulus or threat. Stress may have positive or negative affective outcomes, depending on whether or not a person is able to effectively cope with stress (see [4, 6, 34] for extensive discussions of how stress is operationalized and measured in various health informatics systems).

Longitudinal data about a person's stress levels (and the context(s) in which stress was measured) can be used to facilitate self reflection about patterns of stress embedded in daily routines or as caused by various environmental factors. Systems that promote this kind of data-driven self-reflection are often referred to as *quantified self* or *personal informatics* systems (e.g., [1, 14, 23, 24]). In addition, having knowledge about a user's stress level can be a valuable resource for adapting the interfaces of interactive computing systems or collecting data about the ways that adoption of a particular system affects levels of engagement, attention, or frustration in the real world [31].

Because of the highly subjective nature of perceived stress levels, researchers traditionally have relied upon self-report measures to gather data about people's experiences of stress. These techniques include diary studies (e.g., [1, 13, 34]) and *in-situ* experience sampling method (ESM) studies [22] (also known in some research sub-communities as *ecological momentary assessment* [30]). These approaches, while suitable for short-term research studies, present challenges when incorporated as part of a personal informatics system that is intended to provide benefits to its users over the long term. Diary studies are prone to memory effects and reduced compliance over time [1, 22], and experience sampling can be highly interruptive [17, 36] (which may, itself, become a source of stress for a study participant or a system user). Although the HCI community has developed adaptations to the ESM method, including delivery of surveys electronically and based on sensed contextual information (e.g., [17, 19, 28]), these approaches still require a considerable investment in time and effort, in order to provide insights about everyday stress and stressors over time. Although this style of data collection might be well suited to helping individuals to *discover* sources of stress within a particular time period, it would clearly not be as helpful when reflecting over an arbitrary window of time or when aiming to *maintain* an intended or desired response to stressors [24].

The increasingly pervasive sensing capabilities of our computational devices (cf. [29]) provide a valuable opportunity for continuously and non-intrusively measuring stress levels [1, 12, 13, 15, 31, 32, 34]. These devices are also being adopted by medical professionals and incorporated into long-term clinical treatments and behavioral interventions that are designed to improve healthcare outcomes [9]. However, because stress is a complex and multifaceted health issue, there are a variety of methodologies for automatically collecting data about people's levels of stress. Some

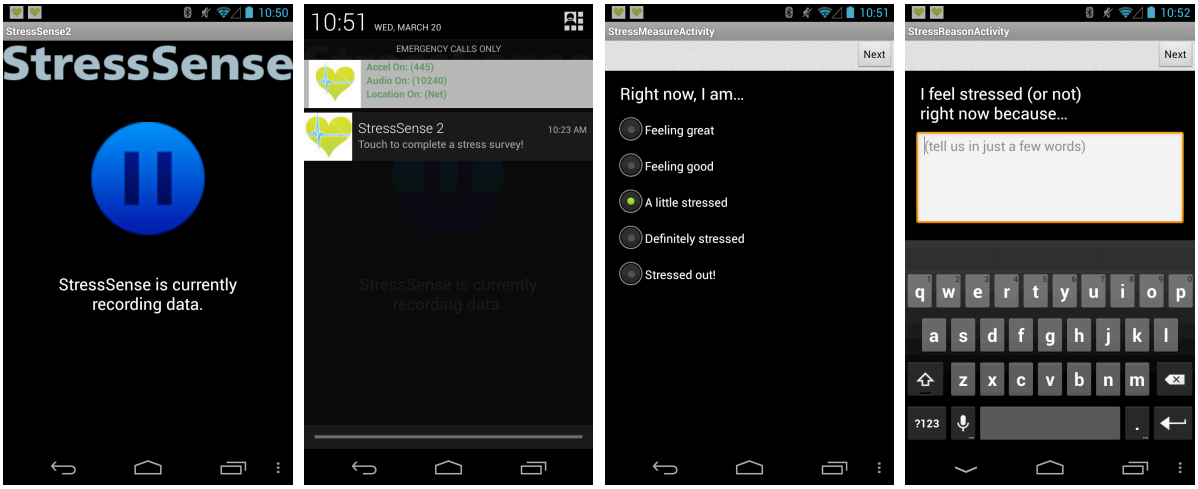

Figure 1. The interface of *SESAME.* Left to right: the control panel for starting and pausing the passive sensing (i.e., StressSense [25]); the notifications the app generates when it is time to complete a self-report survey; and the screens for self-report, including the Taylor single-item stress measure [37] and a free-text region to elicit a subjective rationale for the previous selection.

approaches, such as heart rate or heart rate variability, provide relatively direct and accurate measurements of stress, but come with undesired trade-offs in terms of the intrusiveness of measurement (e.g., [7]). Other approaches rely on secondary physiological signals, such as skin conductivity (referred to as electrodermal analysis or EDA), to detect changes in arousal that may be linked to stress [4, 12, 31, 33, 34] at the cost of some degree of fidelity (e.g., affective valence). Monitoring changes in vocal production [8, 25, 35] is a far less invasive approach but may be limited in the accuracy that can be achieved or the diversity of environments in which it is effective. Because different sensing modalities capture different representations, granularities, and quantities of stress data, most of the previous systems that have been proposed or developed for collecting stress data triangulate among different data collection methods, depending on whether they have aimed to create the most robust possible user model (e.g., [7, 8, 32]) or to explore issues related to sensor deployment and integration (e.g., [12]) or to design visual representations of stress to reflect back to end users (e.g., [13, 27, 34]).

In our research, we are interested in supporting long-term engagement with one's own perceived stress levels. One of the central challenges in creating these types of systems is in determining what kind of stress-related data to collect in order to strike a balance between *reliability*—that is, how closely the data accurately and consistently track a person's perceived stress at the moment that it occurs—and *intrusiveness*—that is, how much effort is required on a participant or user's behalf to provide the data. Previous research has focused on developing techniques for monitoring stress levels, often by triangulating among multiple data sources (e.g., [7, 8, 32]). In this paper, we present the results of a study designed to compare different kinds of data understand how we might collect these data with the minimal cost to the users. Among the kinds of sensing technologies that continuously monitor stress levels with a minimal impact on participants' daily lives, which track one another and participants' self report most closely? Under what conditions or in what contexts? When do these sensing modalities agree with one another, and when do they produce conflicting narratives about daily stress?

We designed and ran a study to answer these questions. Over the course of 10 days, we collected a variety of stress measures from a small group of participants during their everyday activities. We then compared the different data streams produced by traditional self-report measures and minimally invasive sensing devices (EDA and voice-based stress recognition). In addition, we conducted post-study interviews, asking the participants, themselves, to reflect on the accuracy and completeness of the data that had been collected. In this paper, we present several outcomes from this study that represent specific research contributions: (1) we present evidence that voice-based stress sensing tracks with variations in EDA and self-reported stress measures in real-world environments; (2) we describe the range of participant experiences—positive *and* negative—as they reported their stress levels; and (3) we reflect on some of the limitations of the various sensing approaches and the ways that our participants' experiences can help to inform the design of future *personal stress informatics* systems.

2. STUDY DESIGN

Our smartphone app, *SESAME* (*Stress Experience Sampling And Measurement Experiment*), was designed to collect data about individuals' stress levels and the environmental contexts within which this stress is experienced. It runs on the Android mobile operating system. Data is collected in the following three ways: (1) passively via sensors on the mobile phone, (2) via self-report measures also on the mobile device, and (3) via an Affectiva *Q Sensor*, worn on the wrist. The data collected on the smartphone device are cached locally on the device and pushed to the server in batches when the device is both plugged in to a power source and connected to a Wi-Fi network, most commonly during each night.

At short time intervals, SESAME infers an audio profile (*silence*, *non–human-voice noise*, *stressed voice*, *not stressed voice*) from microphone data. As recording audio can raise particular privacy concerns, the audio feature extraction and profile labeling takes place on the device—see previous work for more details (e.g. [25]). Due to limitations of the operating system, audio profile sensing is suspended when the participant makes or receives a phone call.

SESAME's user interface (Figure 1) is minimal, consisting of two icons in the system notification bar. Tapping the first icon presents the option to pause passive data collection. Tapping the second launches the self-report panel, which includes single-item measures assessing momentary stress and momentary affect, as well as an optional short free-text response to the prompt, "I feel stressed (or not) right now because….". To capture momentary stress we used Taylor's 5-item measure [37] that prompts "*right now, I am (1) feeling great! (2) feeling good (3) a little stressed (4) definitely stressed (5) stressed out!*" (Figure 1(c)).

We used the Affectiva device [33] to gather data about participants' electrodermal activity, which provides an indication of physiological arousal; as described above, this measure is associated with momentary stress. The Affectiva is worn on the inside of the wrist and is about the size of a wristwatch. Data gathered by an Affectiva was cached on that device and retrieved by the researchers at the conclusion of the study.

2.1 Experimental Protocol

Prior to the study period, each participant completed a short questionnaire. In addition to basic demographic information and questions about prior smartphone usage experience (including use of personal informatics [1, 23, 24]), participants reported traits for affect (PANAS) [38] and stress (PSS-14) [10], as well as a measure of mindful attention awareness correlated with a variety of wellbeing constructs (MAAS) [5].

During the day preceding the study, each participant was introduced to the SESAME app. After being trained on pausing/restarting sensing and responding to ESM prompts, participants were encouraged to make a handful of sample reports using the app and receive answers to any questions they might have. Because the continuous sensing and sense-making components of the system are computationally intensive and can drain a smartphone battery before the end of a full day, six participants used the system on a secondary, loaned phone (an LG Nexus 4), which they carried with them for the duration of the study. Over the next ten days, participants were asked to run the SESAME app between the hours of 8:00am and 11:00pm (at a minimum) and to make self-reports in response to notifications (issued every half hour with a small random variation) as they were able. Participants were also free to volunteer additional self-reports at any additional time. Because we were most interested in collecting *ground truth* data from our participants using ESM over a relatively short window of time, we opted to increase the prompt frequency to the upper end of what is typically considered acceptable practice [36] and to forego contextual suppression of stress reporting prompts (e.g., [17, 19]). Participants were informed that our goal was to collect as much stress data from them as was practical, but that they should feel free to ignore these frequent prompts if they occurred during attention-sensitive tasks like driving or if responding would be socially inappropriate (e.g., during a family dinner or while on a date).

Due to the limited number of Affectiva *Q Sensor* devices available for deployment during the study, we randomly divided our participants into two groups; participants in Group 1 collected EDA data on study days 1–5 and participants in Group 2 collected EDA data on days 6–10.

At the conclusion of data collection, we conducted a 20- to 30-minute semi-structured interview with each participant to gather qualitative information about their experiences using the SESAME app.

2.2 Participants

We recruited a small cohort of participants in person by convenience and snowball sampling. In an effort to hold stress profiles as constant as possible, we recruited only graduate students and postdoctoral researchers from within a single academic department at our institution. Participants received no compensation for participating in the study.

Of the original 11 participants recruited, n=7 completed the study; two participants chose not to continue beyond the first two days, and two others did not respond consistently to the self-reporting prompts during the data collection phase. Six of the seven participants who completed the study were male and one was female; six were aged 26–35 and one was aged 18–25. All but one participant owned and regularly used a smartphone (six Android, one iOS), and while six participants reported tracking personal information (such as sleep, exercise, spending, or mood) using websites or apps, no participant reported regularly using or wearing external sensors like the Nike FuelBand[1].

2.3 Analysis

Over the ten-day deployment, and with a pre-test questionnaire and a semi-structured interview, we have collected many types of data often at different temporal resolutions. Here, we describe how each form of data was preprocessed and then analyzed.

2.3.1 Data Preparation

Data from the pre-test questionnaire was prepared according to each instrument's directions (see [5, 10, 38]).

For both passively collected and self-reported data, we discarded samples outside of the time range dictated by the study parameters. To maintain consistency, participants were asked to use the system between the hours 8:00am and 11:00pm daily, but some used the system outside this time range (e.g., making self-reports early in the morning or wearing the Affectiva late at night).

We inferred physiological arousal based on the EDA data provided by the Affectiva *Q Sensor*. In stressful situations, the sympathetic nervous system activates the sweat glands. EDA devices like the *Q Sensor* estimate the amount of sweat secreted by measuring changes in the electrical conductance of the skin [33]. Once we retrieved the raw EDA data from the wrist-worn devices, we normalized the time-series data using a Z-normalization technique, which centers the data distribution about a zero-mean and scales it, resulting in unit-variance. We used a 20-second long window with 10-second shift to extract high-level features from the normalized EDA data. EDA has a fast-changing response (startle response) when stressors are present, so some of the features that we extracted include the mean crossing rate, the energy associated with low-frequency filter bank, the slope of the linear regression, and minimum and maximum values, all of which have been shown to capture aspects of this startle response [16]. We trained two single-component Gaussian mixture models (one for modeling the *aroused* condition and another for modeling *neutral* or *non-aroused* conditions) with a full covariance matrix based on a . We then used a GMM-based model to classify the EDA data into a binary value—*aroused* or *not aroused*—every 10 seconds over 20-second windows, using a threshold of .85. (On the pre-existing dataset, this GSR-based stress model has a performance of 74.3%, 76.5% and 78.4% in terms of accuracy, recall and precision, respectively.)

Continuously sensed audio yielded inferences every 1.2 seconds. To aid in comparison across measures, we elected to smooth all passively sensed inferences into 5-minute windows using a simple majority rule; smoothing to tighter (1-minute) windows did not appreciably change the results.

2.3.2 Comparisons over Modalities

One of the central questions that this research seeks to address is the reliability of various widely deployable stress-sensing techniques in a range of real-world scenarios. It is clear that some of these approaches will be more suitable in detecting stress in certain situations. For example, recognizing stress from voice will

[1] http://www.nike.com/cdp/fuelband/

Table 1. A summary of the experience sampling-driven self-report data provided by our study participants.

Participant ID *(EDA Group)*	Number of Stress Self-reports Completed	Completed Reports to Reminders Sent Ratio	Avg. Delay in Self-report Response (min:sec)	Avg. Self-reported Stress Level (5=high, 1=low)
1 *(early)*	62	38%	6:17	2.66
2 *(early)*	88	44%	7:53	2.09
3 *(early)*	128	13%	7:09	1.82
4 *(late)*	173	36%	7:49	1.87
5 *(late)*	99	10%	19:22	1.90
6 *(late)*	162	9%	6:45	2.86
7 *(late)*	172	46%	7:02	2.32
Averages	**126.3**	**28%**	**8:54**	**2.22**

necessarily be more accurate when a person is engaged in a conversation than when they are working alone; the EDA signal will change in different ways during physical exercise than when a person is experiencing emotional or cognitive stresses [33]. Furthermore, while subjective self-assessment of stress levels (e.g., with the PSS-14 instrument [10]) has been shown to have high internal consistency and predictability, many of these types of instruments have been designed to examine stress as a trait, framing stress in the context of life events that take place over the course of weeks or months. Since there is no clear-cut and established gold standard for globally measuring stress levels in an ecologically valid, non-intrusive way, we set out to determine the circumstances in which a variety of established stress measuring mechanisms, including EDA, continuous voice-based stress recognition, and self-reported stress levels, do and do not align with each other.

In comparing self-report measures, captured at a resolution of 30 minutes, inferences from continuously sensed sources were smoothed over 1 hour windows prior to the self-report. This is because the self-report stress literature indicates that momentary psychological stress is a function of the current stress trait, and recent daily stressors; further, we confirmed that this is how our participants made self-report assessments from the semi-structured interviews. Because data most closely co-occurring with the self-report will have a greater effect on experienced stress, we computed a weighted mean over the data in the window, giving preference to the most recent data points; again, feature selection was by simple majority. This happens for each self-reported value for each user; we then normalize and assess the relationship for each self-report value, one to five, with both electrodermal and voice-stress inferences.

2.3.3 Interview data

We referred to the semi-structured interview data to help make sense of discontinuities in the sensor data streams and to inform our understanding of the participants' experiences using the system, including how and when participants provided data over the various modalities that we used.

3. RESULTS AND DISCUSSION

3.1 General Participant Experience / Use

Over the course of the study, 15 hours per day for 10 days, SESAME collected a significant amount of data about stress and stress contexts experienced by our 7 participants. The system recorded some 17,415,310 audio profile inferences, 884 self-reports, 56,837 location measurements, and 9,400,139 EDA measurements. Smoothing and binning the data into manageable windows (as described above) yielded, on average, 1,192 location measurements, 1,066 audio profile inferences, 126 self-reports, and 15,368 *aroused* or *not aroused* inferences from the EDA data *per participant*.

The number of stress self-reports completed over the course of the study (a ~40% response rate) had dramatic variance (1,982), with several participants [P4, P6, P7] each providing more than 150 responses (Table 1). Most of the self-report survey submissions appear to have been direct responses to SESAME's experience sampling prompts—appearance of an auxiliary status bar icon, a short vibration sequence, and the notification LED set to pulse a purple color. However, most participants voluntarily submitted at least a few instances of un-prompted self-reports, with P2 submitting the largest number (11).

During the study, the smartphone application was programmed to issue the participants a reminder to complete the experience-sampling questionnaire approximately once every 30 minutes. In practice, many experience-sampling responses were delayed due to constraints of the threading model used by the operating system or an app crash, or the participants simply did not respond to the notifications. There were a number of reasons why this may have been the case: the vibration pattern associated with the notification was somewhat subtle on some of our participants' phones (particularly on the Nexus 4s), and participants told us during our post-study interviews (see also below) that they would voluntarily ignore the notifications if they were engaged in an activity that demanded their attention (e.g., a conversation or driving) or if their hands were otherwise occupied (e.g., cooking, playing with children). The algorithm that generated the experience sampling notifications was also linked directly to the system timer, rather than being driven by the last time that a self-report survey was submitted. This resulted in a number of occasions in which a participant would notice that they had neglected to complete a survey in response to a prior reminder, submit the survey, and immediately be notified that it was time to complete another; many of these subsequent self-report reminders were ignored.

On average, our participants completed surveys a little less than 1/3 of the time that an experience sampling notification was issued. Compliance ranged from 9% to nearly 50% across our group of participants. For those self-reported surveys that were completed in response to an experience sampling notification, the average delay between the system issuing an experience sampling reminder and the participant invoking the self-report survey mechanism was 8 minutes, 54 seconds (excluding delays longer than 30 minutes, which indicated an error or crash in the application). There were a fairly large number of very quick responses in our dataset (as little as 3 seconds elapsed from notification to invocation of the survey), but much of the time, the participants

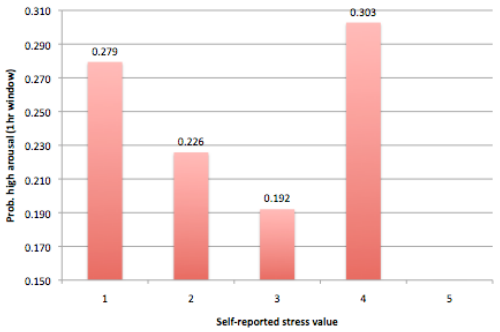

Figure 2. The probability of *aroused* EDA in a 1-hour window associated with self-reported stress.

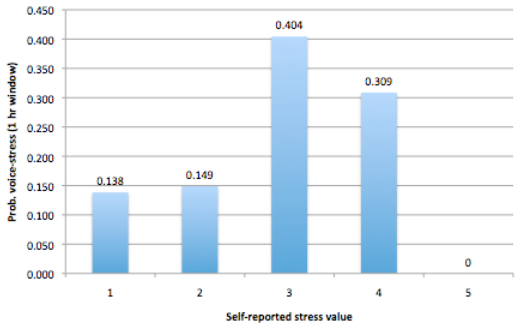

Figure 3. The probability of *voice-stress* presence in a 1-hour window associated with self-reported stress.

simply did not or could not respond until tens of minutes had elapsed. The only real drawback to this response rate is that our self-report data is scattered somewhat unevenly over our data collection window, as our analysis examined the passively sensed stress rates at whatever time the self-report surveys were completed. Interviews suggest that periods of time associated with very high levels of stress are under-reported in the self-report data, leading an artificial downward skew of the self-reported stress levels.

Although participating in the study did result in frequent interruptions due to our use of experience sampling to collect self-report data, participants were able to complete these three-question surveys quickly, with an average start-to-finish time of 20 seconds and a median of 17 seconds; this excludes 5 outliers over 5 minutes where the survey activity appears to have been interrupted by another smartphone function.

We also ran a number of paired *t*-tests in order to determine whether participants' responses to the pre-test surveys (MAAS, I-PANAS-SF, or PSS-14) were effective predictors for the average stress levels reported over the course of the study. We found no significant differences that would suggest a relationship, although we did observe a very weak trend (p=.112) suggesting a correlation between participants' score on the PSS-14 and their average self-reported stress levels across the 10 days of the study (r=.562). A study with a larger sample size would be needed to more rigorously assess the predictive power of PSS-14 in empirically determining a per-person baseline stress level.

3.2 Comparison of Modalities

3.2.1 Scenarios and Stress

We anticipated that each capture modality would, for individuals and in aggregate, reflect similar daily stress rhythms, and this bore out in our data. Voice-based stress measures were most pronounced on weekdays in the early and mid afternoon. There is a second, smaller peak in the late morning; detection of voice stress before mid-morning or after dinnertime is rare. Voice stress is prevalent for several users on one or two particular days (e.g., P4's day 10 and P6's day 3). EDA data also peaks in the early- to mid-afternoon, overall, and dramatically so for participant P3. Overall, self-reported stress remains relatively constant throughout the day, although there are fewer reports of "A little stressed" once the evening begins. Of note is that not a single participant self-reported the most stressed value "Stressed out!" over the course of the study. This unexpected gap in the data highlights one of the key shortcomings of the traditional self-reporting approach to tracking stress levels: in situations where a participant is experiencing the highest levels of stress, they are extremely unlikely to stop what they're doing and attend to an experience sampling prompt. This reporting bias has been noted previously [36], but is of particular interest when the construct being investigated relates so closely to a major factor in survey response compliance.

The sound profile that appeared in conjunction with experiences of stress is also interesting. The audio classification corresponding to the times that self-reports were given was most commonly *noise* or *silence*. P2 and P4 additionally reported feeling "good" when there was unstressed voice present on multiple occasions. On four devices, the audio capture silently failed occasionally, and so self reports, particularly for P1 and P7, were also made at times for which we have no noise profile information; we have no reason to believe this occurred in periods of particularly high or low perceived stress and should not therefore bias our findings.

3.2.2 Associating Stress from the Three Modalities

In investigating whether the minimally invasive passive measures (voice-based stress detection and EDA) could be reliably used to monitor stress levels in the wild, we report the voice-stress and EDA data collected simultaneously with self-reported stress, as well as with one another. Not all self-report points have associated EDA or voice-stress data, because (1) each participant wore the *Q Sensor* for only for half of the study days; (2) as a result of the occasional audio capture crash, as described above, there are some windows of time lacking raw audio data upon which to draw voice-based stress inferences; and (3) there were occasions when participants chose to disable the passive sensors, such as when swimming, washing dishes, or to conserve the phone battery.

3.2.2.1 Self-report Stress with EDA

Over the self-report values 1–4, we see a cup-shaped curve, as responses 1 ("feeling great!") and 4 ("definitely stressed") correspond to higher levels of arousal than responses 2 ("feeling good") and 3 ("a little stressed"). This distinction emerges most clearly at higher EDA classification thresholds; here we report with a threshold of .8, selected by experimentation (Figure 2).

These data confirm that EDA provides an indication of the intensity of perceived stress responses, that is, the more strongly a participant agrees or disagrees that they are under stress, the stronger the EDA signal recorded by the system. However, the main drawback of the EDA approach is also highlighted here—it is statistically impossible to detect from EDA data alone the valence of the perceived stress response, that is, whether increased arousal is associated with pleasant experiences or the negative experiences that we typically associate with being under duress.

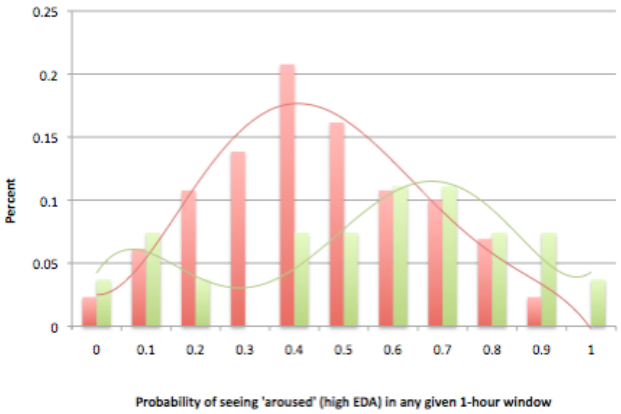

Figure 4. Distribution of *aroused* EDA, both overall (in red) and when *voice-stress* is detected (in green).

3.2.2.2 Self-report Stress with Voice-stress

In associating self-reported stress values with voice-stress, we report a weighted mean of voice-stress (1) and voice-no-stress (0), also computed over a 1-hour window (Figure 3). There is a positive correlation of r=.59 over the self-report values 1–4, which we anticipated. However, the relationship peaks not at 4 ("definitely stressed") but at 3 ("a little stressed"); we believe this is because participants self-reported much less frequently when experiencing higher levels of stress.

This positive correlation revealed in our data suggests that analysis of passively collected voice signals can result in reasonably accurate detection of stress episodes without requiring any participant/user intervention at all. However, the success of this approach is clearly dependent on the presence of a clear voice signal; stressful situations in which vocalizations are not present cannot be inferred at all.

3.2.2.3 Voice-stress with EDA

In Figure 4, we examine the distribution of EDA aroused/non-aroused responses (threshold=.75, selected by experimentation) both independent of, and in the presence of, positive voice-stress recognition. The red bars show the distribution of arousal inferred from EDA in one-hour windows over the entire study. For example, in about one-fifth (y-axis) of the windows, we detected a state of high arousal for 40% (x-axis) of the reports collected during that window. Similar to a histogram, the relatively normal distribution of the red bars illustrate that it is relatively common to observe a relatively even mix of aroused and non-aroused readings from the EDA sensor during any given hour; it is more unusual for a one-hour window to be dominated by aroused or non-aroused inferences.

The green bars also show the distribution of EDA arousal in one-hour windows, but in only those windows in which we also inferred the presence of voice stress for the user in question. The probability mass shifts rightward when we examine only these EDA reports that correspond to sensed stress from the voice channel. This shows that EDA and voice-based stress recognition generally track one another in the positive case; that is, over a one-hour window, when we observe consistent stressed responses using the voice-based recognizer, we are also more likely to see aroused inferences from the EDA sensor during this time.

The small rise at the left end of the green bar distribution is worthy of note, however. This artifact reveals that there were a small (but noticeable) number of one-hour windows in which the EDA sensor was resulting in a much greater number of non-aroused inferences than aroused inferences, but the voice-based stress recognizer was detecting instances of stressed voice. When triangulating raw sensor and location data with interviews, it appears that this disagreement is a result of P6, a teaching assistant, calmly listening to several hours of student presentations, and P7 relaxing at home while eating breakfast or dinner while watching action movies on TV. In both cases, the system was (correctly) detecting stressful vocalizations, just not ones generated by the participants themselves. Adding a speaker ID filter to the system would mitigate these instances of false positives.

3.3 Semi-Structured Interview Feedback

Some of the frustrations that participants reported included the troublesome drain on the phone battery caused by our implementation of continuous passive sensing and a concern that being repeatedly asked to consider and report on one's stress could itself become a stressor—P7 even went so far as to "mentally rename the app *Keep Calm*" so that he could continue participating without feeling overwhelmed. Although, P7 also felt that "sometimes the device felt like a box I could put my stress in, and move on." Several participants agreed with P4 that even through the self-report prompts occurred very frequently, the interaction required to acknowledge the prompt and complete the associated survey was "very light."

The self-report stress measure was criticized in two ways. First, participants felt that there was need for a "not stressed and not unstressed" option between "feeling good" (2 on a 5-point Likert item) and "a little stressed" (3 on a 5-point Likert item). Second, two participants agreed with P6 who wanted greater resolution of the scale:

> *I would compare it to previous times I had answered, and I would feel a little more stressed, but not enough to take it to the next level, so I'd put the same as last time. [P6]*

The experience of wearing the *Q Sensor* during the study was widely reported to be "easy," although one participant found the strap too tight and another commented that the device kept slipping out of contact with her skin. The Affectiva was the only visible aspect of participation in the study and it did spark some conversation about the device, the study, and the stressors that had been experienced (and noted) by the participants during the study.

Participants did not widely report privacy concerns as a result of participation in the study. P6 explained that any privacy concerns were mitigated by the potential reflective benefit of such a tool: "The question [of whether there are privacy issues implicated in use of the system] is, do you believe in the values of the app? Here, I feel like the app is helping me be aware of and control my stress." P5 noted that while she had few concerns since she was informed about and understood what the system was doing, her friends became uncomfortable when they learned that audio sensing was taking place via her smartphone's microphone; friends of P2 expressed similar concerns. This finding illustrates one of the more complex trade-offs in the design of systems like SESAME: even when approaches can be developed to sense stress without directly burdening the user (e.g., voice-based stress recognition), there may be tacit social costs implicated in these decisions.

Table 2. An overview of stress detection methods explored in this study and their contexts of use.

Data Collection Method	Measures	Effective contexts	Less-effective contexts
Experience sampling-driven self-report [22, 30]	Collection of the Taylor single-item stress measure [37], an open-ended rationale for stress level, and affect	Scenarios in which the user's subjective perception of stress is valuable; provides (potentially) finer descriptive resolution	When interruptions are not desirable (e.g. work, driving, social engagement); interruptions may adversely influence the user's stress level
Electrodermal analysis (Affective *Q Sensor* [33])	Physiological arousal via skin conductivity	Most day-to-day contexts; most valuable when contextual valence already known	Physical discomfort of or preferences against wearing device; expense limits scaling of participant pool
Voice-based stress analysis	Variation in vocal characteristics (pitch, speaking speed, vocal energy)	Many day-to-day contexts in which user will be regularly speaking; where on-device feature extraction possible	Ineffective in quiet or noisy spaces; currently only provides coarse metrics

The use of the microphone in a smart phone as a sensor for capturing voice, while effective in the lab [26], had various consequences in the wild. First, multiple participants reported unexpected holes in the audio profile data with respect to voice-stress. Because SESAME could not access the microphone during phone calls, phone conversations with remote friends and family, not unusually an opportunity to talk through experiences of stress, were not captured. Secondly, even though the device is not actually recording audio files, there is a sense among a few participants, and others around them, that this might be a privacy concern. Allowing users to censure sensed data after the fact has been effective in other physiological sensing systems (e.g., [18]); more control over personal sensed data could also help alleviate privacy concerns in SESAME. Third, the voice-stress classifier used in SESAME did not learn, so using an adaptive voice stress recognition algorithm could improve the classification.

Although our data collection app also recorded location and activity during the study, we do not report on the relationship between location or activity and the other sensed/reported data here. Future analysis will consider the automated recognition of physical regions of interest and their association with stress levels, as well as the potential for using accelerometer-based activity recognition for identifying stress or increasing the robustness of voice- or EDA-based approaches.

While we attempted to control for stress profiles and daily stressors for this study, it is possible that characteristics of our participant population impacted the results. For example, much of their work takes place collaboratively in spaces relatively free of background noise. Our results should be confirmed in larger, more diverse populations.

4. CONCLUSION

We examined minimally intrusive mechanisms for measuring, inferring, and eliciting characterizations of a user's stress. We demonstrated that voice- and EDA-based classifiers produce representations of stress that correlate with self-reported measures of perceived stress and with one another in real-world environments. As a result of this research, we identified contexts in which the various sensing approaches are more or less effective (Table 2).

We found that self-report remains the mechanism through which the most accurate representations of *low* and *moderate* levels of stress can be collected from participants, as well as the only mechanism that can be easily augmented to understand the source of stress; however, contextual augmentation could be provided in the personal informatics reviewing interface. Our results indicate that self-report is also useful for validating or correcting stress models constructed based on automatically sensed data. The main drawbacks of self-report are its intrusiveness, which might be mitigated through the use of context-aware prompting (e.g. [17, 19]) and the fact that the method is still unlikely to reveal episodes of intense stress, simply because users can choose to ignore experience sampling prompts during those experiences.

Based on our empirical data, EDA- and voice-based stress recognition both provide less invasive yet still reasonably robust representations of stress in real-world environments; certainly when these two channels agree we can be confident the user is experiencing stress. Further research is needed to develop robust sensing of EDA or the intensity of stress using only those sensors users already carry with them. And future work will also need to address a number of weaknesses of voice-based stress sensing that we have identified, such as a higher incidence of false positives and the potential for raising privacy concerns. Several of these false positives can be mitigated by adding a speaker ID filter to the system, while affording user censure of sensed data (e.g., [11]) should continue to alleviate privacy concerns already partially addressed by the system design.

Stress is a factor in so many facets of health and wellbeing. Our study provides an encouraging starting point for informing the design of minimally invasive ubicomp systems for sensing stress.

5. ACKNOWLEDGEMENTS

We would like to thank our participants and the anonymous reviewers for their thoughtful feedback. This work is partially supported by NSF GRFP (DGE-1144153), Swiss NSF Sinergia grant, NSF IIS#1202141 and the Intel Science and Technology Center for Pervasive Computing.

6. REFERENCES

1. Almeida, D.M. Resilience and Vulnerability to Daily Stressors Assessed via Diary Methods. *Current Directions in Psychological Science 14*, 2 (2005), 64–68.
2. American Psychological Association. *Stress in America: Our health at risk*. APA (2012). http://www.apa.org/news/press/releases/stress/2011/final-2011.pdf.
3. APA Practice Directorate. *Creating a psychologically healthy workplace*. http://www.phwa.org/resources/creatingahealthyworkplace/
4. Ayzenberg, Y., Hernandez Rivera, J., and Picard, R. FEEL: Frequent EDA and event logging – A mobile social interaction stress monitoring system. *Ext. Abstracts CHI '12*, ACM (2012), 2357–2362.
5. Brown, K.W., and Ryan, R.M. The benefits of being present: Mindfulness and its role in psychological well-being. *J. Personality and Social Psychology 84*, 4 (2003), 822–848.

6. Campbell, J. and Ehlert, U. Acute psychosocial stress: Does the emotional stress response correspond with physiological responses? *Psychoneuroendocrinology 37*, 8 (2012), 1111–1134.
7. Carbonaro, N. et al. Wearable biomonitoring system for stress management: A preliminary study on robust ECG signal processing. *Proc. WoWMoM '11*, IEEE Computer Society (2011), 1–6.
8. Chang, K., Fisher, D., Canny, J., and Hartmann, B. How's my mood and stress?: An efficient speech analysis library for unobtrusive monitoring on mobile phones. In *Proc. BodyNets '11*, ICST (2011), 71–77.
9. Chatterjee, S. and Price, A. Healthy living with persuasive technologies: Framework, issues, and challenges. *J. American Medical Informatics Association, 16*, 2 (2009), 171–178.
10. Cohen, S., Kamarck, T., and Mermelstein, R. A global measure of perceived stress. *J. Health and Social Behavior, 24* (1983), 385–396.
11. Epstein, D.A., Borning, A. and Fogarty, J. Fine-grained sharing of sensed physical activity: A value sensitive approach. *Proc. UbiComp '13*, ACM (2013), 489–498.
12. Ertin, E., Stohs, N., Kumar, S., Raij, A., al'Absi, M. and Shah, S. AutoSense: Unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field. In *Proc. SenSys '11*, ACM (2011), 274–287.
13. Ferreira, P., Sanches, P., Höök, K., and Jaensson, T. License to chill!: How to empower users to cope with stress. *Proc. NordiCHI '08*, ACM (2008), 123–132.
14. Freedman, D.H. The Perfected Self. *The Atlantic*, 2012. http://www.theatlantic.com/magazine/archive/2012/06/the-perfected-self/308970/.
15. Gaggioli, A., Pioggia, G., Tartarisco, G., Baldus, G., Corda, D., Cipresso, P. and Riva, G. A mobile data collection platform for mental health research. *Personal and Ubiquitous Computing 17*, 2 (2013), 241–251.
16. Healey, J.A. and Picard, R.W. Detecting stress in real-world driving tasks using physiological sensors. *Intelligent Transport Systems, 6*, 2 (2004), 156–166.
17. Intille, S.S., Rondoni, J., Kukla, C., Ancona, I. and Bao, L. A context-aware experience sampling tool. *Ext. Abstracts CHI '03*, ACM (2003), 972–973.
18. Kay, M., Choe, E.K., Shepherd, J., Greenstein, B., Watson, N., Consolvo, S. and Kientz, J. Lullaby: A capture & access system for understanding the sleep environment. *Proc. UbiCom*p '12, ACM (2012), 226–234.
19. Klasnja, P., Harrison, B.L., LeGrand, L., LaMarca, A., Froehlich, J., and Hudson, S.E. Using wearable sensors and real time inference to understand human recall of routine activities. *Proc. UbiComp '08*, ACM (2008), 154–163.
20. Korhonen, I. and Bardram, J.E. Guest editorial introduction to the special section on pervasive healthcare. *IEEE Trans. Info. Tech. in Biomedicine, 8,* 3 (2004), 229–234.
21. Lane, N.D., Mohammod, M., Lin, M., Yang, X., Lu, H., Ali, S., Doryab, A., Berke, E., Choudhury, T., and Campbell, A.T. BeWell: A Smartphone application to monitor, model and promote wellbeing. *Proc. Pervasive Health '11* (2011).
22. Larson, R. and Csikszentmihalyi, M. The experience sampling method. *New Directions for Methodology of Social & Behavioral Science*, *15* (1983), 41–56.
23. Li, I., Dey, A., and Forlizzi, J. A stage-based model of personal informatics systems. *Proc. CHI '10*, ACM (2010), 557–566.
24. Li, I., Dey, A., and Forlizzi, J. Understanding my data, myself: Supporting self-reflection with ubicomp technologies. *Proc. UbiComp '11*, ACM (2011), 405–414.
25. Lu, H., Frauendorfer, D., Rabbi, M., Mast, M.S., Chittaranjan, G.T., Campbell, A.T., Gatica-Perez, D., and Choudhury, T. StressSense: Detecting stress in unconstrained acoustic environments using smartphones. *Proc. UbiComp '12*, ACM (2012), 351–360.
26. Lu, H., Yang, J., Liu, Z., Lane, N.D., Choudhury, T., and Campbell, A.T. The Jigsaw continuous sensing engine for mobile phone applications. *Proc. SenSys '10*, ACM (2010), 71–84.
27. McDuff, D., Karlson, A., Kapoor, A., Roseway, A., and Czerwinski, M. AffectAura: An intelligent system for emotional memory. *Proc. CHI '12*, ACM (2012), 849–858.
28. Meschtscherjakov, A., Reitberger, W., and Tscheligi, M. MAESTRO: Orchestrating user behavior driven and context triggered experience sampling. *Proc. Measuring Behavior '10*, ACM (2010), 29:1–29:4.
29. Miller, G. The smartphone psychology manifesto. *Perspectives on Psychological Science, 7*, 3 (2012), 221–237.
30. Moskowitz, D. and Young, S. Ecological momentary assessment: What it is and why it is a method of the future in clinical psychopharmacology. *J. Psychiatry and Neuroscience 31*, 1 (2006), 13–20.
31. Picard, R.W. and Liu, K.K. Relative subjective count and assessment of interruptive technologies applied to mobile monitoring of stress. *IJHCS 65*, 4 (2007), 361–375.
32. Plarre, K., Hossain, S.M., Ali, A.A., Nakajima, M., al'Absi, M., Ertin, E., Kamarck, T., Kumar, S., Scott, M., Siewiorek,D., Smailagic, A. and Wittmers, Jr., L.E. Continuous inference of psychological stress from sensory measurements collected in the natural environment. *Proc. IPSN '11*, IEEE Computer Society (2011), 97–108.
33. Poh, M-Z., Swenson, N.C., and Picard, R.W. A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *IEEE Trans. Biomed. Engineering, 57*, 5 (2010), 1243–1252.
34. Sanches, P., Höök, K., Vaara, E., Weymann, C., Bylund, M., Ferreira, P., Peira, N. and Sjölinder, M. Mind the body!: Designing a mobile stress management application encouraging personal reflection. *Proc. DIS '10*, ACM (2010), 47–56.
35. Scherer, K.R. Voice, stress, and emotion. In Appley, M.H. and Trumbull, R. (Eds.), *Dynamics of stress: Physiological, psychological, and social perspectives.* Plenum Press, New York (1986), 157–179.
36. Scollon, C. N., Kim-Prieto, C. and Diener, E. Experience Sampling: Promises and Pitfalls, Strengths and Weaknesses. *J. Happiness Studies 4*, 1 (2003), 5–34.
37. Taylor, S.E., Welch, W.T., Kim, H.S., and Sherman, D.K. Cultural differences in the impact of social support on psychological and biological stress responses. *Psychological Science 18*, 9 (2007), 831–837.
38. Watson, D., Clark, L.A., and Tellegen, A. Development and validation of brief measures of positive and negative affect: The PANAS scales. *J. Personality and Social Psychology 54*, 6 (1988), 1063–1070.
39. Wolf, G. Know thyself: Tracking every facet of life, from sleep to mood to pain, 24/7/365. *WIRED, 17*, 7 (2009), 92–97.