Wearable ESM: Differences in the Experience Sampling Method across Wearable Devices

Javier Hernandez1, Daniel McDuff2, Christian Infante2, Pattie Maes1, Karen Quigley3, Rosalind Picard1

Media Lab, Massachusetts Institute of Technology, Cambridge, United States1,2
Psychology, Northeastern University, Boston, United States3
{javierhr, djmcduff, pattie, picard}@media.mit.edu1, cinfante@mit.edu2, k.quigley@neu.edu3

ABSTRACT
The Experience Sampling Method is widely used for collecting self-report responses from people in natural settings. While most traditional approaches rely on using a phone to trigger prompts and record information, wearable devices now offer new opportunities that may improve this method. This research quantitatively and qualitatively studies the experience sampling process on head-worn and wrist-worn wearable devices, and compares them to the traditional “smartphone in the pocket.” To enable this work, we designed and implemented a custom application to provide similar prompts across the three types of devices and evaluated it with 15 individuals for five days (75 days total), in the context of real-life stress measurement. We found significant differences in response times across devices, and captured tradeoffs in interaction types, screen size, and device familiarity that can affect both users’ experience and the reports made by users.

Author Keywords
Experience Sampling Method; Ecological Momentary Assessment; wearable devices; smartwatch; smartphone; smart-eyewear; Google Glass.

ACM Classification Keywords
H.5.2. Information interfaces and presentation: User Interfaces.

INTRODUCTION
The Experience Sampling Method (ESM) [15, 19], also known as the Ecological Momentary Assessment (EMA) [6, 22], is a commonly used methodology to gather repeated responses from people during their daily activities. While some uses of ESM date back almost 40 years [12], continuous advancements in technology have enhanced the approach and contributed to its common usage in real-life studies. For instance, ESM has been successfully used to better understand real-life emotional experiences, social interactions, time productivity, and personality [9]. In comparison with traditional techniques such as surveys and diaries, ESM relies on asking questions frequently during daily life to minimize retrospective recall biases [4]. ESM also minimizes the amount of interactions required with researchers during data collection, which helps reduce the response biases that their presence can cause.

The introduction and widespread use of consumer wearable devices, such as Google Glass, has created new opportunities to enhance ESM. As an example, wrist and head-worn devices are usually more accessible than a phone and offer a unique opportunity to minimize the disruption associated with ESM. However, new interaction challenges arise with the novelty and unfamiliarity of these devices and their interfaces. This work compares the use of three types of wearable and mobile devices during ESM (shown in Figure 1). In particular, we developed a new Android ESM tool that can be similarly used on different types of devices (smartwatches, head-mounted devices, and smartphones) and quantify how the device form factor can significantly impact relevant aspects of the reporting process (e.g., missed interruptions, response times, and content of report).

In the remainder of the paper we: 1) review previous research and existing challenges of the ESM, 2) review advantages and disadvantages of considering different locations for ESM, 3) describe the new ESM tool we developed, the interactions, and prompting criteria,
4) describe the design of our experiment, 5) provide quantitative and qualitative results for the three devices, 6) discuss the results, and 7) elucidate the devices advantages and remaining challenges.

BACKGROUND AND EXISTING CHALLENGES
Researchers have used a wide variety of devices to enhance the ESM experience. Some of the most commonly used devices include pagers [12], Personal Digital Assistants (PDAs) [7], paper booklets, audio recordings, and cameras [11]. More recently, thanks to the massive adoption of cellphones, researchers have extensively used smartphones as a way to trigger, collect and log the responses of participants, e.g., [1, 2, 5, 7, 8, 14]. Some of the main benefits of using cellphones are that people are already used to carrying them and receiving and sending information with them. Moreover, these devices offer many opportunities to collect a wide range of information (e.g., social interaction, motion data), that can be used to alter the ESM process in meaningful ways, e.g., only collect information when the person is not moving. However, phone-based ESM still has major limitations.

The process of receiving a prompt is arguably disruptive. As soon as a prompt is triggered, the person needs to reach for the device, provide the requested information, and (if it was moved) return it to the original location. If the person is carrying the device inside the pants pocket, e.g., a phone, this process can significantly bias responses. For instance, if the person is sitting down, it is very likely s/he will need to change their body posture to reach for, and use, the device. This process could not only affect answers to specific questions, e.g., “How frustrated do you feel right now?” but also may significantly alter physiological readings that are used to better understand emotions e.g., heart rate usually accelerates upon standing up. The recommended response time to a specific prompt is less than two minutes [11]. Therefore, it is desirable to ensure that participants do not spend extra time using the devices.

A fundamental challenge of ESM tools is to obtain a high response rate, with that rate unbiased by a participant’s level of stress, activity, or attentive resources. While there are well-established notification techniques such as sounds and vibrations, it remains a challenge to make them noticeable without becoming disruptive. This is especially the case during the busy daily activities in which environmental noises or large body motions may obscure even the most disruptive kind of notifications. As a result, relevant and costly information may be dismissed and the content of the assessments may be biased towards specific moments in time (e.g., when there is little noise and body motion, or when the user is most relaxed, significantly affecting the potential generalization of findings) [1].

Finally, one of the main design principles of ESM tools is to minimally disrupt daily activity while collecting as much information as possible. In turn, the length and number of prompts should be as short and infrequent as possible to minimize the burden to participants. To achieve that, researchers usually minimize the response time as much as possible so participants are eager to provide more frequent ratings and leave lengthier questions to other moments in time when participants are expected to be more available [8, 18]. However, successfully achieving these goals can be quite challenging as postponing questions can easily result in well-studied recall biases (e.g., people forget things, report false memories) [17, 22]. In this work we ask lengthier questions at the end of the day and employ a wearable body camera to capture daily activities and minimize the recall lag.

CONSIDERED WEARABLE LOCATIONS
We explore the potential advantages of using different types of wearable devices in the context of experience sampling. In particular, we compare the traditional “phone inside the pocket” approach against wearable devices that are worn on the wrist and on the head. While there are many other body locations that we could have considered, we believe these two represent important trends for wearables that could offer different ESM benefits. Below we highlight the main characteristics for the wrist-worn and head-worn locations.

Wrist-worn. Recent years have shown enormous growth in the use of smartwatches, offering easily-accessible information to wearers. In this work, we use the Gear Live smartwatch (Samsung, Inc.), henceforth referred to as Watch, which is equipped with many sensors similar to those on a smartphone. This type of device offers a unique opportunity to provide information to the users in an easily accessible and concealable location. While smartwatches are not as widespread as smartphones, many people are used to wearing a watch, which potentially minimizes the burden associated with wearing them. The interactions with the Watch are very similar to those with traditional smartphones (e.g., touch surface display to receive and provide information) but the screen size of 1.63-inches (320x320 pixels) is much smaller than that of the smartphone used in this study, fundamentally limited by its body location. In contrast, the smartphone we use in this study (Galaxy S4 by Samsung Inc.) has a screen size of 5-inches (1920x1080 pixels).

Head-worn. A relatively new and fast-growing trend is the availability of head-worn display devices such as the Oculus Rift or the Microsoft Hololens. In this work we explore head-worn technology through the use of Google Glass (Google, Inc.), henceforth referred to as Glass. Glass includes a see-through display located just above the right eye (640x360 pixels), a touch surface on the right side of the glasses, and a bone-conductive speaker above the right ear, as well as many other sensors. This device offers unique opportunities to provide quicker and more intimate information to the user while capturing behavioral information (e.g., eye gaze, head gestures, facial expressions). As this wearable form-factor is relatively new and unfamiliar to users when compared to smartphones or...
smartwatches, the types of interactions are less well-defined and standardized. However, the device offers the opportunity to explore new types of interactive inputs, such as head gestures.

Recent research has begun to exploit the unique advantages of both wrist-worn [23] and head-worn [16, 21] wearable devices to gather reports during daily life; however, those studies usually consider very different experimental conditions, e.g., reporting perceived exertion vs. annotations in wet laboratories, limiting the comparison across different devices. Motivated by these efforts and previous work quantifying the pros and cons of cellphone-based ESM versus earlier approaches of PDAs [7] and pen-and-paper based methods [5], we perform a systematic comparison of three types of devices within the same experimental condition – stress and mood measurement – made repeatedly during five typical workdays for a person.

WEARABLE EXPERIENCE SAMPLING TOOL

While there are a great range of ESM tools available in the market [10], none of them were designed to work across the three platforms we wanted to compare. Therefore, we designed and developed a novel ESM application that could be deployed across all three types of devices, also ensuring the maximum level of control and aesthetic similarity. This section provides more details about the application and the main differences across the devices.

Implementation

We used the Android development platform to create the new ESM tool. Android is already used in a wide variety of devices, such as smartphones and wearable devices, and offers the benefit of significantly reducing the costs associated with multi-device support. Our tool includes two of the most commonly used types of reporting tools: Grid and Likert scale questions.

2D-Grid. This type of question asks the user to point at a specific location on a 2D-Grid to select his/her answer. In our study, we created two different grid questions. The first one asked the user to report her/his affective state in terms of emotional valence (x-axis) and arousal (y-axis), which is based on the commonly used Circumplex model of emotions [20]. The second one asked the user to report his/her current job demands (x-axis) and resources (y-axis), which is based on the Job Demands-Resources model to quantify the wellness of the work environment [3, 13].

Likert-scale. This type of question asks the user to pick a point on a rating-scale. In our application, we used a 5-point Likert scale to ensure the different options could be easily read and accessed on the smaller smartwatch screen. We asked two Likert scale questions. The first was, “How stressed are you feeling right now?” and the second was, “How disruptive was this prompt?” Both Likert scales had end points “Not at all” and “Extremely.” Only the end points were labeled.

To help make the tool easy to use each time the application triggered a prompt, the four questions were presented in the same order. While we could have used two Likert-scale questions instead of one 2D-Grid, we wanted to ensure we included different types to better capture potential differences. Figure 1 shows the application screenshots of a 2D-Grid and a Likert-scale question.

Interaction

While the underlying implementation and appearance was the same across devices, the specific form-factor of each wearable device resulted in different interaction patterns. This was especially true for the head-mounted device, which was the least familiar to the participants. This section describes the interactions performed by users in order to report their answers.

2D-Grid. When using the smartphone and the smartwatch, the user touched their display to select the preferred answer. While the user continues to touch the display, a 3-second countdown starts and a progress circle is shown around the finger of the user. At the end of 3 seconds, the answer is automatically submitted. If the user interrupts the countdown and stops touching the display, the progress circle will disappear and a smaller white circle will appear on the latest touched point, indicating that the countdown has been reset. The time of the countdown can be easily configured but we found 3 seconds was enough to prevent accidental submissions, especially when taking the phone from the pocket. While Glass also provides a display, its touch pad only works on one dimension, which does not easily allow for pointing on a 2D-Grid. However, due to its location on the head and onboard motion sensors, the device offers the opportunity of using head gestures to point toward and report users’ selections. Therefore, we used the Android libraries to estimate the pitch and yaw of the head and mapped the head orientation in real-time to a virtual pointer on the display. To start the response process, the user needs to tap with one finger on the side of the Glass and the virtual pointer appears in the middle of the 2D-Grid. Then, head movements are intuitively associated with the movement of the pointer. During this process, the user can tap again in order to re-center the pointer. Finally, the response is submitted by tapping with two fingers simultaneously. Both the one finger and two finger tap interactions were pre-defined gestures in Glass; the selection of which one was used to re-center the pointer and which one to submit the response was made to minimize accidental submissions (the two finger tap was deemed less likely to happen accidentally). To further minimize accidental submission, one finger and two finger taps needed to be 1 second apart in order to be registered.

Likert-scale. When using the smartphone and the smartwatch, the user similarly uses touch interactions to pre-select the preferred answer and then a virtual submit button to complete the process. In the case of Glass, the user had to swipe one finger forward/backward over the
Figure 2. Prompts could be postponed for 5 minutes by pressing the volume down button of the smartphone (left), the side button of the watch (middle), and by swiping two fingers down on the side of the glasses (right).

Every new question and user interaction with the application triggered an auditory (Glass) or haptic (smartphone and smartwatch) feedback, which is designed to be as unobtrusive as possible while still being noticeable. Once a prompt is triggered, the application automatically turns the display on and a subtle notification appears. If the prompt is not answered, the notification is repeated every 30 seconds to ensure the user notices the prompt and does not forget to respond to it. If the prompt is not answered within a pre-defined amount of time (3 minutes in our study), it is automatically dismissed and the display of the device is automatically turned off. If the person is too busy to respond to the prompt, s/he can delay it by a certain amount of time (5 minutes). To do so, the user needs to either click the physical button located on the side of the smartwatch, the volume down button of the phone, or swipe two fingers down on the Glass (see Figure 2). The application also records time stamps for every user interaction and other relevant events such as the time when the prompt is triggered, the time when the first user interaction occurs, and the time when the reports are provided. Table 1 provides an overview of the main interaction differences for each device.

### Prompt Triggering
Deciding when to trigger a prompt to collect information is directly influenced by the purpose of the study and potential use of the reports. In our study, we wanted to gather information throughout the day to capture stress fluctuations as well as compare how people reported on each of the three devices. Therefore, we decided to trigger the prompts at random times and uniformly across the different devices, which also minimized users’ anticipation for the prompts.

To ensure the different devices would not overload the user by prompting at the same time, we pre-generated files using the randomly-generated triggering times of each device and transferred them to each of the devices before the study. Then, the ESM tool automatically loaded them and used them to trigger the prompts. In our study, the triggering times were generated with a custom-made MATLAB script following these constraints:

- **Time Distribution.** The time between prompts needs to follow a uniform distribution between 30 minutes and 60 minutes when considering all the devices together. Therefore, the user should not get more than one prompt in a 30 minute period and should receive approximately one prompt every 45 minutes.

- **Time Variability.** The standard deviation of the triggering time for each device has to be at least 3 hours ensuring the prompts are distributed throughout the day.

- **Device Variability.** No more than two consecutive prompts can happen on the same device in order to minimize the anticipation of users.

Figure 3 shows an example of triggering times for one day. Each person had a different pattern on each different day.

### Table 1. Main interaction differences across devices.

<table>
<thead>
<tr>
<th>Device</th>
<th>Display</th>
<th>Feedback</th>
<th>Location</th>
<th>Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phone</strong></td>
<td>Rectangle</td>
<td>Haptic</td>
<td>Pocket</td>
<td>Finger</td>
</tr>
<tr>
<td><strong>Watch</strong></td>
<td>Square</td>
<td>Haptic</td>
<td>Wrist</td>
<td>Finger</td>
</tr>
<tr>
<td><strong>Glass</strong></td>
<td>Rectangle</td>
<td>Auditory</td>
<td>Head</td>
<td>Head/Finger</td>
</tr>
</tbody>
</table>

![Figure 3. Example of daily scheduled prompts for the three devices of one of the participants.](image)
EVALUATION

This section describes the experimental protocol which was approved by the Institutional Review Board of the Massachusetts Institute of Technology before the experiment was started.

Protocol

Since this study was embedded in one involving stress measurement, participants were asked to carry several wearable devices during five days of their regular work. The devices included wrist-worn and chest-worn physiological sensors, a Narrative Clip wearable camera (Narrative, Inc.), and the three ESM devices (Glass, Watch and smartphone). Only the latter three prompted the participants with questions. While participants were aware that the main purpose of the prompting tools was to provide self-reported stress levels during the day, they were not aware that their response times were also being studied.

Participants met with the researchers at the start and end of each day to ensure the devices were appropriately charged and worn, and to return the sensors and provide additional information, respectively. Both the start and end times for each day were flexible for each person as long as they tried to work for around 8 hours. At the end of the experiment, participants met with the researcher again to respond to additional questions about the usability of the different devices and to provide comments about the overall experiment. Here, we focus on the analysis of the ESM tools; an extensive analysis focused on real-world stress measurement is the focus of a separate work.

To minimize study drop-outs and ensure participants responded to as many prompts as possible, we provided scaled monetary rewards of up to $200 (in the form of an Amazon gift card). The payments were distributed as follows: $15 for the 1st day, $25 for the 2nd day, $35 for the 3rd day, $45 for the 4th day, and $55 for the 5th day. Additionally, there was a bonus of $25 for completing the whole study successfully.

Data Overview

Fifteen participants (7 females and 8 males) completed the whole study successfully. Another participant started the experiment but after a few hours of wearing the devices decided to discontinue it. Thirteen participants were graduate students, one was a research assistant, and another was an administrator. When participants were asked to briefly describe their work environment, most of them mentioned spending large amounts of time in front of a computer writing text and code, responding to e-mails, and making phone calls. Some of the participants also reported working on electronics and performing laboratory assays. Their work occurred in closed office spaces (with none or two other office mates) and shared collaborative spaces. This population was selected not only due to their exposure to high stress levels but also because they were members of the same technical research lab, facilitating meeting with the researchers at the beginning and end of each day and providing more control over the care of the devices.

The average age of participants was 29 years and 7 months (standard deviation = 6.42) with the youngest being 18 years and 8 months, and the oldest being 41 years and 11 months old. While most participants completed the experiment in five consecutive work days, a few of them had to take one or two days off to recover from sickness and/or to attend to unexpected business away from the workplace.

RESULTS

The experiment triggered a total of 627 prompts across all participants and devices. In order to quantify the differences across devices (independent variable), we extracted the following (dependent) measures for each participant and device: 1) number of triggered prompts, 2) percent of completed prompts, 3) total response time, 4) time between prompt and first user interaction, 5) time between first interaction and final submission, 6) range of 2D-grid responses, and 7) range of Likert responses. When considering all the variables simultaneously, we found a statistically significant difference (MANOVA, F(12, 74) = 4.136, p < .0005; Wilk’s Λ = 0.358, partial \( \eta^2 = .40 \)). The following sections analyze how each of these measures varied with one-way ANOVAs and Bonferroni correction (i.e., statistical significance at p < 0.007).

Response Rates

As mentioned earlier, one of the challenges of ESM is to ensure that participants do not unnecessarily miss prompts. This section explores whether different form-factors yielded different response rates.
The left and right graphs of Figure 4 show the average number of prompts triggered throughout the study for each participant and device, and the percentage of these that were successfully answered, respectively. Green lines indicate the standard error across participants. As can be seen on the left graph, even though the timing of the interruptions was designed to be uniformly distributed across devices, there were significant differences across devices (ANOVA, F(2,42) = 5.71, p = 0.006). A Tukey post-hoc test revealed that there were significantly fewer smartwatch prompts than with the smartphone (p = 0.004). This difference is due to the amount of time each of the devices was working during the experiment, which is affected by several factors. At the beginning of the study, participants were instructed to wear the three ESM devices throughout the work day and charge them during lunch time. However, due to the unpredictability of their schedules (such as meetings during lunch, late breakfast), many participants ended up charging the sensors only when the low battery warning was triggered, resulting in longer charging periods. This problem occurred more frequently for the Watch as its battery life was more limited. Moreover, we noticed that some skin moisturizers would transfer to the bottom of the watch, partially occluding some of the charging ports and, consequently, preventing some of the devices from properly charging.

How many prompts were successfully answered? The total number of answers was very high when considering all the devices (82.3%), yielding a total of 111 unanswered prompts. Both Glass and Watch prompts were answered on average 13% more than the phone prompts (ANOVA, F(2,42) = 3.62, p = 0.035); however, the difference was not statistically significant. At the end of the study, the majority of participants reported feeling comfortable with the number of prompts they received and mentioned that they could have probably provided more reports before feeling disrupted. These comments suggest that slightly lower average response rates on the phone, while consistent with the hypothesis that the Glass and Watch offer more efficient access to the wearer’s attention, were not a statistically significant factor for the participants in our study.

Response Times
Did the response times vary across the devices? We measure these in three ways. Figure 5 shows the average number of seconds between the triggering of the prompts and their completion (left), between the triggering of the prompt and the first user interaction (center), and between the first interaction and the completion of the questions (right).

When considering the total response time (first graph of Figure 5), there were no statistically significant differences across devices (ANOVA, F(2,42) = 2.92, p = 0.065). However, participants took around 42 seconds to report on the Watch and around 52 seconds with the Glass and the phone. When examining the time between the triggering of the prompt and the first user interaction (second graph of Figure 5), there were significant differences (ANOVA, F(2,42) = 6.58, p = 0.003). A Tukey post-hoc test revealed that phone prompts took significantly longer (around 14 more seconds) than the Watch (p = 0.01) and the Glass (p = 0.007). While the location of the device on a more accessible location played an important role, it should be noted that the time to fetch a phone inside the pocket is usually far less than 14 seconds. However, this time captures the amount of time it took for participants to stop their daily activity and focus on answering the prompt, the phone having taken the longest. While 14 seconds may not seem very long, many things can happen in a few seconds during daily life. Indeed, one of the participants reported that something very stressful happened between receiving the notification and starting to respond, which added confusion when completing his stress report. Therefore, minimizing the time needed to access the device is critical to ensure the highest quality of the response. These differences were further supported by some of the comments gathered at the end of the study, e.g., “The phone was by far the worst, because I had to take it out of my pocket,” “Taking the phone out from the pocket was cumbersome.”

Considering the time responding to the questions (right most graph in Figure 5), there were significant differences
ANOVA, $F(2,42) = 15.65, p < 0.001$). The Tukey post-hoc test revealed that responding to the questions with Glass took significantly longer (around 14 seconds) than with the phone ($p < 0.001$) and the Watch ($p < 0.001$). Feedback provided by the participants suggested that this was partly due to the familiarity of the Watch and phone, while the new types of interactions on Glass slowed down the interaction. One of the participants explained, “I had to learn how to use the glasses – I was not used to the head gestures.” To further explore this, Figure 6 shows the average total response times of the first two days versus the last two days for the three devices and found that participants were faster providing their reports by the end of the study, especially for the Glass ($t$-Test, $t(28) = 2.55, p = 0.017$). While we incorporated a practice session at the beginning of the experiment to minimize this effect, longer tutorial sessions or trial lengths could have helped to address this. Note, however, that there may also be some learning effects associated with the type of questions we asked, since people may also get faster at reflecting and reporting their emotional states over time.

To better understand the challenges of interactions, participants were asked to report on a 5-Likert scale how easy it was to interact with each of the devices (with end points “Very challenging” and “Very easy”) at the end of the study. We found that their responses were inversely correlated, although not significantly, with the time it took them to use each of the devices (Pearson’s correlation, $r = -0.16, p = 0.306$). While device familiarity is an important factor to explain this difference, some participants also experienced unexpected problems with some of the designed interactions. For instance, some participants experienced problems with the two-finger tapping gesture to submit responses on Glass. One participant reported “I had problems with the double tapping because of my hair. I realized that one finger was tapping over the hair and the device only detected one finger,” and another participant reported similar problems in which one of the fingers was placed on the non-touch sensitive area close to the display. These comments highlight some of the challenges associated with the design of devices that can work for a wide variety of populations and how hair length, in this case, can be an important factor when designing head-mounted interactions.

Some participants experienced difficulties when responding to 2D-Grid questions on the Watch. In particular, a participant stated, “I found it difficult to point out things [on the Watch] because my finger was on top of it” and another, “The watch was the hardest [to point] because my finger may be too fat.” Finger and screen sizes are important design factors when designing smartwatch interactions.

Overall, these results illustrate how analyzing three different parts of response time can yield different conclusions and how different factors such as location, familiarity, and types of interaction can influence the different aspects of response time.

**Response Distribution**

Ideally, the self-reported answers would not depend on the device used. Figure 7 shows boxplots of the answers for all the questions across each of the devices. Note that each 2D-Grid was separated into two graphs as each captured information along one of the two dimensions. As can be seen, there are some differences across questions, but there is high consistency within each of the questions. To further explore if there were significant differences, we performed a non-parametric Kruskal-Wallis test for each of the questions, and found that none of the comparisons rejected the null hypothesis that the responses belonged to the same distribution ($p > 0.058$). Thus, indicating that the distributions were not significantly different.

When more carefully inspecting the 2D-Grid responses (first four graphs of Figure 7), a small but relevant difference indicates that Watch reports may have used a smaller range of values. To more closely inspect these differences, Figure 8 shows the average range of values for each of the devices and the two types of questions. As can be seen, the 2D-Grid reports on the Watch tended to use around 6% less of the range than the other two devices. However, these difference were not statistically significant (ANOVA, $F(2,42) = 1.07, p = 0.354$). This finding is consistent with the previous comments about the difficulty
of reporting on screens of smaller sizes. While the distributions and ranges of the answers were not significantly different, it is important to keep in mind the type of analysis that will be performed. For instance, if researchers are interested in the extreme points of the 2D-Grid questions (e.g., people with very low arousal and valence), they may need to correct for the screen size of different devices. The range differences were not observed in the Likert-type questions probably because it was easier to point at discrete answers and double check the answer before it was submitted, indicating that different types of interactions and questions can also help reduce the impact of the limited screen size.

At the end of the study, participants also provided ratings about how accurate they thought their reports on each device were. While no significant differences were observed, participants thought that using the phone would probably yield more accurate results, followed by the Watch and then the Glass. As expected, these answers were significantly correlated with how difficult they thought the interactions were with each of the devices (Pearson’s correlation, r = 0.72, p<0.001).

Overall, these findings seem to provide support that our new ESM application was able to appropriately address the different types of device form-factors; most ESM reports did not significantly change across the devices. This was especially interesting to see for the Glass as the interactions with the device were truly novel for most participants.

**Usability of Devices**

An important factor to consider with ESM is whether the prompting devices can not only effectively capture information but also be useful and wearable during daily activities. To further explore the users’ feelings towards each of the considered devices, participants were asked to report any problems they encountered on a daily basis and to respond to usability questions at the end of the study.

Figure 9 shows the distribution of responses for each of the usability questions. In particular, participants were asked to rate whether the devices were comfortable to use, whether they affected their social interactions, whether wearing the device increased their stress levels, and whether they would continue using the device during their daily lives. Both the Watch and the phone received more positive ratings than the Glass across all of the questions. The differences were greatest when reporting about the potential use of the device in the future.

When considering device comfort, the Watch received slightly better results than the phone, and the Glass scores were strongly bimodal. Among the participants who provided negative scores, physical discomfort was a common concern. While we were anticipating it would take a few hours to get used to the new form-factor of Glass, especially with people who did not wear glasses normally, some of the problems persisted until the end of the study.

Part of the discomfort was associated with the physical form-factor of Glass, as several participants commented: “The Glass was a little bit tight on me”, “I did not like [the Glass] because it hurt me a lot. Maybe my ears had something weird...”, “[The Glass] is painful, I wear glasses sometimes and they're not that uncomfortable.” Two of the participants personally addressed this problem by adding some soft padding around the ear on the side where most of the electronics were housed. Some other participants also mentioned that part of the discomfort was associated with the reduction in the field of vision, with a participant stating “I do not like the idea of [the Glass] sitting in your peripheral vision,” and another one “The main problem of [the Glass] is that I was losing the upper part of my view... if anything, I would like wearables that increase my field of
During the first day of study, one participant reported having headaches due to the Glass. However, this problem did not re-occur during the following days. The participants who provided more positive scores in terms of comfort did not seem to experience this type of problem and one of them reported “Surprisingly [the Glass] was not uncomfortable,” and another one “I got used to having [the Glass] on my face.”

A similar split was observed when participants were asked whether the Glass affected (either positively or negatively) their social interactions. In this case, several participants expressed concerns about how they looked and how other people felt about them. One participant explained “Among all the devices, I did not like the Glass because everyone knows you’re wearing it,” and another mentioned “When I was by myself I forgot about [the Glass] but then I became self-conscious about them when I was walking around.” One participant further explained that part of the problem was that wearing the Glass sparked unpleasant conversations about privacy (e.g., “People would feel I was taking pictures of them [with the Glass] and did not enjoy the conversation,” “I went to several meetings with other people, and I would take off the [Narrative] clip and the Glass because I did not enjoy being the topic of conversation”). Indeed, the only participant who dropped the study after a few hours explained that he could not afford having those types of conversations, as his job required him to meet new people very frequently. The participants that provided the more positive ratings seemed to enjoy this type of conversations. One of the participants mentioned “[The Glass] was a nice ice breaker,” and another explained “[Wearing Glass] was a good opportunity to speak about self-tracking.” In either case it is clear that the Glass made people more self-conscious.

Several participants mentioned that receiving notifications on the devices during social interactions was very disruptive, especially for the case of Glass. One participant further explained “I found it very annoying receiving notification through Glass when speaking with people because it was so noticeable.” Finally, two participants reported feeling more distant during social interactions due to the specific form-factors of Glass (e.g., “[The Glass] sort of creates distance between the person you are communicating with and yourself”).

Both device comfort and potential impact on social interactions could partly explain the negative Glass ratings for future use as well as elicited stress due to the device. While the Watch and the phone received more positive comments overall, participants also experienced some problems with them. Even though the band of the Watch was adjustable, the smaller possible configuration was still too large for four of our female participants. Moreover, two of them thought the device was too heavy for their wrist. On the other hand, the same band felt a bit tight for two of the male participants. With regard to the phone, several female participants mentioned they did not enjoy carrying it inside the front pocket of their trousers. One participant stated, “Phones inside the front pocket are not enjoyable as women’s trousers are not designed for it.” The request that women wear it in a pants pocket was made both for data collection purposes for the stress study and to encourage uniformity across the study. While different device form-factors (e.g., elastic wrist-bands, smaller phones) could have partially addressed the problems, these are still important design considerations for practical daily-use.
DISCUSSION

We have conducted a real-life workplace study comparing relevant factors of ESM across different wearable devices. Overall, we found that phones, which have been commonly used in ESM, received the more positive scores in terms of ease of interaction, potential future use, and perceived accuracy of the reports. However, participants of our study took significantly longer to start interacting with the phone application once the prompt was triggered, and there were both verbal reports and a trend towards slower response times indicating that pulling the phone from the pocket was less convenient. These results contrast with the Glass and Watch, which took less time before the first interaction. To interact with the Glass, we used novel head and finger gestures which did not affect the distributions of the answers but significantly increased the response time once interaction was initiated. The specific form-factor of the Glass also received some negative scores in terms of potential future use and participants highlighted different scenarios in which the device could have added some unnecessary stress, especially during social interactions. The Watch received more positive scores across the different usability factors and slightly outperformed the phone in terms of device comfort. However, the limited Watch screen size may potentially influence the use of the full-range of some kinds of response scales.

In the context of stress measurement (the motivating purpose of our study), the smartwatch seemed to prevail. This device not only enabled effective and quick gathering of self-reports during the day but also minimized the burden of participants with a concealable and accessible form-factor that did not contribute to increased stress levels. However, it is important to note that different types of devices may be more appropriate in different experimental conditions. For instance, certain work environments may minimize part of the social stigma associated with head-mounted wearable devices. Also, if the reporting tasks are limited by a small screen (such as reporting on a 2D-Grid) then a smartphone or Glass device may be preferable. Finally, this study explored the use of head gestures and touch interactions; however, different findings may be obtained with different types of interactions. For instance, using voice as an input mechanism may be more invasive but also may yield more similar response times across the three devices.

This work evaluated the ESM tool in a population of 15 people during five work days yielding a total of 627 prompts. While we obtained several significant findings, the sample size and the demographics of our participants may not necessarily generalize to other populations. In our experiment, all of the participants were part of a technical research institution and, therefore, they tended to be familiar with state-of-the-art of technology. Because of this, participants may have been more open or critical to certain aspects of the devices which could explain the divided opinions for some of the questions. Participants of our experiment were also requested to not wear eyeglasses during the study in order to minimize the burden associated with them. As a result, only two out of the 15 participants used regular eye glasses during their regular daily life which could have biased some of the qualitative Glass ratings. Similarly, only three of the participants were used to wearing a watch during their daily lives. Another limiting factor of our study is that participants were not allowed to use the devices for their personal use. While this is a reasonable request when providing ESM tools with limited battery power to participants, it can also affect the perceived utility and potential use of the devices. For instance, if participants could have used the Glass to take photos or receive phone calls, then they may have provided more positive scores. Moreover, a few months before the study took place, the Glass device received some media criticism due to its onboard camera and its potential to invade the privacy of others (e.g., taking photos without people awareness). While we disabled the camera for the purpose of the study, some participants were still concerned. Finally, participants of our study were instructed to carry the devices on the same body locations throughout their work day. While this requirement helped quantify differences across devices, less structured usage of the devices (e.g., the phone can be left on the desk, the Glass can be used for few hours) could yield different findings. Future studies will consider larger sample sizes, more varied populations and potentially other device form-factors to help corroborate our findings and further investigate some of the challenges.

CONCLUSIONS

We have created a new application for ESM that works across smartphones, wrist-worn and head-worn devices. In a real-life workplace deployment, we examined how each device form-factor interacts with several variables of importance to experience sampling.

Overall, our results suggest that there is not a one-solution-fits-all for ESM. Different devices offer different benefits which may make them more adequate for different scenarios. An interesting area of future work would focus on systematically exploring which devices are most suited to specific situations and applications. Moreover, it is very probable that people may be carrying more than one wearable device in the future, creating new opportunities to explore a multi-device ESM that more effectively lowers the burden of participants. In addition, we would also expect that different people may prefer different prompting channels (e.g., people already wearing glasses may prefer the head-worn form-factor) and, therefore, learning and adapting to users’ preferences will be a necessity.

In the long-term, we expect smartphones and wearable devices will become more ubiquitous, creating new opportunities to more effortlessly capture self-reports and to help advance ambulatory scientific discovery.
ACKNOWLEDGMENTS
This material is based upon work supported by the National Science Foundation (NSF 1029585), Google, the Samsung Think Tank Team, and the MIT Media Lab Consortium.

REFERENCES