Investigating Contextual Cues as Indicators for EMA Delivery

Varun Mishra,[1] Byron Lowens,[2] Sarah Lord,[1] Kelly Caine,[2] and David Kotz[1]
[1]Dartmouth College, [2]Clemson University

Dartmouth College Computer Science Technical Report TR18-842*

Abstract

In this work, we attempt to determine whether the contextual information of a participant can be used to predict whether the participant will respond to a particular Ecological Momentary Assessment (EMA) prompt. We use a publicly available dataset for our work, and find that by using basic contextual features about the participant's activity, conversation status, audio, and location, we can predict whether an EMA prompt triggered at a particular time will be answered with a precision of 0.647, which is significantly higher than a baseline precision of 0.410. Using this knowledge, the researchers conducting field studies can efficiently schedule EMA prompts and achieve higher response rates.

1 Introduction

Ecological Momentary Assessment (EMA) [15], also known as the Experience Sampling Method (ESM) [9], is a commonly used technique designed to collect information about a participant's current behavior and experiences while they are in their natural environment. EMA methods ask the participant to answer questions, in the moment, to assist researchers in collecting ecologically valid self-reported data [14]. By collecting responses ''in the moment'', EMA reduces recall bias relative to methods that query the participant at the end of the day or end of the study period. By using technology to collect responses, EMA technology also reduces the labor cost and potential bias incurred in observational studies in which the researcher shadows the participant [15].

The ubiquitous presence of smartphones and wearable devices has enabled the common use of EMA in a broad range of studies. Researchers have successfully used EMA to collect 'ground truth' for annotating sensory measurements and the construction of training data for machine-learning models of human emotion, mood, stress, and personality [20, 4, 21].

The problem, however, is that the researchers are dependent on participants to correctly and diligently answer the EMA prompt. For the participants, responding to frequent or lengthy requests can be burdensome. This burden may decrease participant responsiveness over the course of a study, an effect noted by several researchers [17, 20].

We anticipate that EMA participants would be more responsive if the prompts occur in a context when they are more likely to respond. We propose to time EMA prompts to suit the participant's context -- reducing participant burden and increasing participant compliance. To do so, we must first understand how context affects participant compliance (responsiveness) to EMA prompts.

In this work, we evaluate the context of the participant to determine whether s/he is likely to answer an EMA prompt. We investigate some basic features about participant context (which includes activity, audio, conversation, and location) to determine whether contextual information enables us to predict whether a given EMA prompt is likely to be answered quickly. Our work is the first to use activity, audio, conversation, and location data to predict whether an EMA prompt will be answered. Such a predictive model can help researchers develop effective strategies for delivery of EMA prompts without over-burdening the participant.

We use the publicly available StudentLife Dataset [20]. The dataset consists of longitudinal data from 48 participants over a period of 10 weeks. While the dataset itself contains a wide variety

*This technical report is an extended version of a UbiTtention 2017 workshop paper with the same title [10].

of data (including phone sensor and usage data, EMA data, surveys, dining data, and more), our work focuses particularly on the activity, audio, conversation, and location data along with the self-reported EMA data.

2 Background

Prior work has introduced an assortment of time-based sampling schemes for the delivery of EMA prompts to participants, selected by the researcher based on the goals of the study [1, 6, 15]. The different EMA delivery schemes can be broadly divided into three categories:

1. At a predetermined *fixed time(s)*, EMA prompts may be delivered at a set time during the day. In StudentLife [20], for example, the authors had set 1-2 times during the day for each EMA, and the prompts were delivered at those times for each participant. EMA prompts may also be delivered to participants at equal time intervals (e.g., every 30 to 45 minutes during waking hours of the day), as demonstrated in prior work [7]. In some instances, delivering EMA prompts at equal time intervals permits a particular block of time to function as a unit of analysis. Predetermined fixed-time schemes also support analysis that necessitate consistently spaced EMA prompts to capture participant ongoing experience [15].

   The choice for the fixed time(s) is decided by the researchers based on the needs of their study. In some cases the time of answering the EMA is crucial to the goal of the study, as the researchers might be interested in obtaining the momentary assessment at that time of the day, where as, in some cases the researchers might not have thought of another way, and/or are limited by the EMA delivery system available to them [20].

2. At a *random time*, within certain window(s) of time each day. Participants can be prompted at a random time within a predefined window of time (e.g. morning, afternoon, or evening) [1, 15] or randomly several times a day during a predefined time window (10:00 am to 10:00 pm) [8]. Unlike *fixed time* delivery scheme, where the prompts might be scheduled at 10:00 am and 5:00 pm during the day, a random-time scheme could randomly select a time in the range [9:00 am, 3:00 pm] and another in the range [3:00 pm, 9:00 pm]. These prior studies also show that these 'random'-time schemes are usually not truly random, because there are almost always some parameters constraining the random distribution, the number of prompts, the minimum time between prompts, and the temporal bounds of the day.

3. In *context-sensitive ecological momentary assessments (CS-EMA)* [5] prompts are triggered by contextual information, including time, place, and sensor data [6]. Prior work has demonstrated CS-EMA methods that leverage context like participant location using GPS sensors, participant physical posture using accelerators, participant environmental conditions using environmental sensors, and participant heart rate using wireless chest-strap monitors [6, 19, 2, 3, 16]. Context based approaches trigger a prompt when the sensor detects an activity or state which is of interest to the researchers. [19, 6].

Regardless of the trigger and delivery scheme, the time taken by participants to respond is of great importance to the researchers, because delayed response defeats the goal of obtaining *in-the-moment* response. Hence, it is important to consider the participants' willingness and/or availability (collectively known as the *state of receptivity*) while delivering such prompts.

None of the above methods, however, account for the participant's availability to respond to the prompt. A participant might not be available to answer a particular prompt, for many reasons: the EMA device may not be present, the social context may be requiring the participant's attention, or the participant's activity prevents him or her from seeing the prompt or from responding. In such cases, the participant may respond late (if permitted by the EMA protocol) or never. In some studies, even a delayed response may not meet the research goals.

To address these issues, the EMA delivery policy should pick a 'good' time to trigger prompts -- times when the user is more likely to answer the prompt. Several studies have sought to find

Table 1: Contextual information available in the StudentLife dataset.

Context	Values
Activity	0 : Stationary 1 : Walking 2 : Running 3 : Unknown
Audio	0 : Silence 1 : Voice 2 : Noise 3 : Unknown
Conversation	Start time, End time
GPS Location	Latitude, Longitude
Wi-Fi Location	On-campus Location from WiFi scan

such *opportune moments*, when the user is likely to respond to a notification (any notification, not necessarily EMA prompts) [13, 18, 11, 12].

The most prominent example is *InterruptMe*, an interruption-management library for Android smartphones, designed to allow researchers to look at opportune moments to interrupt the user. They consider contextual information like activity and location [13]. Their analysis also uses features computed from data reported by users, and the researchers achieve a precision of 0.64 in estimating whether a participant will respond to a notification prompt.

Turner et al. investigated whether to push or delay a notification based on contextual information about the phone, including motion, charging state, volume state, ambient light, and phone orientation [18]. They report preliminary results with accuracy of up to 60%.

Other researchers have looked at delivering notifications at activity 'breakpoints' [11] and discovered that delivering a notification at a breakpoint resulted in lower participant cognitive load as compared to those sent out ''immediately'' [12].

In contrast to the above research, we look at a broader set of contextual features, and use passively collected sensing data to predict whether a given prompt will be answered quickly.

3 StudentLife dataset

We use the publicly available StudentLife Dataset, which consists of a wide range of data collected from 48 participants over 10 weeks [20]. Table 1 lists some of the interesting sensor data. The study also used EMA to collect several types of self-report data: stress, affect, behavior, mood, sleep, and activity.

The StudentLife app triggered several EMA prompts each day, based on a predetermined schedule determined by the research team. The schedule was the same for all participants but changed every week. For each EMA response, the dataset recorded the time and content of the response -- but does not indicate which prompt corresponds to which response, or when the prompt was triggered.

We seek to develop a model to predict the *prompt-response time*, that is, the latency between the trigger of a prompt and the participant's response to that prompt. Because the StudentLife dataset does not include the trigger time, we must first estimate the time each EMA prompt was triggered, based on the responses available in the dataset.

In the following sections, we discuss our method to reconstruct the likely trigger times, followed by a discussion of our prediction model and the results obtained.

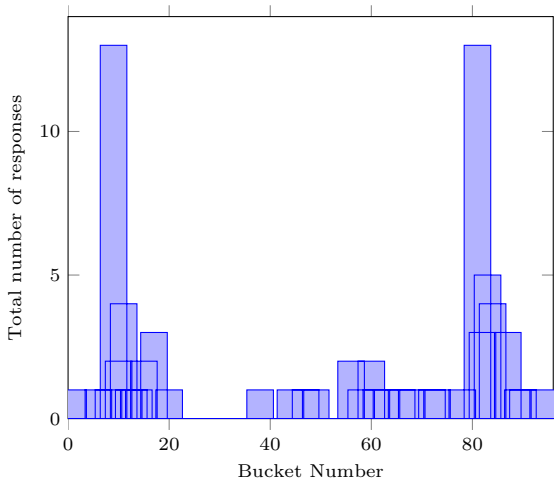

Figure 1: Histogram of responses to the *stress* EMA prompts, across all participants in one day, in 15-minute buckets.

4 Trigger time estimation

In StudentLife, the trigger schedule was identical for all participants, but not recorded in the dataset. The trigger schedule varied from week to week, but the number of prompts per day was small (typically one or two). Our key insight is that the total number of responses to a EMA in a short time immediately following the EMA trigger time should be significantly higher than the number of responses at a later time. In other words, we expect that most participants respond quickly, leading to peaks in the number of responses over time, allowing us to infer that a prompt was triggered shortly before each such peak.

To pursue this approach, we had to verify that EMA prompts were triggered sparingly during the day; if prompts were frequent, responses may be frequent and distributed throughout the day, making it difficult to discern peaks. Although the StudentLife study triggered multiple prompts each day (about 8 per day) we focus on one category -- the *stress* EMA -- to create the *training* dataset for model building and cross-validation. The *stress* EMA was triggered 1-2 times a day. We built two separate *test* datasets, using the *activity* and *sleep* EMA questions. These latter two datasets were used exclusively for testing the model built using the *stress* prompts.

We group the responses by response time into 15-minute buckets, and count the number of responses in each bucket. Figure 1 plots the resulting histogram of responses to the stress prompts in one day of the StudentLife dataset. We can see that the number of responses increases drastically in the 9^{th} and the 81^{st} buckets. Based on our hypothesis, we conclude that the stress prompts were triggered twice in that day -- during the 9^{th} and the 81^{st} blocks.

To find the trigger times, we use a custom peak-detection algorithm to find the blocks in which the stress prompt was triggered. For every such block detected by our algorithm, we assume the corresponding prompt was triggered at the start time of that block. Using this approach, we determined 54 trigger times for the stress prompt over the length of the study. Although the stress prompt may have been triggered more times -- we may have overlooked some peaks -- we are confident of these 54 occasions.

We observed that all 54 occasions discovered by our algorithm were either on the hour (:00) or on the half hour (:30). We contacted the authors of StudentLife and they confirmed our findings, saying that their EMA trigger times were always aligned on the hour or half hour. This confirmation gave us confidence that our estimated trigger times were accurate.

Next, for every participant, we checked for a response within 4 hours of the estimated EMA trigger time. If such a response existed, we assumed the participant answered that prompt, whereas if there was no response within the 4-hour window, we assumed the participant did not answer that

Context	Features
Activity	*before_activity* and *after_activity*, where each can take a value from 0-3, depending on the labels in Table 1
Audio	*before_audio* and *after_audio*, where each can take a value from 0-3, depending on the labels in Table 1
Conversation	*before_convo* and *after_convo*, where each can either be 'true' or 'false', depending on whether there was a conversation detected in that window
Location	*before_loc* and *after_loc*, where each can take be one label depending on the building type: study, dorm, food, gym, etc.
Time	*time* of the day, *day* of the week

Table 2: The features computed for the different contexts.

EMA prompt. (The StudentLife system did not save a prompt 'id' with each response, so if we had looked at a longer time period, the response might have been to a later prompt, instead of the current prompt.) We found a total of 906 responses from 2,179 stress prompts across all participants. We used this dataset for training and building our model, and call it the *train* dataset. We followed a similar process for the activity and sleep prompts. For the activity prompts, we found 24 trigger times, which led us to obtain a total of 374 responses from 1,010 activity prompts ($test_a$ dataset). For the sleep prompts, we found 29 trigger times, and a total of 631 responses for 1,185 sleep prompts ($test_s$ dataset).

With solid estimates for the trigger time of the EMA prompts, we explain our prediction model in the next section.

5 Prediction model

In this section, we give an overview of our modeling approach and feature computation, followed by the experimental results, starting with the model building and cross-validation with the *train* data, and then testing the performance of the trained model with the two *test* datasets -- $test_a$ and $test_s$.

5.1 Modeling overview

In this work, we use contextual information -- activity, audio, conversation and location -- all of which are readily available in the StudentLife dataset. As shown in Table 1, the dataset consists of two different types of locations: (1) GPS-based location, i.e., the latitude and longitude of the participants' current location, and (2) Wi-Fi based location, which provides the on-campus building name in or around which the participant is present. Since the building name can give us more information about a participant's location, we use the Wi-Fi based location in our model. We then

Table 3: Predicting whether a prompt will be answered based on the context at the time of prompt, using the *train* data.

Classifier	Precision	Recall	Area Under ROC
SVM	0.647	0.526	0.654
Random Forest	0.633	0.551	0.711
Naive Bayes	0.635	0.546	0.724
Baseline	0.410	0.420	0.512

map each building name to a particular category (e.g., study, dorm, food, street) and use these labels in our predictive model.

In our model, we look not only at the contextual information leading up to the time when an EMA was triggered, but also if there was any *change* in context during that time. We consider a window of time before an EMA trigger time, and compute the ''before'' and ''after'' contextual features on that window, so that we can capture the context change in that window. Table 2 lists the features we compute. For example, if the time at which the EMA prompt was triggered was t, and the size of the time window we use to compute features is Δt, then the ''before'' features will be computed in the time range $[t - \frac{\Delta t}{2}, t]$, and the ''after'' features will be computed at $[t, t + \frac{\Delta t}{2}]$. For our experiments, we set $\Delta t = 10$ minutes.

Since we aim at modeling receptivity, we predict the following outcomes: (1) whether a participant will respond to a particular EMA prompt, given an indefinite amount of time, (2) whether a participant will respond to a particular EMA prompt within a given time interval, t_d, and (3) the time taken by a participant to respond to a particular EMA prompt. For the first two outcomes, we report *precision*, i.e., the proportion of the receptive instances predicted by our model where the participants indeed were receptive, and the *recall*, i.e., the proportion of all receptive instances correctly identified by our model. We also report the area under the ROC (Receiver Operating Characteristic) curve. For the third outcome, we use a regression-based approach to predict the time to respond, and report the prediction error -- Mean Absolute Error (MAE) -- in seconds. Prior work like InterruptMe used similar metrics for measuring interruptability [13].

5.2 Experimental results

We train models on the *train* data, and evaluate the performance with 10-fold cross-validation for the three different outcome measures. We then use the trained models to evaluate the predictive performance using the two *test* datasets. Finally, we do a brief analysis as to how the change in context affects responsiveness towards EMA prompts.

5.2.1 Training data

To evaluate the first outcome: for each EMA prompt we calculate the notification context for every participant and label it *true* if that participant provided a response to that EMA prompt, and *false* otherwise. We then perform 10-fold cross validation using three different classifiers -- SVM, Random Forest and Naive Bayes -- and report the results -- precision, recall and area under the ROC (Receiver Operating Characteristic) curve -- in Table 3. We also report the baseline classification results for comparison. This baseline is calculated by classifying the instances with a probability based on the proportion of EMA prompts that were actually answered in the training set.

We observe that all the context-based models perform significantly better than the baseline model, consistently achieving a precision above 0.63, with a highest precision precision of 0.647, which is similar to the precision achieved by InterruptMe. Furthermore, for the highest precision, we achieve recall greater than 0.52, which is significantly better than the recall reported in InterruptMe for a similar precision. This suggests that in comparison to InterruptMe, our model finds a greater proportion of *opportune* moments, with comparable precision. We also achieve significant Area under ROC values as compared to a random baseline, suggesting context-based models are actually effective in predicting if a prompt will be answered.

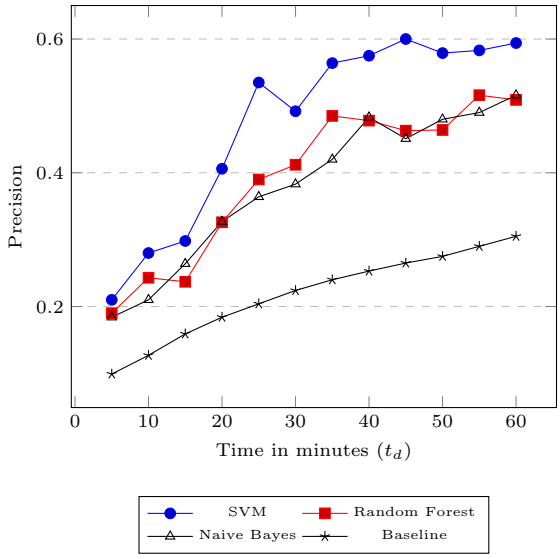

Figure 2: Predicting whether a prompt will be answered within a given time threshold (t_d), based on the context at the time of the prompt.

Table 4: Predicting the time taken to respond to a prompt, based on the context at the time. The MAE is in seconds (lower is better).

Classifier	Mean Absolute Error (MAE)
SVM Regression	2441.98
Random Forest Regression	2589.10
Linear Regression	2506.89
Baseline	2721.69

To evaluate the second outcome: for each EMA prompt we calculated the notification context for every participant and label it *true* if that participant provided a response to that EMA prompt, within a *threshold time (t_d)*, and *false* otherwise. Figure 2 shows the 10-fold cross-validation results across different classifiers. Observe that as we increase the time boundary (t_d), the precision also improves.

To evaluate the third outcome: for each of the 906 EMA prompts that were answered by participants, we compute the notification context for each participant, and label it with the time taken to respond to that prompt (in seconds), i.e., the difference between the response time and the trigger time of the prompt. Having generated the dataset, we perform 10-fold cross validation to predict the time taken to respond, using the following algorithms -- SVM Regression, Random Forest Regression, and Linear Regression. We report the Mean Absolute Error (MAE) achieved using the four algorithms as compared to the baseline, in Table 4. Here the baseline is the MAE obtained by predicting the time to response as the average time to respond across all participants.

As we can see from Table 4, while the MAE obtained from all the classifiers is slightly better than the baseline, they still are high. A MAE of 2441 seconds is an error of ±40 minutes, which is a big error margin, and hence, will be unacceptable for most work. It is possible that our selection of features, focusing on *contextual break-points*, may not be right for predicting the time a participant might take to respond to a prompt. We plan to explore this issue in future work.

5.2.2 Testing data

Having evaluated the predictive performance on the *train* data through cross-validation, we evaluate the trained model on two different datasets, collected from different EMA prompts -- the $test_a$ data

Table 5: Predicting whether a prompt will be answered based on the context at the time of prompt, using the *test* datasets.

	$test_a$			$test_s$		
Classifier	**Precision**	**Recall**	**Area Under ROC**	**Precision**	**Recall**	**Area Under ROC**
SVM	0.534	0.487	0.618	0.691	0.433	0.656
Random Forest	0.498	0.545	0.654	0.721	0.455	0.687
Naive Bayes	0.496	0.561	0.668	0.659	0.513	0.668
Baseline	0.410	0.420	0.520	0.410	0.420	0.501

collected from activity prompts and the $test_s$ data collected from sleep prompts. Here, we predict the presence of a response to a prompt, i.e., given the context of the participant, will the participant answer that prompt. We report our results for both the $test_a$ and $test_s$ datasets in Table 5.

We observe that while Area under the ROC curve value for both the *test* datasets is consistent, it is marginally lower than what was achieved during cross-validation with the *train* data. It is a different story, however, when we look at the precision and recall scores for the two *test* datasets. While $test_a$ has higher recall scores, $test_s$ has significantly higher precision scores. This result leads us to believe that the model generated on one type of EMA prompt cannot be applied *as-is* to another type of prompts; but some tuning of parameters (e.g., the hyper-parameters C and γ in SVM) is required. It may be that participants react to different types of EMA prompts differently; a participant may answer one type of prompt in a particular context, but does not answer a different type of prompt in the same/similar context. Such a behavior might lead to different response rates to different types of prompt, which is exactly the case in this situation. In our dataset, we have across all participants, we see a response rate of 41.50%, 37.02%, and 53.24% for the stress, activity, sleep prompts respectively. It would be interesting to look at what causes such a disparity in EMA response rates, which we leave to future work.

5.2.3 Effect of context on responses

We further sought to understand how context affected response to EMA prompts. For our purposes, we define *responsiveness* as the percentage of prompts a participant answered. We look at how a *change* in context in the time just before a prompt increased or decreased the responsiveness of a participant in that context, as compared to the baseline measure, i.e., overall responsiveness across all prompts. We only consider the *train* data for this analysis. In Table 6 we observe that context changes seem to have only a slight impact on participant responsiveness, when we consider the average across all the participants. We found, however, a substantial change in responsiveness when we look at individual participants (in this table we examine two randomly selected participants): note, for example, how a change in location *seems* to have decreased the responsiveness of Participant 1 (P1) by 13.6%, whereas it increased the responsiveness of Participant 2 (P2) by 16.0%. It is interesting to observe how a context can have opposite effects on the responsiveness of different participants.

6 Discussion

In this section we discuss some of the key insights generated from our work.

- We were able to use some basic contextual cues to predict whether an EMA will be answered, with a precision of 0.647 (which is over 56% higher than just random guessing), and the highest recall of 0.551 (which is over 31% higher than random guessing), as shown in Table 3. The precision and recall reported by the cross-validation of the *train* data are similar to the results obtained in prior work [13]. Further, we achieve a maximum Area under the ROC value of 0.724, which suggests that the model is fairly accurate as compared to random guessing.
- Our work is the first to predict the time taken to respond to an EMA prompt based on the context of an user. While we obtain slightly better results as compared to the baseline, the Mean Absolute Error was too high (±40 minutes) for the results to be useful. This result

Table 6: Affect of context on the change in *responsiveness* towards EMA prompts: across **all** participants, and two randomly chosen participants (**P1** and **P2**).

Contexts	All	P1	P2
Baseline	43.3%	80.0%	59.0%
Activity			
Change	3.9%	2.2%	41.0%
No Change	-0.4%	-0.2%	-1.9%
Audio			
Change	1.2%	10.8%	-34.0%
No Change	-0.3%	-4.0%	3.5%
Conversation			
Change	-2.1%	3.3%	21.0%
No Change	0.4%	-2.0%	-2.6%
Location			
Change	3.2%	-13.6%	16.0%
No Change	-0.7%	2.0%	-1.5%

suggests that the features being used to predict for the presence of a response might not be useful when trying to predict response time. We aim to look at this further, and come up with different features in future work.

- We test our model built with one type of EMA prompt -- the *stress* prompt -- to predict different types of prompts -- *activity* and *sleep* prompts. From the results reported in Table 5, we observe that the model generated using one type of prompt cannot be applied 'as-is' to another kind of prompt, and might require some parameter tuning for it to provide similar results as the *train* data. It is possible that participants react to different prompt types differently, as shown by the wide range in the overall response rates for the different prompts. This observation may be important, as other researchers and previous works tend to assume all EMA prompts to be similar.

- Our results may be limited by our need to estimate the times the prompts were triggered in the StudentLife dataset. While we are confident that the times we have estimated are accurate, we are concerned about the trigger times we could not estimate. Hence, in this work, we only consider prompts for which we identified a trigger time. If we had had all trigger times, and thus been able to evaluate all responses, the results reported above would change and our conclusions may be different.

7 Conclusion and Future Work

In this paper we evaluate the use of contextual information to predict whether a participant will respond to an EMA prompt. Specifically, we explored activity, conversation, audio and location context from the StudentLife dataset. While we understand that interruptability is based on a wide range of factors, our preliminary results give us the confidence to explore deeper. In future work, we hope to explore factors like telephone and SMS logs, phone-app usage, phone-charging events, and calendar events. We also aim to develop an application that triggers EMA prompts according to context so we can evaluate the effect on participant's response time, quality of response, and number of responses.

8 Acknowledgement

This research results from a research program at the Institute for Security, Technology, and Society (ISTS) at Dartmouth, in collaboration with the Center for Technology and Behavioral Health (CTBH)

at Dartmouth, supported by the National Science Foundation under award numbers CNS-1619970 and CNS-1619950. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.

References

[1] Glenn Affleck, Howard Tennen, Susan Urrows, Pamela Higgins, Micha Abeles, Charles Hall, Paul Karoly, and Craig Newton. Fibromyalgia and women's pursuit of personal goals: a daily process analysis. *Health Psychology*, 17(1):40, 1998. DOI 10.1037/0278-6133.17.1.40.

[2] Genevieve Fridlund Dunton, Eldin Dzubur, and Stephen Intille. Feasibility and performance test of a real-time sensor-informed context-sensitive ecological momentary assessment to capture physical activity. *Journal of medical Internet research*, 18(6), 2016. DOI 10.2196/jmir.5398.

[3] Genevieve Fridlund Dunton, Eldin Dzubur, Keito Kawabata, Brenda Yanez, Bin Bo, and Stephen Intille. Development of a smartphone application to measure physical activity using sensor-assisted self-report. *Frontiers in public health*, 2:12, 2014. DOI 10.3389/fpubh.2014.00012.

[4] Karen Hovsepian, Mustafa al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. cStress: Towards a gold standard for continuous stress assessment in the mobile environment. In *UbiComp'15*, pages 493--504. ACM, 2015. DOI 10.1145/2750858.2807526.

[5] Stephen S Intille, John Rondoni, Charles Kukla, Isabel Ancona, and Ling Bao. A context-aware experience sampling tool. In *CHI'03 extended abstracts on Human factors in computing systems*, pages 972--973. ACM, 2003. DOI 10.1145/765891.766101.

[6] Stephen S Intille, AA Stone, and S Shiffman. Technological innovations enabling automatic, context-sensitive ecological momentary assessment. *The science of real-time data capture: Self-reports in health research*, pages 308--337, 2007.

[7] Thomas W Kamarck, Saul M Shiffman, Leslie Smithline, Jeffrey L Goodie, Jean A Paty, Maryann Gnys, and Joey Yi-Kuan Jong. Effects of task strain, social conflict, and emotional activation on ambulatory cardiovascular activity: Daily life consequences of recurring stress in a multiethnic adult sample. *Health Psychology*, 17(1):17, 1998. DOI 10.1037/0278-6133.17.1.17.

[8] David Kimhy, Philippe Delespaul, Cheryl Corcoran, Hongshik Ahn, Scott Yale, and Dolores Malaspina. Computerized experience sampling method (esmc): assessing feasibility and validity among individuals with schizophrenia. *Journal of psychiatric research*, 40(3):221--230, 2006. DOI 10.1016/j.jpsychires.2005.09.007.

[9] Reed Larson and Mihaly Csikszentmihalyi. The experience sampling method. *New Directions for Methodology of Social & Behavioral Science*, 1983.

[10] Varun Mishra, Byron Lowens, Sarah Lord, Kelly Caine, and David Kotz. Investigating contextual cues as indicators for EMA delivery. In *Proceedings of the International Workshop on Smart & Ambient Notification and Attention Management (UbiTtention)*, pages 935--940, September 2017. DOI 10.1145/3123024.3124571.

[11] Mikio Obuchi, Wataru Sasaki, Tadashi Okoshi, Jin Nakazawa, and Hideyuki Tokuda. Investigating interruptibility at activity breakpoints using smartphone activity recognition api. In *UbiComp'16: Adjunct*, pages 1602--1607. ACM, 2016. DOI 10.1145/2968219.2968556.

[12] T. Okoshi, J. Ramos, H. Nozaki, J. Nakazawa, A. K. Dey, and H. Tokuda. Attelia: Reducing user's cognitive load due to interruptive notifications on smart phones. In *PerCom'15*, pages 96--104, March 2015. DOI 10.1109/PERCOM.2015.7146515.

[13] Veljko Pejovic and Mirco Musolesi. Interruptme: Designing intelligent prompting mechanisms for pervasive applications. In *UbiComp'14*, pages 897--908. ACM, 2014. DOI 10.1145/2632048.2632062.

[14] Saul Shiffman. Ecological momentary assessment. In *The Oxford Handbook of Substance Use and Substance Use Disorders*. 1998.

[15] Saul Shiffman, Arthur A Stone, and Michael R Hufford. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4:1--32, 2008. DOI 10.1146/annurev.clinpsy.3.022806.091415.

[16] Jason G Su, Michael Jerrett, Ying-Ying Meng, Melissa Pickett, and Beate Ritz. Integrating smart-phone based momentary location tracking with fixed site air quality monitoring for personal exposure assessment. *Science of The Total Environment*, 506:518--526, 2015. DOI 10.1016/j.scitotenv.2014.11.022.

[17] Vincent W. S. Tseng, Michael Merrill, Franziska Wittleder, Saeed Abdullah, Min Hane Aung, and Tanzeem Choudhury. Assessing mental health issues on college campuses: Preliminary findings from a pilot study. In *UbiComp'16: Adjunct*, pages 1200--1208. ACM, 2016. DOI 10.1145/2968219.2968308.

[18] Liam D. Turner, Stuart M. Allen, and Roger M. Whitaker. *Push or Delay? Decomposing Smartphone Notification Response Behaviour*, pages 69--83. Springer International Publishing, 2015. DOI 10.1007/978-3-319-24195-1_6.

[19] Luuk van Wel, Anke Huss, Philipp Bachmann, Marco Zahner, Hans Kromhout, Jürg Fröhlich, and Roel Vermeulen. Context-sensitive ecological momentary assessments; integrating real-time exposure measurements, data-analytics and health assessment using a smartphone application. *Environment international*, 103:8--12, 2017. DOI 10.1016/j.envint.2017.03.016.

[20] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *UbiComp'14*, pages 3--14. ACM, 2014. DOI 10.1145/2632048.2632054.

[21] Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T. Campbell. Smartgpa: How smartphones can assess and predict academic performance of college students. In *UbiComp'15*, pages 295--306. ACM, 2015. DOI 10.1145/2750858.2804251.