Understanding Emotion Changes in Mobile Experience Sampling

Soowon Kang
sw.kang@kaist.ac.kr
KAIST
Daejeon, South Korea

Cheul Young Park
cheulyop@kaist.ac.kr
KAIST
Daejeon, South Korea

Narae Cha
nr.cha@kaist.ac.kr
KAIST
Daejeon, South Korea

Auk Kim*
kimauk@kangwon.ac.kr
Kangwon National University
Chuncheon, South Korea

Uichin Lee*
uclee@kaist.ac.kr
KAIST
Daejeon, South Korea

ABSTRACT

Mobile experience sampling methods (ESMs) are widely used to measure users' affective states by randomly sending self-report requests. However, this random probing can interrupt users and adversely influence users' emotional states by inducing disturbance and stress. This work aims to understand how ESMs themselves may compromise the validity of ESM responses and what contextual factors contribute to changes in emotions when users respond to ESMs. Towards this goal, we analyze 2,227 samples of the mobile ESM data collected from 78 participants. Our results show ESM interruptions positively or negatively affected users' emotional states in at least 38% of ESMs, and the changes in emotions are closely related to the contexts users were in prior to ESMs. Finally, we discuss the implications of using the ESM and possible considerations for mitigating the variability in emotional responses in the context of mobile data collection for affective computing.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Mobile devices**.

KEYWORDS

Emotion / Affective Computing ; Mobile Devices: Phones/Tablets ; Empirical study that tells us about people ; Experience Sampling

ACM Reference Format:
Soowon Kang, Cheul Young Park, Narae Cha, Auk Kim, and Uichin Lee. 2022. Understanding Emotion Changes in Mobile Experience Sampling. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3491102.3501944

*Corresponding authors.

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3501944

1 INTRODUCTION

Mobile and wearable devices offer opportunities for collecting a wide range of information from users, including their health status and contextual information, such as where they are, what they are doing, and who they are with [84]. Further, with the popularization of low-cost embedded sensors, mobile applications can even extract information that is often not in the immediate grasp of human intuition, such as users' emotions [44], preferences [78], and subtle behavior patterns [63].

In that regard, an in-situ assessment of people's psycho-affective states is essential for mobile and wearable sensing applications. Researchers conducted experiments in controlled settings to study the relationship between various stimuli and people's emotional responses to establish the basis of this technology, including measuring behaviors and bio-signals with stationary equipment while making people listen to emotion-inducing music [36] or watch affective video clips [40]. While such highly controlled studies can acquire fine-grained sensor data with laboratory-grade equipment, they are limited in collecting naturalistic data in diverse and realistic contexts occurring in people's daily lives, which are more challenging to acquire than the data in posed situations. In order to compensate for the limitations of such lab-based studies, the Experience Sampling Method (ESM) [10] has been employed to collect various real-time behaviors and experiences in people's day-to-day activities. As the ESM imposes minimal restrictions on participants' behaviors, it can capture the target behavior/phenomena as naturally as possible.

Nonetheless, emotions collected with the ESM can still be misleading with biases inherent in the method. For example, according to the studies on interruption management for mobile devices [2, 75], unexpected interruptions such as notifications from mobile applications can interfere with people's ongoing tasks, decrease users' efficiency, and even negatively affect their emotions [4]. This calls for assessing the validity of collecting emotion data with the ESM. However, according to our knowledge, no prior study investigated the effect of interruptions introduced by ESM tasks on emotion sampling.

This paper focuses on problems arising from ESM tasks in collecting emotion and stress data. Our study mainly investigated the following three research questions: i) *whether and how ESM tasks affect people's emotions*, ii) *in what conditions ESM tasks cause emotion*

changes, and iii) *what the typical daily contexts are related to emotion changes during ESM tasks*. Towards these goals, we designed a survey to tap into emotional changes during ESM response tasks. Our survey extends existing ESM surveys on user emotions with a new question that directly assesses emotion changes occurring at the survey time. We then conducted rounds of week-long empirical studies and analyzed 2,227 samples from 78 participants to find factors closely related to emotion changes during ESM response tasks.

As a result, we discovered that a considerable amount of changes in users' emotions (i.e., 38.6% of responses in our data) did occur during ESM response tasks. We also conducted repeated measures ANOVA and multilevel regression analysis to identify factors associated with the emotion change and found a total of eight factors—valence, attention level, stress level, task disturbance level, location, smartphone usage time, the standard deviation of heartbeat, the interaction factor between the time of day and day of the week—show a statistically significant correlation with the participants' changes in emotions. Finally, we conducted the thematic analysis of post-study interviews, comparing cases in which there was no emotional change and cases in which positive or negative emotion changes were reported. Overall, the contributions of this work can be summarized as:

- We contribute to the field of Human-Computer Interaction and Affective Computing by demonstrating that ESM response tasks can affect the participant's emotional states even though the method is frequently employed to collect ground-truth labels of instantaneous emotions in naturalistic settings.
- Thus, our work suggests that it is necessary to pay attention to emotion changes during ESM response tasks in those studies that investigate psychological states using the ESM.
- Finally, we offer suggestions to refine experience sampling methodology that would allow researchers better control the effects of surveys on users' emotions.

The rest of this work is organized as follows. Section 2 reviews the related work on emotion assessment and interruptibility. Section 3 describes our research methods in detail, including survey design, sensor data collection, personal trait data collection, and real-world data collection. Section 4 provides our main results and identifies contextual factors related to emotion change. Finally, we conclude by discussing our findings and suggestions to improve the ESM in the context of emotion data collection in Section 5 and 6.

2 BACKGROUND AND RELATED WORK

2.1 Experience Sampling Method

The Experience Sampling Method (ESM) [10], also known as the Ecological Momentary Assessment (EMA) [73], is widely used as an observational tool to study people's behavior and experiences in their daily lives. For example, it is used to record people's emotions, stress levels, habits, productivity, and personality [29, 51, 84]. The advantage of using the ESM is that a researcher does not directly restrict a participants' actions and thoughts, so it is possible to study them naturally. There is also an advantage in that a deep understanding of a participant's own experiences is reflected in self-reports since it is a method in which a participant directly records his/her behaviors, experiences, and thoughts in a self-motivated manner without intervention from a researcher. There is also an advantage of reducing retrospective recall bias [5, 32] since ESM aims to sample immediately or closely after target events in real-time.

Due to the advancement of mobile devices, research using the ESM is becoming more widespread. For example, there were attempts to use mobile devices such as pagers [17], PDAs [79], and cameras [16] as tools to record a participant's daily life. Furthermore, with the spread of smartphones and wearable devices, research on collecting environmental and contextual information is also increasing. Since users almost always carry these devices, researchers can conveniently understand participants' contexts by simultaneously collecting the target experience (e.g., behaviors, emotions) and ambient data via sensors discretely embedded in mobile devices (e.g., accelerometer, thermometer, location information) with the ESM.

Table 1 shows recent studies that collected people's psychological states (e.g., emotions, stress) via mobile-based ESMs. Under the broad umbrella of sampling protocols, prior studies have mainly considered one of four sampling protocols: signal-contingent, event-contingent, interval-contingent, and voluntary samplings. For signal-contingent sampling, participants respond to ESM questionnaires according to a request at various (or randomized) times throughout the day. These randomized requests help avoid the bias and effects (e.g., systematic bias) [10, 69, 84]. However, randomized requests at inopportune moments can be disturbing since participants must stop their current tasks to respond to the ESM questionnaires. Section 2.3 further reviews how the requests at inopportune moments can potentially impact user responses.

For event-contingent sampling protocol, participants are required to assess what happened at the time of the target events. This protocol is applicable when the research aims to capture specific moments such as tweets [42] and mobility changes [8]. However, using such a collection method for frequently occurring events may increase user interference, so caution is required.

For interval-contingent sampling, participants respond to ESM questionnaires according to requests at fixed times throughout the day (e.g., morning, afternoon, and evening intervals; or night daily). This can be less disturbing than signal-contingent sampling since requests are delivered at predictable times, allowing participants to prepare themselves [10, 39, 69]. For example, participants can rearrange their schedules around requests. However, the anticipation for experience sampling can compromise the reliability/validity of responses, as it allows participants to prepare themselves physically, cognitively, and even emotionally, known as expectancy effects [10]. In addition, a fixation on the request time can cause systematic bias [90]; e.g., participants may always report themselves as being lethargic when requests are constantly delivered after dinner.

Participants do not receive any ESM request signals or notifications for the voluntary sampling protocol but initiate responding to ESM questionnaires spontaneously at their will [84]. Similar to interval-contingent sampling, this sampling is also less disturbing since participants can freely decide when to record their states. For example, participants can choose not to respond to ESM questionnaires when physically and cognitively overloaded. However, this may introduce sampling bias, having an unequal chance of different states being measured. For example, responses may only

Table 1: ESM studies measuring emotion or stress. The average number of daily responses was reported for event-based and voluntary sampling strategies. Event = ESM triggered when a predefined event occurs. Interval = ESM triggered at a predefined interval or schedule. Signal = ESM triggered at a random interval. t/day = times per day, n/s = not specified.

Studies		Mood/stress items	Scale references	# Total ESM items	Sampling protocols	# Daily requests
[54]		8-item mood scale 4-item stress scale	PANAS like custom scales [55]	13	Signal (10 ran. t/day)	10
[53]		1-item stress scale	PANAS like custom scales [55]	25	Signal (10 ran. t/day)	10
[86]		9-item mood scale 1-item stress scale	PANAS like custom scales [55]	19	Signal (12 ran. t/day)	12
[41]	(Study 1)	21-item mood scale	Circumplex mood model [62], PANAS [88]	71	Signal (10 ran. t/day)	10
	(Study 2)				Signal (50 ran. t/day)	50
[42]		1-item mood scale	Basic emotions [22]	1	Event (tweet)	22
[44]		2-item mood scale	Circumplex mood model [62]	2	Interval (every 3 hours)	n/s
[6]		1-item stress scale	n/s	7	Interval (every evening)	n/s
[87]		1-item mood scale 1-item stress scale	Photographic affect meter [60], Single item stress measure [76]	7	Signal (random but, different day-by-day)	8
[29]		5-item stress scale	Perceived stress scale (PSS) [14]	5	Signal (15 ran. t/day)	15
[23]		6-item mood scale	MDMQ [92]	78	Interval (hourly), Event (message, call, etc.), Voluntary (anytime)	11
[26]		2-item mood scale	Circumplex mood model [62]	2 (17 optionally)	Voluntary (at least 2 t/day)	n/s
[51]		2-item mood scale 1-item stress scale	Circumplex mood model [62] Tense arousal [66]	3	Signal (20 ran. t/day)	20
[19]		12-item mood scale	Profile of mood states [70], PANAS [88]	32	Signal (8 ran. t/day)	8
[74]		1-item mood scale	n/s	7	Voluntary (at least 3 t/day)	n/s
[38]		12-item stress scale (PSS 4-item, 8 extra items)	PSS 4-item (PSS-4) [13]	13	Signal	n/s
[9]		6-item mood scale	PANAS like custom scales (n/s)	19	Signal (8 ran. t/day)	8
[91]		12-item mood scale	PANAS like custom scales (n/s)	14	Signal (10 ran. t/day)	10
[68]		8-item mood scale	PANAS like custom scales (n/s)	19	Signal (10 ran. t/day)	10
[89]	(Study 1)	3-item mood scale	PANAS like custom scales (n/s;	8	Signal (6 ran. t/day)	6
	(Study 2)		only negative affect scale)	11	Signal (12 ran. t/day)	12

be registered at specific periods (e.g., evening) or psychological state (e.g., boredom), resulting in a skewed dataset.

This study considers signal-contingent sampling protocol since it is most widely used to collect psychological state data (e.g., emotions, stress) in-situ. For this sampling method, researchers need to decide on an appropriate number of daily requests. As shown in Table 1, the previous literature shows considerable variations in the number of daily requests, as different research objectives each require a different level of detail on the collected data. Nevertheless, studies on average opt-in for 10 ESM requests per day, as sufficient observations are essential to capture substantial fluctuations in daily experience (e.g., emotion and stress). Furthermore, the number of daily requests tends to increase when a questionnaire consists of a few items (or takes a short time to complete). For example, Kuppens et al. [41] designed a study where participants complete a single-item questionnaire fifty times a day. Similar to our review, the literature suggests that people are more willing to respond to an ESM questionnaire frequently when the questionnaire has a small number of items and/or does not demand considerable mental effort [16, 84]. In line with this suggestion, Eisele et al. [20] showed that people perceive a relatively lower user burden for an ESM questionnaire with fewer items. In this study, we used a comparably small number of items for the ESM questionnaire (i.e., 7 items), and we expect the participants' perceived burden would be small or moderate. More details of our ESM questionnaires and settings can be found in Section 3.1.

2.2 Emotion and Stress Assessment

Table 1 summarizes recent studies that collected psychological states (e.g., emotions, stress) of participants with the ESM. Previous studies mainly employed the following surveys and emotion scales to collect emotional states: positive and negative affect schedule (PANAS) [88], the basic emotions [22], and the circumplex model [62]. The PANAS is a questionnaire assessing positive and negative affect, with scores measuring the level of agreement with each of the listed positive and negative words. Ekman's basic emotion theory summarizes various human emotions into one of the predefined six categories of basic emotions (i.e., happiness, surprise, anger, disgust, sadness, and fear) [22]. Russell's circumplex model projects emotions onto a two-dimensional space defined by two

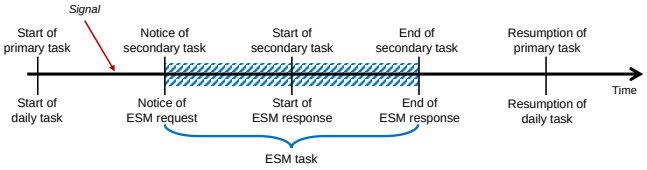

Figure 1: A sequential description of the interruption in the context of ESM response tasks.

independent emotion vectors (i.e., valence and arousal). Note that although valence and arousal are strictly referred to as affects in the literature, we described them as emotions to our participants in the experiment as it is a more prevalent word; hence our paper will use affect and emotion interchangeably to denote valence and arousal. The methodologies for assessing emotions with a combination of affect vectors have been studied in various ways. For example, Thayer [77] suggested two additional emotion vectors: energetic and tense arousal, arguing that Russell's valence vector is not independent but instead expressed as a combination of active and tense arousal. Studies extending the combination of emotion vectors in two axes, further to multidimensional axes, presented the concept of multidimensional mood questionnaire (MDMQ) and reported the results of studying basic dimensions of affect to assess emotions. With MDMQ, researchers have mainly explored three major affect dimensions (i.e., valence, calmness, and energetic arousal) and reported measuring them with the different numbers of survey items or translating the survey items using various auxiliary emotion words [48, 72, 92].

For the stress level, the surveys were mainly conducted using questionnaires based on Cohen's Perceived Stress Scale (PSS) [14]. In the case of PSS, after the first 14-item questionnaire was developed, a shorter questionnaire (e.g., 10 and 4 items) has been studied and used more frequently. In addition, researchers also used a single-item questionnaire to reduce the burden in mobile environments (e.g., what was your stress level over the last few hours? [1-5 points] [38, 51]).

Our work carefully reviewed ESM survey items to capture emotion states and stress levels and changes in them by adapting the questionnaires discussed above. The detailed survey items are presented in Section 3.

2.3 Interruptibility and Emotion Change

ESM for emotion research aims to sample users' psychological states (e.g., emotion and stress) in real-time immediately following the occurrence of target events. As shown in Figure 1, when an ESM response task is delivered amid an ongoing task, a user needs to suspend the ongoing task to switch tasks; thus, delivering an ESM task at random will likely interrupt a user's ongoing tasks. The impact of interruptions has been studied by prior studies on interruption management [1, 64], which refers to an ongoing task as a primary task and an interrupting task (e.g., ESM task) as a secondary task.

Interrupting a primary task at an inappropriate time may considerably influence the performance of primary and secondary tasks and user experiences [64]. At the interruption, a user's attention is drawn to a secondary task (e.g., noticing notification sound or vibration signal). The influence of this interruption can last until a user finishes a secondary task and resumes a primary task, and even after the resumption. However, an inopportune interruption may delay the transition interval that separates the point of noticing and starting the secondary task (i.e., interruption lag) and the time interval separating the end of the secondary task and resuming the primary task (i.e., resumption lag). Thereby, such lags can lead to decreased performance of primary tasks [52]. Furthermore, prior works studied influences of transition lags due to task interruptions on users' emotions, cognition, and contextual determinants and reported that interruptions could induce negative feelings (e.g., stress [46], annoyance [3], anxiety [4]) and increase workload [46] in a variety of environments (e.g., office work [75] and driving [35]). In addition, many studies have reported that task interruptions are influenced by various contextual factors, such as changes in physical activity [25], changes in conversation [33], calendar information [71], messages from different social relationships [49], and difference in personality traits [50].

In prior studies on interruption management, mobile application response tasks (e.g., replying to text messages) were widely used as secondary tasks. Since these tasks are similar to the ESM response task in our study, we can expect that ESM response tasks delivered at unexpected moments can interfere with participants' primary tasks and influence their emotional states as participants need to suspend their primary task and switch to an ESM response task. Thus, emotional states before and after task-switching can be different. In addition, many studies have assessed users' emotions and stress levels using ESMs, but, to our knowledge, no study has investigated how emotions may change due to responses to ESM tasks. Therefore, this paper aims to study whether and how ESM response tasks and associated factors cause changes in emotions. Specifically, we study if and how ESM response tasks affect people's emotional states (RQ1), what factors are related to emotion changes (RQ2), and what are typical contexts when emotion changes occur due to ESM response tasks in daily circumstances (RQ3).

3 METHODOLOGY

In this section, we explain our methodology for collecting the data presented in this paper. Our primary goal in constructing the dataset was to collect emotions arising in various daily circumstances and contextual information associated with them. Towards that goal, we collected various mobile and wearable sensor data and developed a questionnaire to understand how ESM response tasks lead to changes in people's emotions. This section covers the following: (i) the procedure of designing an ESM questionnaire for collecting psychological states (e.g., emotions and stress levels), (ii) our method for collecting contextual information associated with experience sampling responses, including physiological signals and smartphone usage histories, (iii) our approach for categorizing participants' personality traits, and (iv) the process of collecting people's daily life data in the wild.

3.1 ESM Survey Questionnaire Design

3.1.1 ESM Questionnaire. We designed a new ESM questionnaire based on existing questionnaires that assess psychological states.

Table 2: Final version of the ESM questionnaire. (Q1: Valence, Q2: Arousal, Q3: Attention level, Q4: Stress level, Q5: Emotion duration, Q6: Task disturbance level, Q7: Emotion change)

My emotion right before doing this survey was			
Q1. very negative (-3)	~	very positive (+3)	[ ]
Q2. very calm (-3)	~	very excited (+3)	[ ]
My attention level right before doing this survey could be rated as			
Q3. very bored (-3)	~	very engaged (+3)	[ ]
My stress level right before doing this survey was			
Q4. not stressed at all (-3)	~	very stressed (+3)	[ ]
My emotion that I answered above has not changed for recent __ minutes.			
Q5. [5, 10, 15, 20, 30, 60 min / I am not sure]			
Answering this survey disturbed my current activity			
Q6. entirely disagree (-3)	~	entirely agree (+3)	[ ]
How did your emotions change while you are answering the survey now?			
Q7. I felt more negative (-3)	~	I felt more positive (+3)	[ ]

Table 3: Description of ESM design variables and corresponding settings in this study.

Design variables	Description	Study Setting
Daily activation duration	The time between the first request and the last request (e.g., 8 AM~8 PM)	12 hours of regular waking hours (i.e., 10 AM~10 PM)
Request period	How often ESM tasks are delivered during the activation duration (e.g., 60 minutes on average)	An average of 45 minutes
Minimum request time interval	The minimum time interval between two consecutive ESM tasks (e.g., 25 minutes)	30 minutes
Response expiry time	The response time limit after the ESM task arrival (e.g., every ESM questionnaire expires 5 minutes after its arrival)	10 minutes
The number of request limits	The maximum number of response requests delivered during the activation duration (e.g., 20 times per day)	16 times per day

In particular, it includes a question to measure changes in emotions by asking participants to compare the difference in emotions before and after they received ESM response tasks, using phrases commonly used in questionnaires previously used for studies on interruptibility. We validated our new ESM questionnaire with 4 pilot studies in which 4~5 volunteers participated per study. We briefly report how we reviewed each item in our questionnaire in the following.

Similar to prior studies [38, 51], we used multidimensional probing for our ESM questionnaire (Section 2.1) since emotion vectors require fewer survey items than other long-form assessments such as PANAS [88] or PAM [60]. After reviewing and testing the emotion words from these surveys and Russell's circumplex model (valence-arousal) [62], we used two emotion vectors of valence (negative-positive) and arousal (calm-excited).

As discussed in Section 2.3, ESM response tasks can influence participants' task performance and emotional states. Accordingly, by adding an emotion change (negatively-positively) questionnaire item, we guided participants to report changes in their emotional state. We explicitly asked for emotion change similar to prior interruptibility studies where researchers explicitly asked participants to label how disruptive incoming messages are [43, 50], which is known as the explicit labeling of interruptibility [83]. To capture the extent current activity is interrupted by the ESM response tasks at its arrival, we also included questionnaire items on stress level (low-high), attention level (bored-interested), and task disturbance level (low-high). In addition, we added an item that probes how long the current emotion has lasted for a potential study of emotion label imputation for missing ESM responses, but this item was excluded from the scope of the current paper.

3.1.2 Pilot Studies. Prior to data collection, we reviewed the questionnaire via four iterations of pilot studies. We verified whether the questionnaire items could be easily understood from the first to the third rounds of pilot studies ($N_{1,2,3} = \{4, 5, 4\}$). All participants were invited on-site and were instructed to respond to our questionnaire every 30 minutes while doing their business, such as reading research papers, doing homework, and watching videos for two hours. Participants reported that the question about emotion change is comprehensible; they could assess the change in their emotions by with the question asking whether their emotions changed. In addition, *interested*, one of the auxiliary words was replaced by *engaged* since several participants suggested that the word was difficult to differentiate between *excited* for arousal and *interested* for attention level.

In the fourth round ($N_4 = 5$), we verified our ESM requesting mechanism through a 3-day test that delivered survey requests 10 times per day at an average interval of one hour during the daytime. As a result, we increased the average number of requests per day as we received only 20 responses per person out of 30 survey requests. The interval between requests was also shortened so that participants could respond to over 10 ESM responses per day. Finally, we adjusted the length of our questionnaire according to the suggestion to adjust survey lengths that it can be completed within two minutes [16]. In addition, Eisele et al. [20] showed that people perceive less user burdens for an ESM questionnaire with a smaller number of items. In this study, we considered a small number of items for the ESM questionnaire (i.e., 7 items) that can be answered in approximately one minute. See Table 2 for the final version of the questionnaire.

3.1.3 ESM Application and Setting. For delivering ESM response tasks, we used PACO [18], an open-source app for conducting ESM-based research. We instructed participants to install the PACO app on their smartphones and respond to questionnaires delivered at preset intervals. We carefully selected at which intervals ESM response tasks should be delivered to participants. While giving the tasks as often as possible can allow acquiring rich ESM responses, sending too many can increase the response burden. Based on the findings from recent studies on ESM request delivery settings [67, 85], the following design variables were considered: (i) *daily activation duration*, (ii) *request period*, (iii) *minimum request time interval*, (iv) *response expiry time*, (v) and *the number of request limits*. The description of these design variables and corresponding settings in our study are shown in Table 3.

As shown in Table 1, there are variations in the number of requests. In this study, we aimed to collect at least 10 responses per participant per day, which is slightly lower than the average number of observations in prior studies (Mean=13.9, SD=10.2; Mdn=10; range=6 ∼ 50). Therefore, we asked our participants to respond to at least 10 requests among 16 randomly triggered requests daily. In addition, participants were notified that they could change the ringer modes of their phones to silent when they seemed notification alarms inappropriate for the circumstances.

Similar to recent ESM studies [23, 51], the daily activation duration was set to 12 hours of regular waking hours (i.e., 10 AM∼10 PM). The request period was set to an average of 45 minutes, and the minimum request time interval was set to 30 minutes (60 minutes at maximum). We set the maximum number of requests per day to 16 times, so each participant received 15 or 16 ESM tasks per day. In the case of ESM surveys in a mobile environment, a prior study reported an average response rate of about 70% [84]. Assuming a similar response rate in our study, we expected 11 ($= 16 * 0.7$) responses per person each day, which satisfies our goal as mentioned earlier (i.e., 10 or more responses per participant per day). As Scollon et al. [69] and Eisele et al. [21] described that the greater time lag between ESM response and the corresponding request could compromise the quality of ESM data due to recall bias, we set our response expiration time to 10 minutes to reduce the bias.

3.2 Contextual Data

We collected physiological signals widely used in emotion and stress-related research for contextual information, including *electrocardiogram* (ECG), *electroencephalogram* (EEG), *galvanic skin response* (GSR), *plethysmography* (PPG), and *human skin temperature* (HST). Conventionally, stationary medical-grade equipment is used for measuring these signals under stationary laboratory settings. While such devices allow measuring multi-channel electrophysiology data at a higher sampling rate, they are cumbersome to use in daily settings with low mobility as they require an additional power source to operate and are often large in their sizes. Therefore, such devices were not suitable for our study, which intends to investigate the effect of experience sampling on emotions in a daily setting. Instead, we chose commercial wearable devices, *Microsoft Band 2* and *Polar H10* in particular, to collect physiological signals and used heart rates (beats per minute and inter-beat interval), GSR, and HST as primary features in our analysis.

Along with physiological signals, we also collected smartphone usage data, which is also commonly used in affective computing research. We developed and used an app that collects and stores (1) smartphone usage data and (2) sensor data from wearable devices. Before data collection, we instructed users to install the app along with the *PACO app*. After the data collection, the data stored on smartphones were transferred to researchers. Primarily, data related to changes in the location of participants [44], user activities, application usage [11] were used for the analysis.

3.3 Personality Trait Data

Research on the relationship between personalities and smartphones reported individual differences in perceived disruption from smartphone interruptions [50, 93]. For example, extroverts are more likely to experience greater disturbance from interruptions [93]. On the other hand, neurotic or conscientious individuals are faster at responding to interruptions. Based on such findings, we expected that participants would display different response patterns towards ESM based on their personality traits and used the *Big Five Inventory* (BFI) personality test to measure the personality traits of each participant.

BFI is widely employed in personality research, comprised of 44 items measuring an individual's disposition towards five distinct personality traits [31]. Based on that, a shorter version (BFI-S) was developed, and its Korean translation by Kim et al. [37] is a concise Korean BFI survey (K-BFI-15) that we used in our study. K-BFI-15 is comprised of 15 questions in total with three questions for each of five personality traits, which are: (i) openness—how accepting an individual is to intellectual curiosity, changes, and diversity, (ii) conscientiousness—how inclined an individual is to comply with social rules, expectations, and norms, (iii) neuroticism—a degree to which an individual exerts control upon the external environment seeking for mental stability, (iv) extraversion—how much an individual seeks for a relationship, interaction, and attention from others, and (v) agreeableness—an extent to which an individual maintains a comfortable and harmonious relationship with others.

3.4 Real-world Data Collection

Participants were recruited on our university's online bulletin board for the final data collection. Only smartphone owners with an Android smartphone over Android version 6.0 were selected. All recruited participants participated in the experiment for a week and were compensated approximately 70 USD at the end of data collection. Before participating in data collection, all participants attended an offline session where they were given a detailed description of the study and data collection equipment. During the one-hour session, all participants read and signed a consent form approved by our institution's internal Institutional Review Board. In the consent form, we detailed any personal data collected, such as (i) participants' gender and age and (ii) all data collected for the study. Once they signed a consent form, participants were instructed to fill out a K-BFI-15 survey and install required data collection applications and the *PACO app* following detailed instructions from researchers. Researchers assisted participants in installing applications and using data collection wearable devices (*Microsoft Band 2* and *Polar H10*) correctly throughout the session. Finally, the researchers also explained each ESM question item and how to respond using common emotional expressions (e.g., anger, pleasure, frustration, and peace) which participants are familiar with in daily circumstances.

4 RESULTS

Here we present an overview of the collected data and summarize the analysis results for each research question presented in Section 2. Specifically, we describe (i) the overview of ESM survey responses, (ii) how ESM response tasks induced changes in emotions, (iii) what contextual factors are related to changes in emotions, and (iv) what daily instances are associated with emotion changes.

Table 4: Overview of ESM responses.

	Mean	SD	Min	Max
Valence	0.6	1.4	-3	3
Arousal	-0.2	1.7	-3	3
Attention level	0.4	1.6	-3	3
Stress level	-0.3	1.6	-3	3
Task disturbance level	0.1	1.8	-3	3
Emotion change	0.0	0.9	-3	3

4.1 Dataset Overview

Over three 1-week data collection sessions (Apr. 30~May 6, May 8~May 14, and May 16~May 22 respectively), 80 participants were recruited, and a total of 5,753 ESM responses were collected. Overall, more than 71 responses on average (SD = 17, MAX = 110, MIN = 20) were collected per participant. The average number of daily responses was 10.2 (SD = 0.3, MAX = 10.7, MIN = 9.7), meeting our goal of sampling at least 10 samples per person a day. We excluded invalid responses from the initial pool of responses first by excluding ones that were responded after the expiration time (i.e., 10 minutes; see Section 3.1.3), and next by excluding responses without corresponding wearable sensor data. As a result, the remaining 2,227 ESM samples from 78 participants (23 females and 55 males)—with two participants removed due to a lack of valid samples—were used for the analysis. The participants in the final samples were 21.9 years old on average (SD = 3.8, MAX = 38, MIN = 17). In addition, we further excluded *Q5: Emotion duration* from our analysis, as noted in Section 3.1.1. Table 4 summarizes the descriptive statistics of responses to the survey items (Q1~Q7).

Table 5 summarizes the Pearson's correlation coefficient between Q1 to Q7. The last row shows that changes in emotions due to ESM response tasks are significantly correlated with psychological states (valence, arousal, stress level, and attention level) before the experience sampling task and the level of disturbance incurred by the task. In particular, there are moderate correlations of emotion change with valance (r = .368, p < .001) and stress levels (r = −.362, p < .001) but low correlations with arousal (r = .182, p < .001), attention (.123, p < .001), and task disturbance levels (r = −.173, p < .001).

4.2 Amount of Emotion Change

This section reports the result of our analysis in response to RQ1: how much emotion changes occur due to ESM tasks. For this, we focused on responses to *Q7: Emotion change*, "How did your emotions change while you are answering the survey now?" For an intuitive understanding of the responses to Q7, we divided them into three groups: (i) responses reporting a negative emotion change, (ii) reporting no change, and (iii) reporting a positive emotion change. Then within these three groups, we calculated the proportion of responses belonging in a group for each participant, and averaged that number across all participants (see Equation 1, n = 78).

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\#\text{ of participant } i\text{'s responses in a group}}{\text{total }\#\text{ of participant } i\text{'s responses}} \quad (1)$$

Among the 2,227 responses, 900 responses (40.4%), without averaging across participants, reported changes in emotions while answering the survey. In addition, as shown in Table 6, when averaged across participants, the proportion of responses that reported any emotion change was 38.6% (SD = 19.1%). This result suggests that emotion changes induced by answering the survey should not be overlooked when collecting emotions via experience sampling. In other words, ESM responses can be unintentionally biased due to changes in emotions as participants respond to ESM tasks. Thus, to obtain an accurate record of a participant's emotion at a certain point, a researcher must understand the possibility that emotions may change due to the ESM response tasks themselves and devise an approach that will enable participants to reflect upon their emotions carefully.

The 4th and 5th rows in Table 6 show the maximum and minimum value of the response ratio, respectively. It shows that some individuals reported emotion changes in all cases (100.0%), some others reported no emotion change in any case (0.0%). Besides, standard deviations are greater than 19%. This possibly indicates that each participant's proportion of responses across three groups (negative change, no change, and positive change) varies significantly among individuals. Thus, we can infer that individual differences should be considered crucially in analyzing factors related to emotion changes due to ESM response tasks, especially individual-specific characteristics such as personality traits, which we present in Section 4.3.

4.3 Contextual Factors for Emotion Change

According to the analysis on RQ1 in Section 4.2, participants reported that emotion changes occurred in more than 38% of instances while responding to ESM questionnaires, thus confirming that emotional states when answering ESM questionnaires are affected by the very act of engaging in experience sampling. In this section, we examine the contextual factors related to emotion changes occurring amid experience sampling (RQ2). We first define a set of contextual factors. Then with the repeated measures ANOVA, we

Table 5: The correlation matrix among ESM responses. (*: p < .05, **: p < .01, ***: p < .001)

	Valence	Arousal	Attention level	Stress level	Task disturbance level
Valence					
Arousal	.390 ***				
Attention level	.349 ***	.465 ***			
Stress level	-.602 ***	-.230 ***	-.178 ***		
Task disturbance level	.006	.122 ***	.209 ***	.120 ***	
Emotion change	.368 ***	.182 ***	.123 ***	-.362 ***	-.173 ***

Table 6: Proportion of ESM responses reporting changes in emotions, averaged across participants.

	Emotion change negatively ①	No change	Emotion change positively ②	Emotion change ① + ②
Mean	20.3%	61.4%	18.2%	38.6%
SD	19.1%	24.8%	19.2%	19.1%
Max.	100.0%	100.0%	69.4%	100.0%
Min.	0.0%	0.0%	0.0%	0.0%

observe the relationship between emotion changes and contextual factors. Finally, we examine and compare the influence of all factors through multilevel regression.

4.3.1 Definition of Contextual Factors. We selected a set of contextual factors following a procedure as shown below.

(1) **Selecting a feature set**: We first selected features that commonly appear in studies involving emotions, stress, and interruptibility. The common features primarily include physiological signals (e.g., heart rate, body temperature, sweating) and smartphone & wearable sensor data (e.g., physical movement, location change, phone usage). Specifically, the following features were commonly used for estimating emotions and stress: heart-beat related features (e.g., heart-rate variability (HRV), blood volume pulse (BVP)) [23, 47], electrodermal activity (EDA) [65], human skin temperature (HST) [15], and smartphone usage and embedded sensors data (e.g., location, app usage, activity type, communication history, environmental contexts including sound and light, etc.) [6, 23, 42, 44, 65].

(2) **Feature extraction**: First, we extracted the following contextual features from the physiological data collected with *Microsoft Band 2*: heart-beat per minute (BPM), inter-beat interval (IBI), EDA, HST, and the wrist acceleration (ACC). Specifically, the mean and standard deviation of BPM, IBI, EDA, HST, and ACC were calculated from signals collected within a one-minute window preceding an ESM response [58]. Note that the choice of one-minute window size was based on an empirical observation that gave the most significant result for our regression analysis. Also, we only considered signals collected in low-mobility situations such as standing and sitting for the analysis since the quality of wristband data is known to be poor under high mobility circumstances [57, 61], using the automatic motion type detection of *MS Band 2*. Next, we included the phone usage and the location change features extracted from participants' smartphones. For each ESM response, we calculated a phone use time as the cumulative duration a participant used mobile applications, between the most recent smartphone unlock preceding an ESM response and the response. The location feature was calculated in two steps: (i) identify top three locations that a participant most frequently stayed from smartphone's GPS logs [94], and (ii) compute whether a participant visited these top three locations in the past one hour from the time of an ESM response with one-hot encoding. We call these features *Location 1, 2, and 3*. We also considered temporal contexts of ESM responses, which include the following: whether the responses were answered at weekend or weekday, time of day ($0-23$ hour), day of week (Day 1 to 7), the number of experiences sampled, number of previous responses in the same day, and number of prior requests in the same day.

4.3.2 Repeated Measures ANOVA Analysis. We conducted a series of one-way Repeated Measures ANOVA (RM-ANOVA) to analyze how responses to *Q7: Emotion change* depend on psychological states (valence, arousal, attention level, stress level) prior to the ESM response task and the level of disturbance induced by the task (see Table 7). As in the previous section, the responses were divided into three groups: *emotions changed negatively*, *no change*, and *emotions changed positively*. Particularly, we calculated the average for each group and used them in the analysis. For each RM-ANOVA, we adjusted the degree of freedom according to the significance of Mauchly's Test of Sphericity. For the effect size, we report partial η^2 (Eta squared). For post-hoc comparisons, we adjusted the p-values based on the Bonferroni correction.

Table 7 summarizes the RM-ANOVA results. The table shows there are noticeable patterns associated with different independent variables. Based on Cohen's guidelines on effect size interpretation [12], we find that all psychological states (*Valence*, *Arousal*, *Attention level*, *Stress level*) prior to the ESM response task and the *Task disturbance level* show high effect size, indicating these factors have significant influences on emotion changes. That is, participants' emotions tended to positively change when they were feeling positive, i.e., when they were feeling emotionally aroused with a low stress level, and the ESM survey did not interfere with their tasks.

4.3.3 Multilevel Regression Analysis. Next, we conducted a multilevel regression analysis to explore how each contextual factor accounts for emotion changes. Before the analysis, we checked whether the dataset met the major assumptions for regression analysis (e.g., normality, homoscedasticity, and multicollinearity). In this process, we excluded the *IBI mean* due to the violation of the multicollinearity test. We also excluded the majority of variables describing temporal contexts, such as the number of prior responses in the same day, as they failed the multicollinearity test with a high correlation with each other. We only included *Time of day* and *Day of the week*. For the measure of goodness-of-fit, we report an adjusted R-squared value, which indicates variance explained by both fixed and random effects. ESM responses or subjective contextual factors represent the psychological states (e.g., valence) prior to the ESM response task and the disturbance level of the task. As demonstrated in Table 8, in addition to subjective contextual factors, individual-specific factors (e.g., age), which can represent individual differences in participants, and objective contextual factors were considered together. We applied min-max normalization for the objective contextual factors.

As shown in Table 8, most of the subjective contextual factors except for Arousal were statistically significant ($p < .05$). In comparison, none of the individual-specific factors were statistically significant, while only some objective contextual factors were statistically significant ($p < .05$).

Valence, *Attention level*, *Stress level*, and *Task disturbance level* were statistically significant among the subjective contextual factors. The coefficient values of *Valence* ($\beta = .133$, $p < .001$) and *Attention level* ($\beta = .027$, $p < .05$) are positive. Thus, we can say that when a participant had a high level of valence or attention level prior to the ESM response task, the participant's emotions were more likely to change positively. The negative coefficients of *Stress level* ($\beta = -.100$, $p < .001$) and *Task disturbance level* ($\beta = -.108$, $p < .001$) suggest the opposite. A participant's emotions tended to change negatively when a participant had a high level of stress prior to the ESM response task, and similar with the disturbance level, a participant's emotions tended to change negatively if the perceived disturbance of the ESM task was high.

Among the objective contextual factors, *Location 1*, *Phone use time*, *BPM std*, and the interaction factor between *Weekend* and *Time of day* were statistically significant. Given that the coefficient of *Location 1* ($\beta = -.099$, $p < .05$) is negative, we can infer that

Table 7: Results of repeated measures ANOVA.

	Emotion change			DF	F-value	p-value	Partial η^2
	Negatively	No change	Positively				
Valence	-0.06 (1.02)	0.74 (0.80)	1.27 (0.90)	1.850, 77.704	35.718	<.001	0.460
Arousal	-0.54 (0.96)	-0.18 (0.94)	0.27 (1.16)	1.768, 74.264	11.961	<.001	0.222
Attention level	0.03 (1.15)	0.31 (0.77)	0.85 (1.03)	1.729, 72.620	11.724	<.001	0.218
Stress level	0.62 (1.00)	-0.43 (0.83)	-0.95 (0.99)	1.684, 70.708	53.009	<.001	0.558
Task disturbance level	0.62 (1.37)	0.12 (1.09)	-0.27 (1.55)	1.484, 62.342	8.893	<.001	0.175

Table 8: Results of GLMM analysis. (*: $p < .05$, **: $p < .01$, ***: $p < .001$)

Independent variables		Coefficient		t value	Significant	
		β	SE			
(Intercept)		-0.154	0.481	-0.319	.750	
Subjective contextual factors	**Valence**	0.133	0.017	8.259	<.001	***
	Arousal	0.020	0.013	1.528	0.127	
	Attention level	0.027	0.013	2.074	0.038	*
	Stress level	-0.100	0.014	-7.144	<.001	***
	Task disturbance level	-0.108	0.012	-9.136	<.001	***
Individual -specific factors	**Age**	-0.003	0.011	-0.293	0.771	
	Gender (Female)	0.019	0.092	0.210	0.835	
	Openness	0.012	0.014	0.832	0.409	
	Conscientiousness	-0.009	0.018	-0.503	0.616	
	Neuroticism	0.008	0.015	0.493	0.624	
	Extraversion	-0.003	0.014	-0.243	0.809	
	Agreeableness	-0.019	0.017	-1.148	0.255	
Objective contextual factors	**Location 1**	-0.099	0.045	-2.200	0.028	*
	Location 2	0.082	0.049	1.688	0.092	
	Location 3	-0.010	0.048	-0.206	0.837	
	Phone use time	-0.386	0.190	-2.031	0.042	*
	ACC mean	0.079	0.309	0.254	0.799	
	ACC std	0.210	0.138	1.520	0.129	
	EDA mean	0.223	0.276	0.809	0.419	
	EDA std	-0.252	0.191	-1.315	0.189	
	BPM mean.	0.170	0.154	1.100	0.271	
	BPM std	0.477	0.163	2.932	0.003	**
	IBI std	-0.227	0.136	-1.666	0.096	
	HST mean	0.127	0.165	0.771	0.441	
	HST std	-0.382	0.441	-0.868	0.386	
	Weekend (T/F)	0.154	0.082	1.885	0.060	
	Time of day	0.072	0.075	0.961	0.337	
	Weekend×Time of day	-0.300	0.146	-2.057	0.040	*
Adjusted R^2						0.297

participants' emotions tended to change negatively if they stopped by their most frequently visited place (i.e., workplace) within an hour. Similarly, the longer the participants used their smartphones before answering ESM questions ($\beta = -.386$, $p < .05$), the more likely their emotions changed negatively. On the other hand, the coefficient of *BPM std* is positive ($\beta = .170$, $p < .01$), indicating that emotions tended to change positively if a participant has high variability in heartbeats. In addition, the coefficient of the interaction factor between *Weekend* and *Time of day* is negative ($\beta = -.300$, $p < .05$). This crossover interaction [45], where the interaction effect is statistically significant while the two main effects are not significant, indicates that the effect of *Time of day* on emotion changes was opposite depending on *Weekend*, quite unsurprisingly showing that emotions tended to change negatively as time passed during a weekend.

4.4 Real-life Instances of Emotion Change

Through the analysis of RQ1 and RQ2, we found that emotion changes occur when responding to ESM questionnaires, and various factors are associated with such changes in emotions. In this section, we report our findings from a qualitative investigation on the typical daily contexts in which emotion changes occur in the context of experience sampling (RQ3). For this, we interviewed 15 participants (7 females and 8 males) who were randomly selected after the data collection (23.9 years old on average ($SD = 3.2$, $MAX = 31$, $MIN = 19$). We conducted a qualitative analysis with structured interview transcripts to find common themes among the statements of subjects related to emotion changes using affinity diagramming [28]. From the qualitative analysis of interview responses, we sorted the contexts in which emotion changes occur into three categories: no change, positive change, and negative change.

Our quantitative results showed that the disturbance level of ESM tasks was an important factor. Similarly, our participants reported that in the case of negative emotion change following an ESM survey, ESM tasks induced annoyance when delivered at inopportune moments and disturbed their primary task. In detail, the most negative emotion changes were reported in the following two cases (denoted as NC, negative case): NC1) an ESM task interrupting a primary task to diminish an individual's attention and NC2) the action of engaging in the ESM task being misaligned with social norms.

Most negative emotion changes fall into the first case (NC1); interviewees responded that they experienced negative emotion changes when their daily tasks such as office work, studying, gaming, and cooking were disturbed with an ESM task. For example, P1 responded "*I think it was a little more disturbing if this [ESM task] came in when I was reading a [research] paper*" and P11 noted "*I felt a bit annoyed when I received the notification as I was working on my assignment.*"

The second case (NC2) was reported predominantly in the context of public spaces and face-to-face conversations. In public spaces such as libraries, theaters, and offices, sudden ESM alarms are likely to cause disturbance and be considered a violation of a social norm. P7 reported "*I felt sorry to disturb others with the notification vibration in our office*" and P12 said "*I tended to have a negative change*

if the notification suddenly came in when I was around many people." Interviewees also responded that they were embarrassed and felt sorry when their phone emitted notification sounds or vibrations during social interactions such as meetings and conversations. Most interviewees responded that they felt such negative feelings stronger when interacting in a small group. P7 felt sorry because she had to pretend to be listening to a lecturer while responding to the questionnaire. P14 answered that he felt stressed because it was difficult to look at the survey screen, especially when he was in one-on-one conversations.

We also summarize the cases of positive changes in emotions that were relatively less frequently mentioned in our interviews than no change and negative changes. There were primarily the three cases (denoted as PC, positive case): PC1) when an ESM task helped to avoid the current uncomfortable (or stressful) situation, PC2) when an ESM task led to a better understanding of their current emotions, and PC3) when an ESM task helped to recall positive memories in the past. While the first positive case (PC1) is similar to the first negative case (NC1) in that they both break the current state of attention, PC1 differs from NC1 as a positive change is induced as an ESM task helps escape from the current tasks that cause stress and boredom by providing an opportunity to switch contexts. In other words, the task disturbance or interruption did not always result in negative emotion changes. Most interviewees responded that they felt refreshed as they could distance themselves from pressing tasks such as tedious paperwork by responding to ESM questionnaires. P6 responded "*I was blocked while composing, but it was nice to be interfered [by ESM task] and be able to take some rest.*" P14 reported "*When I was stressed, I felt relaxed while I answered the questions*" as that helped to break away from the current task.

The second case (PC2) was when an ESM task helped participants better understand their current emotions. It was previously suggested that self-reporting emotions could promote behavior change, as self-quantification provides an opportunity to reflect upon internal feelings [27]. For example, P6 reported "*I felt good about winning a computer game, and I felt better when I checked this (Q7: I felt more positive)*".

The third case (PC3) was when an ESM task acted as a reminder. We further identified two sub-themes under this case: i) ESM as a reminder of some tasks and ii) the reminiscence of past positive memories. According to the encoding specificity principle [82], revisiting a particular behavior that was encoded together with a memory in the past can trigger retrieval of the memory later. It was also suggested that sounds and vibrations, like as included in ESM notifications, can induce priming effects [81] to resurface memories and experiences. For example, P4 responded during the interview that he felt better when he received an ESM notification, as it helped him remember the positive feeling that he felt yesterday when he accomplished an achievement while playing a video game.

5 DISCUSSION

As the experience sampling method (ESM) is widely used in studies of psychological states (e.g., emotion and stress) for collecting affective data, our study examined the relationship between mobile experience sampling and changes in emotions it induces. In this paper, we presented quantitative and qualitative evidence for emotion changes during sampling of psychological states via the ESM and further identified contextual factors associated with emotion changes. Our results showed that responding to experience sampling questionnaires can influence users' emotions positively and negatively. Statistically significant factors from our analysis were valence, attention level, stress level, task disturbance level, location, smartphone use time, heartbeat variation, and time of day during the weekend. In the following, we further discuss the implications of our main findings, guidelines for mobile ESM studies that aim to work with emotions, possible future research directions, and our work's limitations.

5.1 Function of ESM Response Tasks on Emotion Changes

Our results suggest that when collecting emotions via the ESM, responses could be unintentionally biased due to the emotion changes the ESM itself induces—thus, responses do not reflect unaffected, pure emotions. In 40.4% of ESM responses we collected, participants experienced changes in emotions due to experience sampling. Our participants' emotions changed negatively when their primary tasks (e.g., office work, studying) were disturbed by an ESM task. Similarly, prior studies also showed that interrupting a primary task induces negative emotions (e.g., stress [46], annoyance [3]). Interestingly, for approximately half of the cases involving emotion changes, emotions became more positive. Emotions were reported to have become more positive when ESM tasks helped people distance themselves from stressful daily tasks. Prior work also reported a similar finding, where in workplaces, off-task behaviors or non-work-related secondary tasks affected workers' emotions positively as they helped workers escape stressful situations and release stress [30]. Beyond the benefit of off-task multitasking, we also observed that ESM tasks could amplify existing emotions by enabling self-reflection. For example, Hollis et al. [27] reported that self-reflection of emotions (or emotion tracking) increases awareness of emotional consequences; for instance, those who realize current positive emotions can feel more positive going forward. Similarly, our participants often reported that their emotions changed more positively when ESM tasks allowed them to understand their emotions better.

5.2 Lowering Emotion Change with Context-aware Scheduling

As discussed in Section 4.3, the task disturbance level also significantly influences emotion changes. Therefore, reducing the disturbance an ESM task induces can help collect stable responses, and the disturbance level can be controlled by adjusting the point in time ESM tasks are delivered by predicting opportune moments when users can seamlessly engage in ESM tasks [34, 59]. Previous works [59, 95] suggest that the level of task disturbance or interruptibility can be predicted from physiological sensor data. An alternative approach is to request experience sampling at moments already known to be opportune. For example, we can deliver ESM response tasks at activity breakpoints (e.g., when a user's activity status changes from "walking" to "stationary") [56], i.e., context-switching points. However, for collecting affective data,

this selective sampling method based on opportune moments can result in skewed data in favor of certain emotions, as the interruptibility at a given moment depends on emotions at the moment. For example, experiencing negative affective states or a depressive mood, such as when the mental workload is high [2, 75] or sick, is known to be inopportune moments. Thus, it is likely that the selective sampling approach based on interruptibility will sample emotional states highly correlated with preexisting emotional states, i.e., a high level of auto-correlation. Further studies are needed to investigate the impact of selective sampling at opportune moments on the balance of the distribution of collected emotions and the intensity of emotion changes in collected samples.

5.3 ESM Settings and Emotion Changes

An ESM's configuration can vary depending on the purpose of the research. In this study, we considered a signal-contingent sampling, which has been, by far, the most widely used in prior studies that collected emotions and stress through ESM (see Section 2.1). Alternatively, interval-contingent or voluntary samplings are also viable options. These sampling methods are considered less disturbing than signal-contingent sampling [10, 39], so we expect a lower occurrence of emotion changes with these methods. However, caution is required as adverse bias and effects (e.g., systematic bias and expectancy effects) can happen [10, 39]. In addition to a sampling protocol, given that a higher number of items for an ESM task is associated with a higher degree of perceived burden [20], the level of emotion change may increase with the number of items. Given that prior studies often had a relatively higher number of items for their questionnaires (see Table 1), one can expect a higher perceived burden for the ESM tasks in these studies compared to ours with a seven-item questionnaire. We also theorize that the level of emotion changes can differ significantly by the existence of ESM trigger (i.e., notification-initiated response vs. user-initiated response), the medium of ESM trigger (e.g., sound vs. vibration vs. light), and the variation of ESM sampling frequency. In this study, as the first step to identify potential factors related to emotion changes, we relied on a quasi-experiment design where only relative within-subject differences are observed (i.e., comparing emotion changes before and after the ESM tasks) [80]. It was challenging to design a controlled experiment since the number of samples and the data collection periods were too limited to configure diverse combinations of ESM parameters. Therefore, future works could investigate how the aforementioned different combinations of ESM settings affect the level of emotion changes.

5.4 Influence of Personality Traits on Emotion Changes

In our analysis using multilevel regression (Section 4.3), the task disturbance level (assessed by Q6 in our survey) was a statistically significant variable affecting emotion change. It was previously reported that personality traits determine the extent to which an individual is susceptible to a task disturbance [50]. Thus, we too expected that personality traits would similarly be associated with emotion changes during emotion data collection. Nonetheless, our multilevel regression analysis found that personality traits were not statistically significant with respect to emotion changes. We further investigated the relationship between reported task disturbance and personality traits using a multilevel linear regression model with task disturbance as a dependent variable, but personality traits were still not statistically significant in estimating task disturbance. We posit that the difference in the sample size of the previous research (N=11) [50] and ours (N=78) and the difference in statistical models used for analysis may have resulted in such a diverging result, thus calling for a further study to verify the relationship between individual differences, such as personality traits, and emotions or emotion changes acquired with ESM.

5.5 Suggestions for Mobile ESM Studies on Emotions

We provide the guidelines for future mobile ESM studies that aim to collect emotion and stress data. Our findings clearly show that the following details should be considered while planning ESM studies to reduce errors and biases in emotion and stress data: (i) clearly specified reference points for which the users will be asked to report their emotions or stress, and (ii) easy-to-understand emotion words as references that would help users to report their emotions accurately.

Prior interruptibility studies assumed that users could judge whether they would be available for context switching [46, 75]; likewise, our work identified that users could correctly differentiate between their feelings during the primary tasks and those during the secondary task of ESM answering. Therefore, when designing an ESM questionnaire, it is crucial to clarify the distinction between primary and secondary tasks by including the reference phrase such as "right before doing this survey."

During the pilot session of our ESM questionnaire, we found that our participants had difficulties fathoming the emotion words used in the literature (e.g., activation), SAM-based visual explanations [7], and how unconventional affective dimensions such as *valence* and *arousal* map to day-to-day emotions. Therefore, we had to carefully review existing emotion words in the original two-dimensional emotion vector model and check whether our participants had any difficulties understanding those words. In our final design, we denoted the endpoints of the two emotion dimensions in a 7-point Likert scale: *valence* (negative and positive) and *arousal* (calm and excited).

5.6 Limitations

There are several limitations of this work. First of all, there could be non-response bias as subjects may have selectively responded to the ESM requests. For example, during interviews, P2 responded that "*when I did not have time to look at the phone, then I just passed the survey notification.*" Some participants could have intentionally overlooked responding to surveys in certain circumstances (e.g., being under extreme stress or experiencing negative emotions). Overall, it is expected that participants are more likely to avoid answering when they experience extreme emotions both positively and negatively. These neglected responses likely have introduced some bias in our data, which could have resulted in lowering our sample's representativeness. Nonetheless, this does not invalidate our main findings as such neglected responses, while being outliers, are likely associated with even greater changes in emotions levels,

thus further strengthening our finding that experience sampling itself can cause changes in emotions.

Next, the data was collected from a population whose mean age is 21.9 (mainly undergraduate and graduate students). Such subpopulation was selected as they are more familiar with studies like ours, utilizing technical applications and devices. Readers should be aware of the potential limits of our findings' applicability, especially if they seek to apply our findings to a population with characteristics different from our sample population.

Finally, we acknowledge that frequent ESM probing and sensor wearing may have increased user burden, possibly lowering the number of ESM responses, even though a user burden is a well-investigated issue in ESM studies [24, 84] and similar device configurations [47] have been used for user studies. However, despite the concerns that user burdens may have influenced emotion ratings, all responses in our analysis were collected under the same controlled setting. Thus, the influence of burdens induced by ESM probing and sensor wearing on emotion change would likely have been neutral without confounding our analysis results.

6 CONCLUSION AND FUTURE WORK

This paper analyzed 2,227 responses from 78 participants to quantify possible emotion changes associated with ESM tasks. To our surprise, over 38% of the samples from participants reported emotion changes while answering ESM questionnaires. This result indicates that in ESM studies collecting emotion data, it is necessary to acknowledge changes in emotions induced by ESMs, namely the observer effect of an ESM that aims to sample psychological states. A multilevel regression analysis revealed a total of eight statistically significant factors related to changes in emotions when responding to ESM tasks: 1) valence; 2) attention level; 3) stress level; 4) task disturbance level; 5) most recently visited location; 6) smartphone use time; 7) change in heartbeats; and 8) time of day during the weekend.

The experience sampling method (ESM) is widely used in studies that collect the psychological state of participants, and thus, it is necessary to consider an approach that can mitigate changes in users' emotions due to the action of responding to a survey. In our work, participants were instructed to report their emotions before and after an ESM task. In particular, explicit phrases were used in the survey questions to distinguish between the time of data collection and the time of answering an ESM task, and it was confirmed through interviews during pilot studies that participants could recognize whether their emotions changed due to ESM interruptions. This distinction could be leveraged to lower the likelihood of users misreporting their psychological states (e.g., emotion and stress). In future works using the ESM to study emotions, special attention should be paid to carefully devise an ESM instrument that considers the possibility of users' psychological states being influenced due to an act of being interrupted to make self-reports.

ACKNOWLEDGMENTS

This research was supported by the KAIST-KU Joint Research Center, KAIST, and by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Korea government (MSIT) (2020R1A4A1018774).

REFERENCES

[1] Erik M Altmann and J Gregory Trafton. 2002. Memory for goals: An activation-based model. *Cognitive science* 26, 1 (2002), 39–83.
[2] Brian P Bailey and Shamsi T Iqbal. 2008. Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction* 14, 4 (2008), 1–28.
[3] Brian P Bailey and Joseph A Konstan. 2006. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in human behavior* 22, 4 (2006), 685–708.
[4] Brian P Bailey, Joseph A Konstan, and John V Carlis. 2001. The Effects of Interruptions on Task Performance, Annoyance, and Anxiety in the User Interface.. In *Interact*, Vol. 1. 593–601.
[5] Daniel J Beal. 2015. ESM 2.0: State of the art and future potential of experience sampling methods in organizational research. *Annual Review of Organizational Psychology and Organizational Behavior* 2, 1 (2015), 383–407.
[6] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Sandy Pentland. 2014. Daily Stress Recognition from Mobile Phone Data, Weather Conditions and Individual Traits. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 477–486.
[7] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.
[8] Narae Cha, Auk Kim, Cheul Young Park, Soowon Kang, Mingyu Park, Jae-Gil Lee, Sangsu Lee, and Uichin Lee. 2020. Hello There! Is Now a Good Time to Talk? Opportune Moments for Proactive Interactions with Smart Speakers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–28.
[9] Yawen Chan, Suzanne Ho-wai So, Arthur Dun Ping Mak, Kewin Tien Ho Siah, Wai Chan, and Justin C. Y. Wu. 2019. The temporal relationship of daily life stress, emotions, and bowel symptoms in irritable bowel syndrome—Diarrhea subtype: A smartphone-based experience sampling study. *Neurogastroenterology & Motility* 31, 3 (2019), e13514.
[10] Tamlin Conner Christensen, Lisa Feldman Barrett, Eliza Bliss-Moreau, Kirsten Lebo, and Cynthia Kaschub. 2003. A practical guide to experience-sampling procedures. *Journal of Happiness Studies* 4, 1 (2003), 53–78.
[11] Matteo Ciman and Katarzyna Wac. 2016. Individuals' stress assessment using human-smartphone interaction analysis. *IEEE Transactions on Affective Computing* 9, 1 (2016), 51–65.
[12] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.
[13] Sheldon Cohen. 1988. Perceived stress in a probability sample of the United States. *The Social Psychology of Health* (1988).
[14] Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. 1983. A global measure of perceived stress. *Journal of health and social behavior* (1983), 385–396.
[15] Christian Collet, Evelyne Vernet-Maury, Georges Delhomme, and André Dittmar. 1997. Autonomic nervous system response patterns specificity to basic emotions. *Journal of the autonomic nervous system* 62, 1-2 (1997), 45–57.
[16] Sunny Consolvo and Miriam Walker. 2003. Using the experience sampling method to evaluate ubicomp applications. *IEEE Pervasive Computing* 2, 2 (2003), 24–31.
[17] Mihaly Csikszentmihalyi, Reed Larson, and Suzanne Prescott. 1977. The ecology of adolescent activity and experience. *Journal of Youth and Adolescence* 6, 3 (1977), 281–294.
[18] Paco Developers. 2018. PACO - Apps on Google Play. https://play.google.com/store/apps/details?id=com.pacoapp.paco&hl=en
[19] Kari M Eddington, Chris J Burgin, Paul J Silvia, Niloofar Fallah, Catherine Majestic, and Thomas R Kwapil. 2017. The effects of psychotherapy for major depressive disorder on daily mood and functioning: a longitudinal experience sampling study. *Cognitive therapy and research* 41, 2 (2017), 266–277.
[20] Gudrun Eisele, Hugo Vachon, Ginette Lafit, Peter Kuppens, Marlies Houben, Inez Myin-Germeys, and Wolfgang Viechtbauer. 2020. The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment* (2020), 1073191120957102.
[21] Gudrun Eisele, Hugo Vachon, Inez Myin-Germeys, and Wolfgang Viechtbauer. 2021. Reported affect changes as a function of response delay: Findings from a pooled dataset of nine experience sampling studies. *Frontiers in psychology* 12 (2021).
[22] Paul Ekman. [n.d.]. Expression and the nature of emotion. *Approaches to emotion* 3, 19 ([n. d.]), 344.
[23] Anja Exler, Andrea Schankin, Christoph Klebsattel, and Michael Beigl. 2016. A wearable system for mood assessment considering smartphone features and data from mobile ECGs. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 1153–1161.
[24] Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2019. Designing an experience sampling method for smartphone based emotion detection. *IEEE Transactions on Affective Computing* (2019).

[25] Joyce Ho and Stephen S Intille. 2005. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 909–918.

[26] Victoria Hollis, Artie Konrad, Aaron Springer, Matthew Antoun, Christopher Antoun, Rob Martin, and Steve Whittaker. 2017. What does all this data mean for my future mood? Actionable analytics and targeted reflection for emotional well-being. *Human–Computer Interaction* 32, 5-6 (2017), 208–267.

[27] Victoria Hollis, Artie Konrad, and Steve Whittaker. 2015. Change of heart: emotion tracking to promote behavior change. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 2643–2652.

[28] Karen Holtzblatt and Hugh Beyer. 1997. *Contextual design: defining customer-centered systems*. Elsevier.

[29] Karen Hovsepian, Mustafa al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 493–504.

[30] Heming Jiang, Mikko Siponen, and Aggeliki Tsohou. 2021. Personal use of technology at work: a literature review and a theoretical model for understanding how it affects employee job performance. *European Journal of Information Systems* (2021), 1–15.

[31] Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of Personality and Social Psychology* (1991).

[32] Daniel Kahneman, Alan B Krueger, David A Schkade, Norbert Schwarz, and Arthur A Stone. 2004. A survey method for characterizing daily life experience: The day reconstruction method. *Science* 306, 5702 (2004), 1776–1780.

[33] Nicky Kern and Bernt Schiele. 2003. Context-aware notification for wearable computing. In *Seventh IEEE International Symposium on Wearable Computers, 2003. Proceedings.* IEEE, 223–230.

[34] Auk Kim, Woohyeok Choi, Jungmi Park, Kyeyoon Kim, and Uichin Lee. 2018. Interrupting Drivers for Interactions: Predicting Opportune Moments for In-Vehicle Proactive Auditory-Verbal Tasks. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 175 (dec 2018), 28 pages.

[35] Auk Kim, Jung-Mi Park, and Uichin Lee. 2020. Interruptibility for in-vehicle multitasking: influence of voice task demands and adaptive behaviors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–22.

[36] Jonghwa Kim and Elisabeth André. 2008. Emotion recognition based on physiological changes in music listening. *IEEE transactions on pattern analysis and machine intelligence* 30, 12 (2008), 2067–2083.

[37] Ji-Hyeon Kim, Bok-Hwan Kim, and Moon-Sun Ha. 2011. Validation of a Korean version of the Big Five Inventory. *Journal of Human Understanding and Counseling* 32, 1 (2011), 47–65.

[38] Zachary D King, Judith Moskowitz, Begum Egilmez, Shibo Zhang, Lida Zhang, Michael Bass, John Rogers, Roozbeh Ghaffari, Laurie Wakschlag, and Nabil Alshurafa. 2019. micro-Stress EMA: A Passive Sensing Framework for Detecting in-the-wild Stress in Pregnant Mothers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 91.

[39] Predrag Klasnja, Beverly L Harrison, Louis LeGrand, Anthony LaMarca, Jon Froehlich, and Scott E Hudson. 2008. Using wearable sensors and real time inference to understand human recall of routine activities. In *Proceedings of the 10th international conference on Ubiquitous computing*. 154–163.

[40] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing* 3, 1 (2011), 18–31.

[41] Peter Kuppens, Zita Oravecz, and Francis Tuerlinckx. 2010. Feelings change: accounting for individual differences in the temporal dynamics of affect. *Journal of personality and social psychology* 99, 6 (2010), 1042.

[42] Hosub Lee, Young Sang Choi, Sunjae Lee, and IP Park. 2012. Towards unobtrusive emotion recognition for affective social communication. In *2012 IEEE Consumer Communications and Networking Conference (CCNC)*. IEEE, 260–264.

[43] Hao-Ping Lee, Kuan-Yin Chen, Chih-Heng Lin, Chia-Yu Chen, Yu-Lin Chung, Yung-Ju Chang, and Chien-Ru Sun. 2019. Does who matter? Studying the impact of relationship characteristics on receptivity to mobile IM messages. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[44] Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. 2013. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. ACM, 389–402.

[45] Geoffrey R Loftus. 1978. On interpretation of interactions. *Memory & Cognition* 6, 3 (1978), 312–319.

[46] Gloria Mark, Daniela Gudith, and Ulrich Klocke. 2008. The cost of interrupted work: more speed and stress. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 107–110.

[47] Gloria Mark, Yiran Wang, and Melissa Niiya. 2014. Stress and multitasking in everyday college life: an empirical study of online activity. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 41–50.

[48] Gerald Matthews, Dylan M Jones, and A Graham Chamberlain. 1990. Refining the measurement of mood: The UWIST mood adjective checklist. *British journal of psychology* 81, 1 (1990), 17–42.

[49] Abhinav Mehrotra, Mirco Musolesi, Robert Hendley, and Veljko Pejovic. 2015. Designing content-driven intelligent notification mechanisms for mobile applications. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 813–824.

[50] Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. 2016. My Phone and Me: Understanding People's Receptivity to Mobile Notifications. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1021–1032.

[51] Abhinav Mehrotra, Fani Tsapeli, Robert Hendley, and Mirco Musolesi. 2017. MyTraces: Investigating Correlation and Causation between Users' Emotional States and Mobile Phone Interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 83.

[52] Christopher A Monk, J Gregory Trafton, and Deborah A Boehm-Davis. 2008. The effect of interruption duration and demand on resuming suspended goals. *Journal of experimental psychology: Applied* 14, 4 (2008), 299.

[53] Inez Myin-Germeys, Machteld Marcelis, Lydia Krabbendam, Philippe Delespaul, and Jim van Os. 2005. Subtle Fluctuations in Psychotic Phenomena as Functional States of Abnormal Dopamine Reactivity in Individuals at Risk. *Biological Psychiatry* 58, 2 (2005), 105–110.

[54] I. Myin-Germeys, F. Peeters, R. Havermans, N. A. Nicolson, M. W. DeVries, P. Delespaul, and J. Van Os. 2003. Emotional reactivity to daily life stress in psychosis and affective disorder: an experience sampling study. *Acta Psychiatrica Scandinavica* 107, 2 (2003), 124–131.

[55] Inez Myin-Germeys, Jim van Os, Joseph E. Schwartz, Arthur A. Stone, and Philippe A. Delespaul. 2001. Emotional Reactivity to Daily Life Stress in Psychosis. *Archives of General Psychiatry* 58, 12 (12 2001), 1137–1144.

[56] Mikio Obuchi, Wataru Sasaki, Tadashi Okoshi, Jin Nakazawa, and Hideyuki Tokuda. 2016. Investigating interruptibility at activity breakpoints using smartphone activity recognition API. In *Proc. of the 2016 ACM Intern. Joint Conf. on Pervasive and Ubiquitous Computing: Adjunct*. 1602–1607.

[57] Simon Ollander, Christelle Godin, Aurélie Campagne, and Sylvie Charbonnier. 2016. A comparison of wearable and stationary sensors for stress detection. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 004362–004366.

[58] Cheul Young Park. 2020. *Toolbox for Emotion Analysis using Physiological signals (TEAP) in Python*. https://github.com/cheulyop/PyTEAP.

[59] Martin Pielot, Bruno Cardoso, Kleomenis Katevas, Joan Serrà, Aleksandar Matic, and Nuria Oliver. 2017. Beyond Interruptibility: Predicting Opportune Moments to Engage Mobile Phone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 91 (Sept. 2017), 25 pages. https://doi.org/10.1145/3130956

[60] John P Pollak, Phil Adams, and Geri Gay. 2011. PAM: a photographic affect meter for frequent, in situ measurement of affect. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 725–734.

[61] Ho-Kyeong Ra, Jungmo Ahn, Hee Jung Yoon, Dukyong Yoon, Sang Hyuk Son, and JeongGil Ko. 2017. I am a" smart" watch, smart enough to know the accuracy of my own heart rate sensor. In *Proceedings of the 18th International Workshop on Mobile Computing Systems and Applications*. 49–54.

[62] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[63] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research* 17, 7 (2015), e175.

[64] Dario D Salvucci, Niels A Taatgen, and Jelmer P Borst. 2009. Toward a unified theory of the multitasking continuum: From concurrent performance to task switching, interruption, and resumption. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1819–1828.

[65] Akane Sano and Rosalind W Picard. 2013. Stress recognition using wearable sensors and mobile phones. In *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE, 671–676.

[66] Ulrich Schimmack and Reisenzein Rainer. 2002. Experiencing activation: Energetic arousal and tense arousal are not mixtures of valence and activation. *Emotion* 2, 4 (2002), 412.

[67] Philip Schmidt, Attila Reiss, Robert Dürichen, and Kristof Van Laerhoven. 2018. Labelling Affective States" in the Wild" Practical Guidelines and Lessons Learned. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 654–659.

[68] Maude Schneider, Thomas Vaessen, Esther DA van Duin, Zuzana Kasanova, Wolfgang Viechtbauer, Ulrich Reininghaus, Claudia Vingerhoets, Jan Booij, Ann Swillen, Jacob AS Vorstman, et al. 2020. Affective and psychotic reactivity to daily-life stress in adults with 22q11DS: a study using the experience sampling method. *Journal of Neurodevelopmental Disorders* 12, 1 (2020), 1–11.

[69] Christie Scollon, Chu Kim-Prieto, and Ed Diener. 2003. Experience Sampling: Promises and Pitfalls, Strengths and Weaknesses. *Journal of Happiness Studies* 4, 1 (2003), 5–34.

[70] Saya Shacham. 1983. A shortened version of the Profile of Mood States. *Journal of personality assessment* (1983).

[71] D Siewiorek, A Smailagic, J Furukawa, A Krause, N Moraveji, K Reiger, J Shaffer, and Fei Lung Wong. 2003. SenSay: a context-aware mobile phone. In *Seventh IEEE International Symposium on Wearable Computers, 2003. Proceedings.* IEEE, 248–249.

[72] Rolf Steyer, Peter Schwenkmezger, Peter Notz, and Michael Eid. 1997. Der Mehrdimensionale Befindlichkeitsfragebogen MDBF [Multidimensional mood questionnaire]. *Göttingen, Germany: Hogrefe* (1997).

[73] Arthur A Stone and Saul Shiffman. 1994. Ecological momentary assessment (EMA) in behavorial medicine. *Annals of Behavioral Medicine* (1994).

[74] Yoshihiko Suhara, Yinzhan Xu, and Alex'Sandy' Pentland. 2017. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In *Proceedings of the 26th International Conference on World Wide Web.* ACM, 715–724.

[75] T. and Brian P. Bailey. 2005. Investigating the Effectiveness of Mental Workload as a Predictor of Opportune Moments for Interruption *(CHI EA '05).* ACM, New York, NY, USA, 4. https://doi.org/10.1145/1056808.1056948

[76] Shelley E Taylor, William T Welch, Heejung S Kim, and David K Sherman. 2007. Cultural differences in the impact of social support on psychological and biological stress responses. *Psychological science* 18, 9 (2007), 831–837.

[77] Robert E Thayer. 1990. *The biopsychology of mood and arousal.* Oxford University Press.

[78] Artem Timoshenko and John R Hauser. 2019. Identifying customer needs from user-generated content. *Marketing Science* 38, 1 (2019), 1–20.

[79] Peter Totterdell and Simon Folkard. 1992. In situ repeated measures of affect and cognitive performance facilitated by use of a hand-held computer. *Behavior Research Methods, Instruments, & Computers* 24, 4 (1992), 545–553.

[80] William Trochim and James Donnelly. 2008. Research methods knowledge base. *Donnelly. Mason, OH: Cenage Learning* (2008).

[81] Endel Tulving and Daniel L Schacter. 1990. Priming and human memory systems. *Science* 247, 4940 (1990), 301–306.

[82] Endel Tulving and Donald M Thomson. 1973. Encoding specificity and retrieval processes in episodic memory. *Psychological review* 80, 5 (1973), 352.

[83] Liam D Turner, Stuart M Allen, and Roger M Whitaker. 2015. Interruptibility prediction for ubiquitous systems: conventions and new directions from a growing field. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing.* 801–812.

[84] Niels van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2018. The Experience Sampling Method on Mobile Devices. *ACM Computing Surveys (CSUR)* 50, 6 (2018), 93.

[85] Niels van Berkel, Jorge Goncalves, Lauri Lovén, Denzil Ferreira, Simo Hosio, and Vassilis Kostakos. 2019. Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports. *International Journal of Human-Computer Studies* 125 (2019), 118–128.

[86] Ruud Van Winkel, Cécile Henquet, Araceli Rosa, Sergi Papiol, Lourdes Fañanás, Marc De Hert, Jozef Peuskens, Jim van Os, and Inez Myin-Germeys. 2008. Evidence that the COMTVal158Met polymorphism moderates sensitivity to stress in psychosis: An experience-sampling study. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 147B, 1 (2008), 10–17.

[87] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing.* 3–14.

[88] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063.

[89] Mario Wenzel, Zarah Rowland, and Thomas Kubiak. 2021. Like clouds in a windy sky: Mindfulness training reduces negative affect reactivity in daily life in a randomized controlled trial. *Stress and Health* 37, 2 (2021), 232–242.

[90] Ladd Wheeler and Harry T Reis. 1991. Self-recording of everyday life events: Origins, types, and uses. *Journal of personality* 59, 3 (1991), 339–354.

[91] Raf Widdershoven, Marieke Wichers, Peter Kuppens, Jessica Hartmann, Claudia Menne-Lothmann, Claudia Simons, and Jojanneke Bastiaansen. 2019. Effect of self-monitoring through experience sampling on emotion differentiation in depression. *Journal of Affective Disorders* (2019), 71–77.

[92] Peter Wilhelm and Dominik Schoebi. 2007. Assessing mood in daily life. *European Journal of Psychological Assessment* 23, 4 (2007), 258–267.

[93] Fengpeng Yuan, Xianyi Gao, and Janne Lindqvist. 2017. How busy are you? Predicting the interruptibility intensity of mobile users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* 5346–5360.

[94] Vincent W Zheng, Yu Zheng, Xing Xie, and Qiang Yang. 2010. Collaborative location and activity recommendations with GPS history data. In *Proceedings of the 19th international conference on World wide web.* 1029–1038.

[95] Manuela Züger and Thomas Fritz. 2015. Interruptibility of software developers and its prediction using psycho-physiological sensors. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.* 2981–2990.