Exploring Context-Aware Mental Health Self-Tracking Using Multimodal Smart Speakers in Home Environments

Jieun Lim*
KAIST
Daejeon, South Korea
jieun.lim@kaist.ac.kr

Youngji Koh*
KAIST
Daejeon, South Korea
youngji@kaist.ac.kr

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

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

ABSTRACT

People with mental health issues often stay indoors, reducing their outdoor activities. This situation emphasizes the need for self-tracking technology in homes for mental health research, offering insights into their daily lives and potentially improving care. This study leverages a multimodal smart speaker to design a proactive self-tracking research system that delivers mental health surveys using an experience sampling method (ESM). Our system determines ESM delivery timing by detecting user context transitions and allowing users to answer surveys through voice dialogues or touch interactions. Furthermore, we explored the user experience of a proactive self-tracking system by conducting a four-week field study (n=20). Our results show that context transition-based ESM delivery can increase user compliance. Participants preferred touch interactions to voice commands, and the modality selection varied depending on the user's immediate activity context. We explored the design implications for home-based, context-aware self-tracking with multimodal speakers, focusing on practical applications.

CCS CONCEPTS

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

KEYWORDS

Self-tracking, Mental Health, Multimodal Smart Speakers, Experience Sampling Method (ESM)

ACM Reference Format:
Jieun Lim, Youngji Koh, Auk Kim, and Uichin Lee. 2024. Exploring Context-Aware Mental Health Self-Tracking Using Multimodal Smart Speakers in Home Environments. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3613904.3642846

*Equal contribution.
†Corresponding authors.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642846

1 INTRODUCTION

Mental health has become a global concern in recent years. According to a Global Burden of Disease (GBD) study, an estimated 970 million people, or approximately one-eighth of the world's population, are affected [19]. In addition, mental health disorders have a considerable socioeconomic cost, as they negatively affect productivity. A World Health Organization-led study found that depression and anxiety disorders lead to 1 trillion dollars in lost productivity worldwide annually [10].

Self-tracking involves individuals collecting and reflecting their own data [37] to enhance self-awareness and understanding of their health and well-being [29, 32] For mental health issues such as depression and anxiety disorders, there is a need for periodic monitoring of dynamic changes on a daily or hourly basis rather than being observed at occasional hospital visits [53]. Self-tracking is well suited for identifying such dynamic changes because it can continuously observe the user's state. Self-tracking in daily life is widely used as a research and clinical methodology because it helps bridge the information gap between healthcare stakeholders and patients for clinical decision-making [60].

This study used the experience sampling method (ESM) to self-track mental health states. ESM is a well-known methodology that asks users to complete short surveys at various points in their daily lives to collect data on their feelings, thoughts, and behaviors [13]. ESM can deliver surveys at regular intervals, which can be effective for observing and understanding highly volatile mental states [53], such as stress and anxiety levels while mitigating recall bias [51]. Furthermore, ESM can be structured to collect associated contextual information, which helps gain insights into mental health [1].

ESM has traditionally been based on the paper-and-pen method. With advances in mobile and wearable technology, ESM has been actively used with mobile phones [55, 56] and smartwatches (known as microESM) [17, 25]. The issue with these technologies is that people do not always carry smartphones [14] or wear smartwatches at home [21]. As an alternative approach, smart speakers have recently been explored for ESM in domestic settings [57, 59] because there has been a significant increase in the adoption of speakers at home. The advantage of speakers is that they are often placed in key areas of the home (e.g., the living room and bedside), where they can easily interact with users.

While traditional smart speakers have centered on voice interfaces, the market has recently expanded to include multimodal speakers, such as Amazon's Echo Show and Google's Nest Hub, which support both voice and touch interactions. The number of users is increasing rapidly [35]. These multimodal speakers provide new opportunities to perform various types of mental health self-tracking tasks with visual elements that are not feasible with traditional voice-only speakers. For example, users were tasked with describing a given image to diagnose depression [22, 34] and cognitive impairment [40]. In this sense, multimodal speakers can collect different types of mental health data, making them effective at conducting mental health research and self-reflection. Despite the immense potential of multimodal speakers as mental health self-tracking tools, we observed that HCI studies are still to investigate user experiences of mental health self-tracking with multimodal speakers.

One important aspect of ESM design is to determine the opportune timing for delivering ESM surveys because inappropriate timing can result in negative emotions such as stress [30], irritation [4], and anxiety [5], as well as bias users in their responses to the survey [23]. Previous studies identified task breakpoints and user activity transitions as opportune moments to interrupt (or proactively interact with) users [3, 15, 20]. While previous research has investigated opportune timing for interactions with smart speakers in domestic environments [11], to the best of our knowledge, no research has explored user interruptibility in response to ESM based on context transitions detected by various Internet of Things (IoT) sensors (such as noise and light) in a field study setting.

Therefore, we developed a context-aware self-tracking system using multimodal speakers that proactively deliver ESM requests (e.g., mental health questionnaires) at home by detecting context transitions based on IoT sensing at home (e.g., CO2, light, noise, and motion). Users can then respond to ESM requests through voice or touch. We set the following research questions (RQs):

- (RQ1) How do users evaluate their overall experience of proactive mental health self-tracking using multimodal speakers?
- (RQ2) How do users' compliance rates change across different context transitions, and what are their perceptions of context-aware ESM triggering in home environments?
- (RQ3) What are the preferred interaction modalities for responding to ESM requests and what is the usability of these modalities?

To answer these questions, we conducted a four-week field study with 20 participants who had mild depressive symptoms in their homes. Our results show that our context-aware self-tracking system facilitated participants' responses to ESM requests, enhancing both convenience and compliance rates. Participants perceived human likeness in voice-based interactions; however, the repetitive nature of the interaction content lowered their expectations. Participants generally preferred the touch modality over the voice modality, and the modality selection varied depending on the user's activity context before interactions with the system. Based on the results, we explored the design of home-based, context-aware self-tracking systems with multimodal speakers. The key contributions of this study are as follows:

- We propose a self-tracking system that detects contextual transitions in a home environment based on IoT sensor data.
- We conducted a field study (n=20, four weeks) in a home setting to explore system usage patterns and user experiences.
- Finally, we discuss several practical implications of designing a proactive mental health self-tracking agent for home environments.

2 BACKGROUND AND RELATED WORK

2.1 Mental Health Self-Tracking with Experience Sampling

Existing self-tracking technologies have observed mental health states such as emotions, depression, anxiety, and stress [12, 31, 55], using two methods: self-reporting and automatic sensing [28]. Self-reporting is a user-driven method used for recording such states through one-time or periodic surveys. Automatic sensing is a method used for recording such data by collecting sensor data on behavioral indicators (e.g., sleep duration, movement, and activity) and physiological indicators (e.g., heart rate and skin conductance).

The ESM is commonly considered a self-reporting method. ESM is a naturalistic method for collecting user data in various daily life situations by requesting self-reported user states at regular intervals [13]. Collecting self-reported data through ESM offers several advantages. First, ESM allowed us to observe changes in mental health states. Such states are subject to fluctuations due to various factors such as the time of day and external factors [53]. Therefore, we must monitor states periodically on a daily or hourly basis rather than through one-time observations [53]. Consequently, when ESM is used for self-tracking, users can observe changes in their mental health, thereby enhancing their understanding of it. Recall bias poses a significant issue in self-reporting approaches because users may not accurately remember past events or emotions. Second, ESM minimizes recall bias by capturing data close to the actual mental health status [51], as it prompts users to complete surveys immediately when requested. Finally, ESM allows users to record contextual information, helping them gain insights into their mental health status and enhance their understanding of themselves [1, 53].

2.2 Interaction Techniques for Experience Sampling

ESM is commonly performed on mobile and wearable devices (e.g., smartphones [52, 56] and smartwatches [7, 25]). However, such devices are limited to performing ESM in the home environment because users may not always carry such devices on their arm's reach [14]. In addition, some groups of users tend to wear wearables less often at home than in other locations such as the workplace [21]. Therefore, mobile and wearable devices are less suitable for self-tracking in home environments. Alternative, smart-speaker-based ESM is becoming a popular tool for mental health tracking in home settings.

Recently, smart speakers have been widely used in home settings [6]. They are commonly placed at convenient locations in the home (e.g., living room or bedside) to interact with their users. Therefore, they have numerous opportunities to interact naturally

with their users when they enter the radius of the smart speaker during their daily activities. They can also increase user engagement by providing a natural conversational experience similar to interacting with a person [58]. Based on these advantages, several studies have explored ESM using smart speakers in a home setting [57, 59]. However, these studies only explored voice-based ESM through speakers to understand the overall user perception of proactive smart speakers [57] or to analyze errors with voice-based interactions [59]. In this study, we extend the existing studies on voice-based ESM using smart speakers for multimodal interaction-based ESM. We explored (RQ1) the overall user experience and usability of mental health self-tracking with multimodal speakers.

2.3 Opportune Interaction Timing for Proactive Services

The opportune timing of delivery is considered important in proactive services because interrupting users at inappropriate times can affect their ability to perform tasks [5, 62] or their emotional state (e.g., irritability and anxiety) [2, 5, 62]. As ESM also interrupts users several times a day to ask them to complete mental health surveys, we must consider the opportune timing of ESM requests. If users are requested to respond, they may not provide inaccurate responses, or drop out of the user study [42]. This can significantly impact the quality and quantity of self-tracking data. Previous research identified context transitions as opportune moments that interrupt users [3]. For example, in desktop environments, task breakpoints [2, 20], where a user transits from one task to another, have been strategically used as such opportune moments. In mobile environments, activity transitions when a user transits from one activity to another (e.g., standing, sitting, and walking) are widely detected and leveraged as such opportune moments [15, 44]. For instance, the transition from walking to sitting was determined to be an appropriate moment for sending a smartphone notification [18, 44]. Similarly, smart speakers can detect activity transition contexts in a home setting based on IoT sensors and make ESM requests to increase interaction opportunities. Thus, we explored (RQ2) how user response (or compliance) rates vary across different activity transition contexts and how users perceive each trigger type.

2.4 Design Opportunities for Experience Sampling with Multimodal Smart Speakers

While the smart speaker market has traditionally centered on the voice user interface (VUI), it has recently expanded to include multimodal speakers that offer a graphical user interface (GUI) alongside the VUI [35]. These multimodal speakers open new opportunities for performing a variety of mental health self-tracking tasks with a visual component that cannot be performed with traditional voice-based speakers. For example, depression can be detected by using speech features (e.g., pitch changes and speech rate) of users when the users describe pictures on a screen [22, 27, 34]. Cognitive impairment can also be diagnosed based on the semantic content or syntactic complexity of users' picture description [40]. Despite these advantages, there has been a lack of consideration for developing a context-aware self-tracking system using multimodal smart speakers in a home environment. Therefore, this study explores (RQ3) the essential factors for the interaction design of a multimodal speaker-based ESM system by answering the following questions: (1) Which interaction modalities do users prefer to use, (2) in what context do users use each interface, and (3) does the modality preference change over time?

3 SYSTEM DESIGN

As shown in Figure 2, our system comprises three hardware components: an IoT sensor, a smartphone, and a multimodal smart speaker. The phone was equipped with a wide-angle camera to capture images of the entire living space. Earphones connected to the phone were attached to the microphone section of the speaker. Figure 1 shows the two main parts of our system: ① context-aware ESM scheduling, and ② a multimodal ESM survey. User context transitions have been widely utilized as opportune moments to interact with users [2, 15, 18, 20, 43, 44]. ① The context-aware ESM scheduling component determines opportune moments for ESM requests by detecting user context transitions using IoT sensors. ② multimodal ESM survey allows users to respond to ESM requests (or survey questions) through voice and touch interactions via a multimodal smart speaker. In this section, these aspects are reviewed in detail.

3.1 Context-Aware ESM Scheduling Using Sensors

Detecting Context Transitions: We determined the opportune moments for ESM requests by detecting user context transitions using sensors. Specifically, we detected changes in one of the two user contexts: auditory channel availability and proximity to smart speakers. These contexts were previously identified as important contexts relevant to opportune moments for smart speakers to proactively interact with users at home [11]. For the sensors, we considered CO_2, camera, light, and noise. A noise sensor was used to detect changes in the auditory/verbal channel availability. Light, camera, and CO_2 sensors were used to detect changes in human presence (e.g., movement) near the speaker. Environmental conditions (e.g., average noise levels) may vary between homes. Therefore, four sensor-based trigger conditions were set at different thresholds. To ensure that these thresholds accurately reflected real-life conditions in users' homes during the installation, we individually calibrated the thresholds for each home as follows:

- *Noise sensor:* We set the trigger condition when the noise level changed from high to low (or noisy to quiet). For example, when a user stops watching a video, the environment becomes quiet and ESM prompts are triggered. To set the thresholds of the high and low levels, we played a video for 5 seconds in locations (e.g., on a bed or at a desk) where users typically watched videos and measured the sound level in decibels before and after the play.
- *Light sensor:* We set the trigger condition as when the light level shifts from low to high (or dark to bright) to detect the presence of a person, because a user may turn on the lights when they enter the room where the speaker is installed. To set the high- and low-level thresholds, we set the light levels

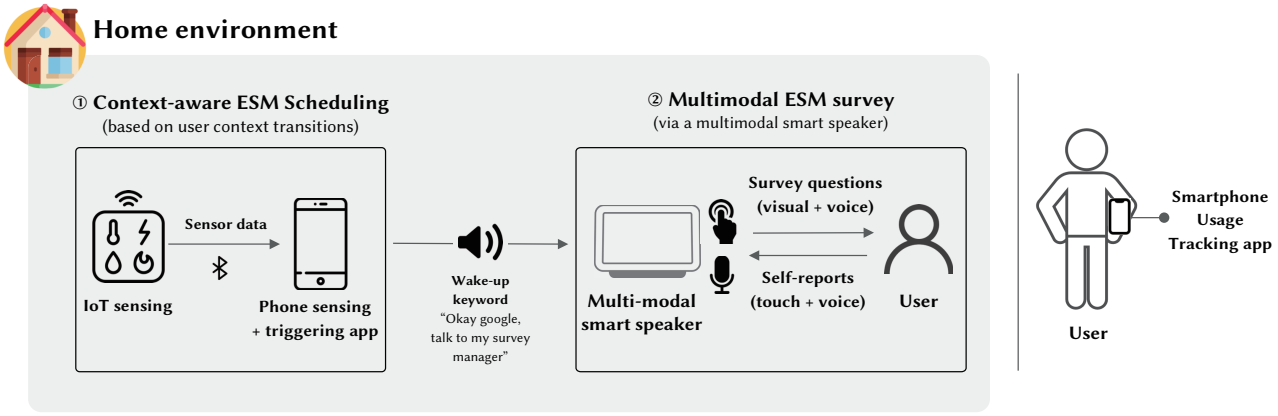


Figure 1: System overview of context-aware mental health self-tracking using multimodal smart speakers

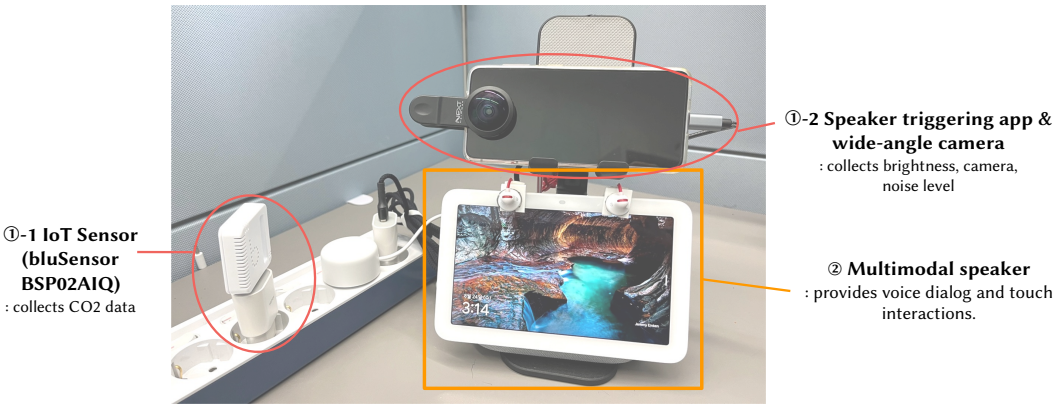


Figure 2: Overall hardware configuration of the system

before and after turning off the lights in the room in which the speaker was located.

- *CO_2 sensor:* We set the trigger condition as when the CO_2 concentration level shifts from a low to a high level, as the CO_2 level in the room increases rapidly owing to breathing and activity when people are in the room. To set the high- and low-level thresholds, we measured the CO_2 level before and after a human spoke for 30 s within a 1-meter radius of the speaker.
- *Camera sensor:* We set the trigger condition to when one or more people were detected by the camera. To monitor the number of people in a room, we continuously processed the images from the camera using the OpenCV library.

Triggering Algorithm and Daily Operating Hours: As shown in Figure 4, after a minimum request interval had elapsed since the last request, ESM surveys were requested (1: sensor-based trigger condition) when opportune moments were determined (i.e., any of the four sensor-based trigger conditions were met) or (2: time-out trigger condition) when the maximum request interval had timed out from the last request. We set the minimum and maximum intervals to 30 min and 90 min, respectively. This configuration enabled ESM to be requested every 60 min on average. Our triggering algorithm was operated only during operation hours. For daily operating hours, users set 10 hours of regular waking hours (i.e., 8 am–6 pm). They can set the operating hours differently on weekdays and weekends. When ESM surveys were requested continuously at the maximum request interval (i.e., 90 min) during a 10-hour operational period, six ESM surveys were requested per day. If a minimum of one context transition was detected within a day, the system requested at least six ESM surveys per day.

3.2 ESM Survey with Multimodal Smart Speakers

ESM surveys were requested via commercial multimodal speakers (2nd Generation Google Nest Hub). As speakers provide the survey in both voice and visual modalities, users can choose one of them to answer. As shown in Figure 5, our ESM task consists of four steps: (1) start and greeting, (2) previous activity inquiry, (3) mental health survey inquiry, and (4) picture card description. Figure 6 shows the detailed user interfaces for each step of the conversation. In user

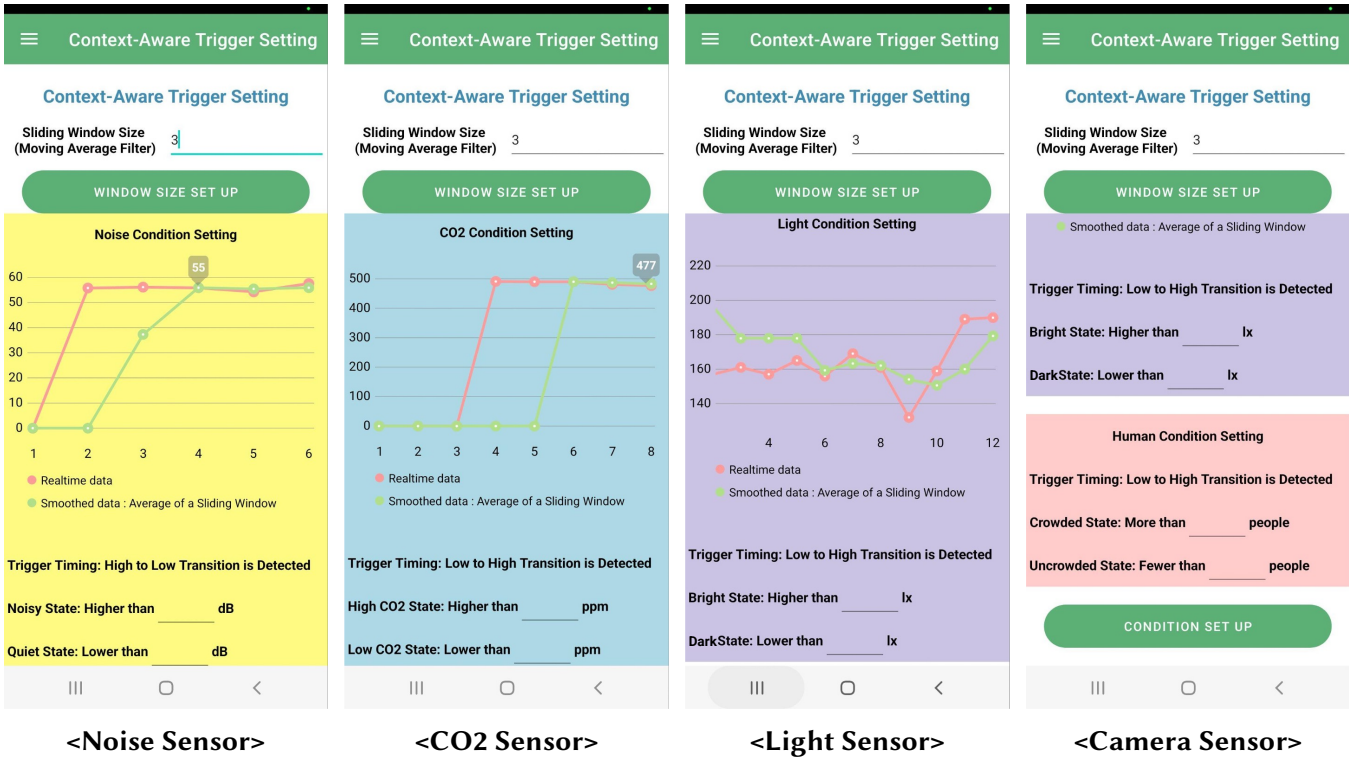


Figure 3: Mobile user interfaces for setting thresholds of four sensor-based trigger conditions.

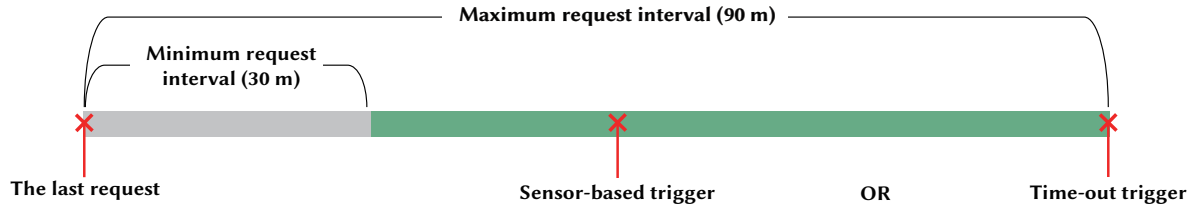


Figure 4: Triggering algorithm for ESM survey requests. ESM surveys are requested when any of the sensor-based trigger conditions are satisfied within 60 minutes (green bar) after a minimum request interval (30 mins; gray line) has passed since the last request. Otherwise, it requests when a maximum request interval (90 min) elapses since the last request.

interfaces, visually displayed text is constantly accompanied by a robot avatar to facilitate human–likeness in intelligent agents [48]. For designing voice interactions, we referred to existing design guidelines [47, 59] to address well-known interaction errors. Furthermore, a series of pilot and field tests were conducted to improve the multimodal interactions of the ESM. In the following section, we present the key interaction elements for each step.

Start and Greeting: For the beginning of the conversation, the speaker greets the user by saying, "Hello, if you want to start the survey, press the start button on the screen or say 'start.'" Once the user presses the on-screen 'start button' (or say 'start'), the main part of the survey begins. By default, the Google smart speaker stops listening to the user after 8 seconds. To address the no-speech timeout issue, we implemented a method of starting the survey by tapping the start button on the screen.

Activity inquiry: Next, to capture the user's context, the speaker asks about the user's activity before the conversation with the speaker by saying, "Please describe the activity you were engaged in just before the survey." Users can answer the question verbally.

Mental health survey: Next, the speaker asked the user to self-report their health states on four mental health scales, as shown in Table 1: depression, anxiety disorders, stress, and mood states. These scales were considered because depression and anxiety disorders are the most common mental health illnesses [19], mood instability is a common symptom associated with mental illnesses [46], and stress can affect physical health or be a major contributor to mental health illnesses [16, 38]. The survey items (or questions) were presented

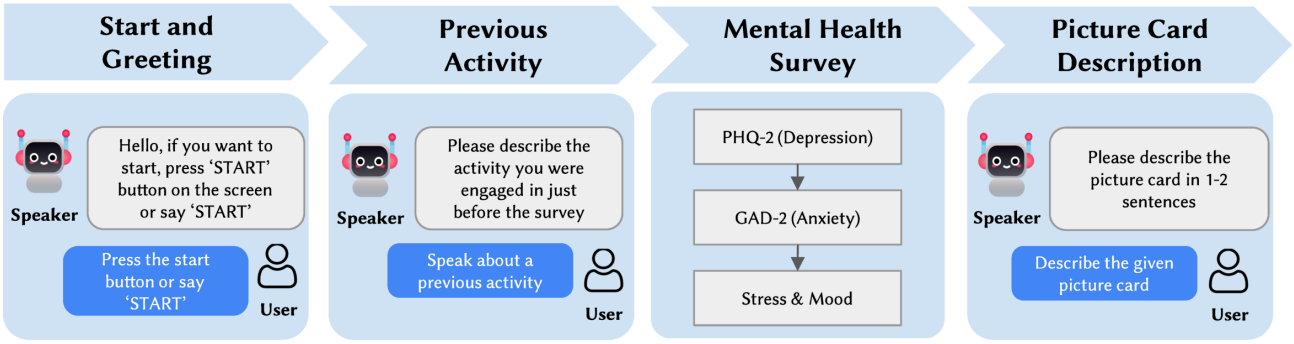
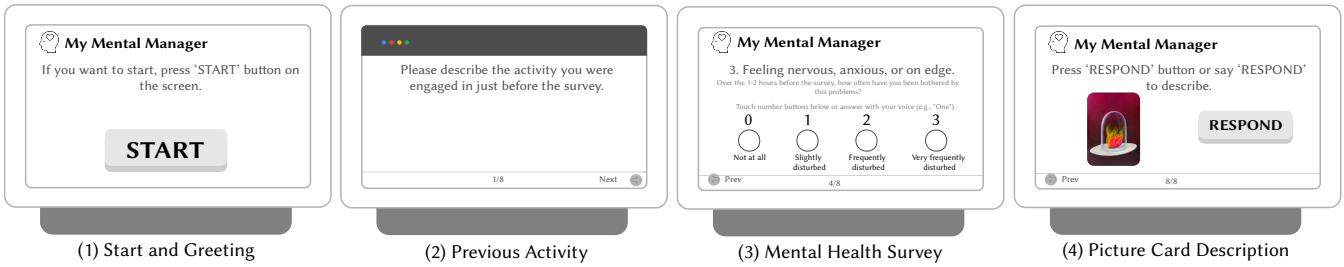


Figure 5: ESM conversation of four ESM task steps (English translated version)



(1) Start and Greeting (2) Previous Activity (3) Mental Health Survey (4) Picture Card Description

Figure 6: Speaker user interfaces of four ESM task steps (English translated version)

to the user in audio and text formats (for example, "In the last 1–2 hours, how nervous, anxious, or impatient have you felt?" [0-3 points]). Each question can be answered either verbally (e.g., speaking) or manually (e.g., touching).

Picture description: In these studies, in addition to mental health surveys, we considered picture description tasks. For the diagnosis of mental health diseases, picture description tasks have been widely utilized to collect and analyze the language and vocal characteristics (e.g., pitch changes and speech rate) of individuals during the tasks [22, 27, 33, 34]. For the description tasks, we used images of cards from the Dixit card game. These images depict various scenarios. In the game, players create sentences based on images. Similarly, in our study, the users were asked to verbally describe an image displayed on a speaker screen. The Dixit game offers an extensive collection of cards, allowing our system to present different images of ESM requests throughout the day. This variety was a key factor in choosing images from the Dixit game for the description tasks.

In this step, the speaker says, "Please describe the picture card in 1-2 sentences." Simultaneously, the screen shows the text, "Press the 'RESPOND button' or say 'RESPOND' to describe." As the users needed time to observe the picture, we implemented a method for them to press the response button on the right side of the screen to start the voice description when they were ready. As mentioned earlier, after 8 s of user response, the speaker stops listening and turns off the microphone (i.e., no-speech timeout). Consequently, if they take a short pause or hesitate, they may not be able to complete their verbal responses. Accordingly, the system offered two response opportunities.

After the first response, the speaker said, "If you could not finish your answer, please continue." If additional answers were required, the user could continue to describe the picture. If they have already completed their explanation, they can end their survey by saying "No." Finally, the speaker thanks the user and exits the program.

All user responses are stored as text in Google Cloud Storage. Additionally, if a participant responded for the second time, the second response was concatenated with the initial response. For example, if a user's first response is "A girl is staring at something" and the second response is "And she's holding a candle," the final stored response would be stored as "A girl is staring at something + And she's holding a candle."

4 FIELD STUDY METHODS

We conducted a field study that was approved by the ethical review board of our institute. This section describes the participants, study procedures, a data overview, and activity categorization.

4.1 Participants

We recruited 20 participants (10 females and 10 males; average age, 27 years) by posting online and offline announcement posters. Our recruitment criteria included people (1) who were diagnosed with at least mild depression (a score of 5 or higher), (2) who had

Related mental illnesses	Survey questions	References
Depression	Over the 1-2 hours before the survey, how often have you been bothered by the following problems? 1. Little interest or pleasure in doing things 0) Not at all 1) Slightly 2) Frequently 3) Very frequently 2. Feeling down, depressed or hopeless 0) Not at all 1) Slightly 2) Frequently 3) Very frequently	PHQ-2 [49]
Anxiety disorders	Over the 1-2 hours before the survey, how often have you been bothered by the following problems? 3. Feeling nervous, anxious, or on edge 0) Not at all 1) Slightly disturbed 2) Frequently disturbed 3) Very frequently disturbed 4. Not being able to stop or control worrying 0) Not at all 1) Slightly disturbed 2) Frequently disturbed 3) Very frequently disturbed	GAD-2 [50]
Stress	5. Over the 1-2 hours before the survey, what was your stress level? 1) Very Low 2) Low 3) Medium 4) High 5) Very High	[24]
Mood	6. Over the 1-2 hours before the survey, how was your feeling? 1) Very bad 2) Bad 3) Neither bad nor good 4) Good 5) Very Good	[24]

Table 1: Mental health questionnaires for measuring depression, anxiety, stress, and mood.

private spaces at home or were single-person households, and (3) who spent a minimum of 5 hours daily in their room, excluding sleep time. To screen for depression severity, we used the PHQ-9 questionnaire [49]. The mean PHQ-9 score was 9.45 (SD, 4.6). The distribution of participants according to their depression severity levels [45] was as follows: No depression (1–4 points), n=0; mild depression (5–9 points), n=12; moderate depression (10–19 points), n=7; and severe depression (20–27 points), n=1.

Criterion 2 was considered because, when using a camera for person detection, it was challenging to distinguish whether the detected person was a participant or not. Among the participants, 17 lived alone, and the remaining three lived with others. Fourteen participants resided in studio apartments or dormitory rooms, four were in two-bedroom residences, and the remaining two were living in houses. Finally, Criterion 3 was considered to ensure that the participants responded to the survey a minimum of five times a day. Participants who completed the entire field study process and achieved the required survey response rate were compensated with 400,000 KRW (approximately 300 USD).

4.2 Procedures

We conducted a four-week field study. Before this study, we asked the participants to watch an instructional video of the study (e.g., a general overview, a list of data collected, and the objectives of the study). For the mental health survey, the participants were instructed to consider their mental health status in the last 1–2 hours. For picture card descriptions, the participants were asked to respond to 1–2 complete sentences. Next, we visited each participant's house to set up an ESM device. The devices were installed in locations that offered a view of the entire personal space, such as desks, in accordance with the participants' preferences. We also installed a Wi-Fi tracking application on a smartphone. The app periodically collects Wi-Fi signals (e.g., network name and signal strength) around the phones. We later used the collected Wi-Fi signals to detect the user's home presence. Four weeks post-installation, participants were instructed to complete ESM surveys five times daily. Weekly reminders were sent via instant messages to those with lower participation rates to encourage consistent engagement.

After the four-week field study, we collected the devices and conducted post-surveys and interviews. The interviews lasted for approximately one hour. In the post-surveys, we asked about system usability using the usefulness, satisfaction, and ease of use (USE) questionnaire [36]. During the interviews, we asked about the overall user experience. We recorded and transcribed the interviews and conducted a thematic analysis [8]. We initially coded user responses without predefined criteria. After the initial coding, we iteratively reviewed and refined the code, merged similar code, and named the themes. This process was repeated until a consensus was reached.

4.3 Types of Collected Data

Three types of data were collected, as listed in Table 2. First, we collected user-speaker conversation data using multimodal speakers. These data contained the trigger type (i.e., sensor-based or timeout trigger condition), responses for activity inquiry, mental health survey, picture card description task, and the time spent for each response. Next, we collected IoT sensor values, such as the number of people near speakers, levels of noise, brightness, and CO2, at one-second intervals. Finally, we collected the Wi-Fi signal data from our Wi-Fi tracking application. Even when a participant was not at home, our speaker requested an ESM survey and recorded the user-speaker conversation data. Thus, Wi-Fi signal data were collected to filter data corresponding to the time when the user

Figure 7: Photos of installed smart speakers: The system was installed in places such as desks and bedsides, where the wide–angle cameras of the mobile phone can scan the entire room and users feel comfortable interacting.

was at home. We first extracted home Wi-Fi networks from the Wi-Fi signal data by selecting the three most frequently detected Wi-Fi networks during early morning hours (12 AM to 6 AM). We presumed these to represent the users' home Wi-Fi networks, as people are mostly at home early in the morning. Based on the selected Wi-Fi networks, we extracted the time of day when each user was at home and selected only the data collected during that time period for further analysis. Through this process, 3,863 ESM cases were selected from the original dataset. Among these, 2,201 ESM surveys were completed, while 1,662 remained unanswered.

4.4 Activity Categorization

Based on user responses in the activity inquiry step, we categorized activities before user engagement in the mental health survey, using affinity diagrams. We first created a taxonomy for activity categorization by performing affinity diagramming and identified 12 activity categories. These categories include using media, resting, doing chores, sleeping, and social interactions, as shown in Table 3. Next, the two researchers individually reviewed and classified 2,201 responses into categories. Cohen's kappa value [39] (k = .97) showed a high agreement between their classification results. Disagreements were resolved through discussions.

4.5 Statistical Analysis

We conducted a series of multilevel logistic regression analyses, along with the final statistical analysis. For the multilevel logistic regression analysis, while the dependent variable and fixed effects were varied for each analysis; their details can be found in the corresponding sections. To account for the non-independence of the data, we included participants as a random effect. As a summary statistic to quantify the goodness-of-fit, we presented marginal R^2 and conditional R^2 [41]. The marginal R^2 shows variance explained by fixed factors, whereas the conditional R^2 shows variance explained by both fixed and random effects. For the final statistical analysis, we used repeated measures analysis of variance (ANOVA) for our statistical analysis, detailing the variables in the corresponding section.

5 RESULTS ON USER EXPERIENCES OF MENTAL HEALTH SELF-TRACKING USING MULTIMODAL SMART SPEAKERS

In this section, we first present (1) the overall user experiences of mental health self-tracking using multimodal speakers and (2) system usability using the USE questionnaire [36]. We then report on (3) how users interact with multimodal speakers when performing picture description tasks.

5.1 Overall User Experiences

In our interviews, users reported that the proactive and periodic nature of the system helped them to gauge their mental health status. First, these features allow them to reflect on their mental health. P6 mentioned, "*I think it was good to be able to reflect on my mental health. It was nice to realize that my mental health was not good during the experiment.*" P18 also noted, "*You don't usually get a chance to ask yourself these questions (related to mental health). But every hour or two, the system asks you how you're feeling or how stressed you are, and it gives me more opportunities to think about whether you've just gotten stressed.*" These findings suggest that our proactive self-tracking service can increase self-awareness and self-reflection regarding mental health. Some users also mentioned the effectiveness of proactive self-tracking. P11 said, "*I liked that the speaker spoke to me first and that I had time to reflect each time. I also thought it would be useful if it became commercialized later.*" P12 mentioned that "*I think it's much better to ask first. If I should start the service by myself, I might not use (the system) endlessly.*"

Some users reported that the system gave them opportunities to reflect on the circumstances or reasons for their mental health conditions. P4 said, "*It always asked me if I was stressed or depressed, giving me a chance to think about reasons related.*" P8 also noted, "*It was helpful since you could think about how and why you felt that*

Categories	Data Types	Data Fields
User-speaker survey conversation data	Mental health survey request/response data	Types of ESM survey triggers
		Users' previous activity before surveys
		Responses to mental health questionnaires and picture card description tasks
		Time and timestamp for responses to each item
		Response methods for each question item (voice or touch)
		User id
Smart home environment data	Camera	Number of people near speakers
	Environmental information	Noise level (dB)
		Brightness (lx)
		CO2 (ppm)
Smartphone usage data	Network and device data	Scanned Wi-Fi information (name, signal strength, etc.)

Table 2: List of collected data types during the field study.

No.	Main categories	Activities	Ratio (%)
1	**Using media**	Reading comics/webtoons, watching videos, listening to music, playing games and using the internet	34.0
2	**Studying/Working**	Studying, doing assignments, presentation preparation and working	16.8
3	**Resting**	Relaxing and doing hobbies (knitting or playing a musical instrument)	10.7
4	**Eating**	Having meals, snacks, drinks and alcohol	9.6
5	**Sleeping**	Getting ready for bed and sleeping	3.6
6	**Doing chores**	Cleaning, washing dishes, taking out the trash, cooking, doing laundry and meal preparation	5.2
7	**Leaving/Returning**	Getting ready to go out, returning to home and entering a room	4.3
8	**Social interactions**	Having conversations, calls or messaging and using social media	4.1
9	**Personal hygiene**	Doing a shower, washing your face or hands, brushing teeth, getting ready for a shower, combing hair, clipping nails	3.4
10	**Miscellaneous**	Unknown, etc.	3.6
11	**Self-caring**	Makeup, skincare, cutting nails and hair drying	2.4
12	**Working out**	Stretching and doing exercise	2.3

Table 3: Previous Activity Category. 12 previous activity categories were extracted from the affinity diagramming process.

emotion while answering." Some users found the survey process to be emotionally beneficial as it helped them release their emotions and refresh their mood. P11 mentioned that "*I have a very stressful life, and it was very nice to have a system that helps me to reflect on myself like this every time.*" P19 commented, "*Before, I had no idea about my moods. But when I got a chance to think about it (through the survey), I was like, 'I see ... what was happening' and could relieve negative emotions.*"

Users generally agreed that the multimodal interactions were engaging. As discussed later, they predominantly used touch screens to answer ESM. However, the speaker verbally asking questions along with the screen display made the speaker feel like a person, allowing for a sense of connection and increasing immersion when responding to a series of survey questions. P6 mentioned, "*The speaker said hello to me and asked me about my mental health periodically. I felt like someone cared about my condition. So I couldn't ignore the speakers like other system notifications.*" and P12 said, "*I was more focused on the question because the speaker asked questions verbally. Also, there's only one question on the screen. It makes me concentrate on each question.*" Furthermore, a speaker's avatar and random ESM timing were considered additional human-like factors. P13 said, "*I felt like it's a person because the timing was not exactly regular. It's usually unpredictable when someone will contact you. So, the timing of the speaker talking to me made me feel like a person.*"

Users negatively evaluated machine-like interaction styles, such as repeating the same surveys without any tone or content variations or a lack of feedback. The users became bored by repeating the same questionnaire. P13 mentioned, "*It's annoying that there are a lot of repeated questions.*" P15 said, "*The questions and pictures are repeated over and over again. As the experiment progressed, I felt bored because the system became more habitual and predictable.*" Besides, P7 mentioned, "*I think it was annoying to keep asking the same questions over and over again. So, there was a decrease in the*

sincerity of responses." This suggests that boredom can affect the quality and quantity of the data collected through self-tracking.

Our users were generally less concerned with collecting the sensor and interaction data associated with self-tracking. Participants only expressed privacy concerns regarding camera data collection, despite assurances that the system only counted people and did not store images. P3 noted, "*The camera is capturing the whole room, so even though it doesn't save photos, it's kind of creepy.*" Similarly, P5 said, "*Other than the data collected through the camera, I wasn't too worried that much. I know that no photos or videos are recorded. But just the fact that the camera was installed made me feel a little creepy.*" P4, P12, and P18–P20 mentioned that they coped with such privacy concerns by moving to a different area or covering their camera during sensitive moments.

5.2 Usability of Context-Aware ESM with Multimodal Smart Speakers

To evaluate the usability, users rated our systems from three perspectives: Ease of Use, Ease of Learning, and Satisfaction (USE questionnaires; 1 = completely disagree to 7 = completely agree). Overall, users positively assessed the system (M = 5.05, SD = 0.48). The ease of learning (M = 6.44, SD = 0.09) was significantly higher than the other factors. Most users rated the system as easy to learn and remembered how to use it. In the interviews, most users found the voice prompts and on-screen content organization to be intuitive and simple, making it easy to learn how to use them. P2 noted that "*Voice prompts and on-screen text are intuitive.*" P6 commented, "*The way of answering by touching the screen or voice is very simple and repetitive, so it is easy to learn (how to use) after a day or two days of use.*"

Compared to ease of learning, satisfaction (M = 4.07, SD = 0.53) and ease of use (M = 4.64, SD = 0.9) were lower. Users reported that system errors such as touch malfunctions and voice recognition issues hindered their ease of use. P12 said, "*Sometimes I saw the speaker misrecognize what I was saying through the screen. So I had to go back to the previous screen and answer again. So I don't think it's easy to use.*" They also reported that system errors reduced satisfaction. P20 noted that "*There were times when the screen didn't come up, or (the survey program) restarted again after answering a question.*" There were also cases where the survey program did not run because of an error in Google's voice recognition; that is, Google Assistant misunderstood our wake-up command and executed the wrong commands. P6 noted, "*Sometimes, the survey program didn't come out (from the speaker) right away, but other unrelated contents like novels came out.*" P10 said, "*Sometimes, Google search results came out (when the ESM survey was triggered).*" Such errors are attributed to Google Assistant misunderstanding the wake-up commands.

5.3 Picture Card Description Task

In the picture description task, participants were given two opportunities to describe a picture card. We explored whether these opportunities were sufficient to capture the full responses of the participants. Of the 2,201 responses, 899 were answered twice. Out of these 899 responses, 296 were cases where they simply indicated completion with phrases like "I'm done" or "Completed." The remaining 603 responses either added more information that they had not mentioned initially or corrected their first responses. In this regard, P11 commented, *"I was able to tell the speaker more information about the picture because it asked me twice. I used this function often."* P1 said, *"Having an opportunity to repeat or clarify my statement was helpful when the speaker sometimes misrecognized what I said."*

6 RESULTS ON RESPONSE RATES AND USER PERCEPTION OF CONTEXT-AWARE ESM REQUESTS

In this section, we first analyze the response rates (or compliance rates) for context-aware ESM requests, followed by user perceptions of the contexts in which ESM prompts were requested.

6.1 ESM Response Rates

Overview: During the entire field study period, ESM surveys were requested 193 times per user (SD = 57), with an average of 110 responses (SD = 30). Figure 8 shows the number of ESM surveys to which each user responded or did not respond.

Response rates across trigger conditions: Next, we analyzed the ESM response rates across trigger conditions. The response rates were calculated by dividing the total number of responses per condition by the total number of requests. As shown in Table 4, all conditions except for the noise sensor had over 100 responses. In addition, all sensor-based conditions had higher response rates than the time-out trigger condition (i.e., the maximum request interval elapsed).

Response rates across time periods of the day: Next, we analyzed the response rates across different periods of the day, separated by 6-hour intervals: dawn (2:00 AM to 7:59 AM), morning (8:00 AM to 1:59 PM), afternoon (2:00 PM to 7:59 PM), and night (8:00 PM to 1:59 AM). As shown in Table 5, consistent with the findings of a previous study [11], the response rates were lower in the morning (52.3%). However, response rates were higher in the afternoon and at night (55.3% and 62.4%, respectively).

Statistical comparisons: Finally, as shown in Table 6, we statistically compared whether (1) the sensor-based trigger conditions and (2) the time periods of the day significantly affected the users' responses to the surveys. We included an indicator of whether or not to respond to ESM surveys as the dependent variable, participants as a random effect, and trigger conditions and time periods as fixed effects. The conditional R^2 and marginal R^2 were 0.203 and 0.046, respectively (for details of these metrics, see Section 4.5). Our results showed that our participants were significantly more likely to respond to ESM surveys in the afternoon (β=0.24, OR=1.27, $p = .01$) and at night (β=0.43, OR=1.54, $p < .001$) compared to the morning. In addition, they were significantly more likely to respond when prompted by CO_2 (β=0.60, OR=1.80, $p < .001$), human presence (the number of people) (β=0.92, OR=2.52, $p < .001$), and light (β=1.24, OR=3.44, $p < .001$) sensor-based conditions than the time-out conditions. However, the noise sensor-based conditions were not statistically significant. This could be due to the lower number of trigger attempts (n=21) compared to the other trigger conditions.

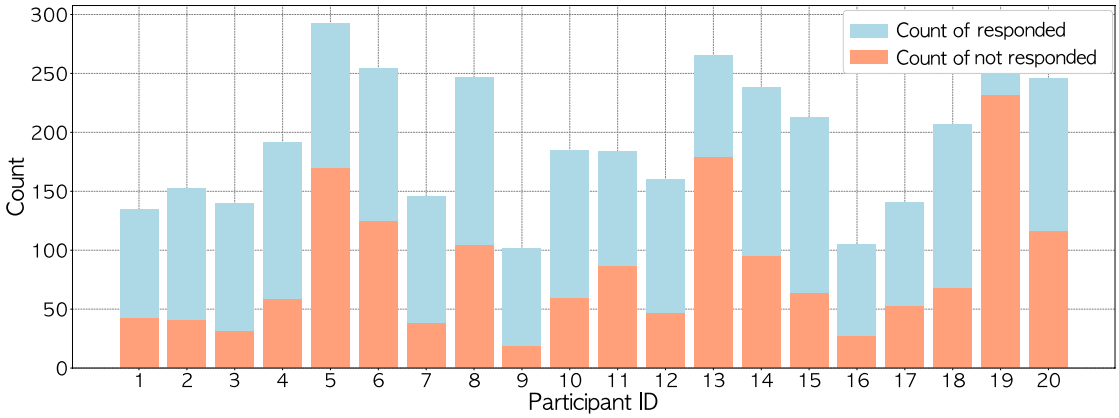


Figure 8: Number of responded and non-responded surveys. Over the course of the entire field study, an average of approximately 193 (SD: 57) ESM surveys were requested, and users responded an average of 110 (SD: 30) times.

Trigger type	Num. responses	Num. requests	Response rate
Maximum time interval	1,502	2,815	53.4%
CO_2	164	272	60.3%
Human	364	549	66.3%
Light	157	206	76.2%
Noise	14	21	66.7%
Total	2,201	3,863	57.0%

Table 4: Number of ESM responses, requests, and response rates by survey trigger type. On average, users showed 57.0% of response rate. A total of 3,863 ESM responses were requested, and 20 users responded to the survey a total of 2,201 times.

Time of day	Num. responses	Num. requests	Response rate
Dawn (2:00~7:59)	35	64	54.7%
Morning (8:00~13:59)	549	1049	52.3%
Afternoon (14:00~19:59)	767	1388	55.3%
Night (20:00~01:59)	850	1362	62.4%
Total	2,201	3,863	57.0%

Table 5: Number of ESM responses, requests, and response rates by time of day.

6.2 User Perception of Trigger Conditions

To understand how users perceived sensor-based trigger conditions, in our post-interviews, we asked in what context they thought the ESM surveys were requested. Users generally recognized light-based (eight users: P1, P5, P8, P9, P12, P15, P16, and P19) or camera-based trigger conditions (five users: P4, P6, P16, P17, and P18). However, they are less aware of other types of sensor-based trigger conditions. The lower response to noise sensor prompts (only 21 instances) compared to CO2, light, and camera sensors could be due to noise triggers being less frequent and CO2 changes being less perceptible to individuals than variations in light or presence detected by cameras.

Next, to understand the user experience (compliance), we analyzed the percentage of activities that users performed across sensor-based trigger conditions. As shown in Figure 9, overall media usage and studying/working are the most frequent activities. For light and camera (human) conditions, doing chores, eating food, and maintaining personal hygiene are the most common activities. In the noise condition, social interaction appeared distinctly when compared with that in the other conditions. For the CO_2 condition, studying/working activities and eating food were more common than for the other conditions.

In the interviews, we also asked our participants about specific scenarios in which they did not respond to ESM surveys. Our participants reported that they were unwilling or unable to respond in the following situations: sleeping (five users: P1, P9, P11, P13, P19, and P20), social interactions (four users: P2, P4, P7, and P10),

Predictors	B (SE)	z-statistic	*95% CI for odds ratio*			p-value
			Lower	*Odds ratio*	*Upper*	
(Intercept)	-0.04 (0.20)	-0.20	0.66	0.96	1.41	0.84
Time of day						
Dawn (2:00–7:59)	0.50 (0.31)	-1.61	0.90	1.64	3.00	0.11
Afternoon (14:00–19:59)	0.24 (0.09)	2.58	1.06	1.27	1.53	**0.01**
Night (20:00–1:59)	0.43 (0.10)	4.54	1.28	1.54	1.86	**<0.001**
Trigger type						
CO_2	0.60(0.15)	3.95	1.35	1.80	2.42	**<0.001**
Human	0.92 (0.12)	7.98	2.01	2.52	3.16	**<0.001**
Light	1.24 (0.19)	6.65	2.39	3.44	4.95	**<0.001**
Noise	0.35 (0.49)	0.72	0.55	1.42	3.71	0.47

Table 6: Statistical analysis of the relationship between modality and previous context and time of periods using multilevel logistic regression. Three sensor-based trigger conditions (CO_2, Human, and Light) and two time periods (Afternoon and Night) were significant predictors of user responses.

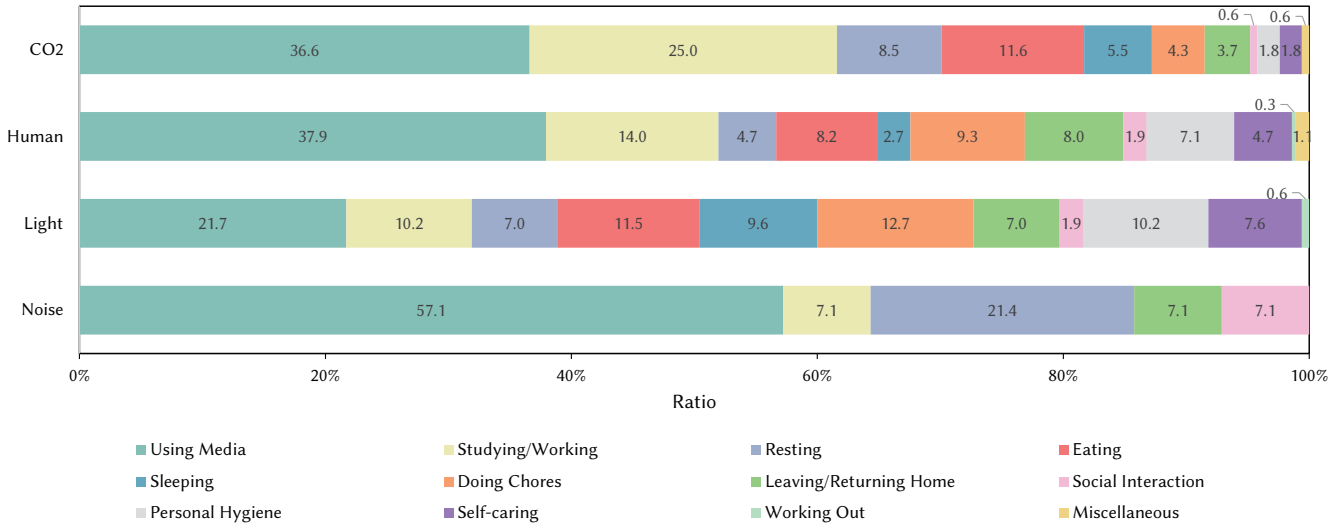


Figure 9: Distributions of previous activities across sensor-based trigger conditions. Overall, the ratios of using media, studying/working, and resting activities appear prominently across all types of triggers.

and performing tasks that require high concentration (e.g., studying/working or online meetings) (five users: P15, P6, P7, P12, and P14), and being in other areas such as the restroom and kitchen. (1 participant: P12). Similar to a previous study [11], these findings suggest that contextual factors (e.g., the cognitive load of ongoing activities and auditory/voice availability) before engaging in proactive interactions are important in determining opportune moments for proactive smart speakers.

7 RESULS ON INTERACTION MODALITY FOR ESM RESPONSES

In this section, we first analyze user preferences in interaction modality selections (e.g., which modality users prefer under what circumstances), followed by the user contexts that influence modality selections. Finally, we analyzed changes in modality selections over time.

7.1 Interaction Modality Preferences

Users can answer multiple-choice mental health surveys (e.g., PHQ-2 and GAD-2) in three different ways: by clicking on the screen

(GUI), by voice (VUI), or by alternating between the two modalities (MIXED). The analysis showed that most responses (93.8%, 2,065/2,201) were obtained via the GUI during self-tracking. In contrast, they responded via MIXED in 3.5% of the cases (76/2,201) and via voice-only in 2.7% of the cases (60/2,201).

In our interviews, most users reported that for the multiple-choice questionnaires, they responded via the GUI because they had to press the start button at the beginning of the survey and look at the picture for the card picture description tasks. Twelve users mentioned that they preferred the GUI because of the limitations of voice modality: (1) the necessity to wait until the speaker completes speaking, (2) low performance in voice recognition, and (3) high familiarity with GUI interactions. First, for the voice-only task, they had to wait while the speaker's guidance audio was played, and it took an extended period to recognize the user's voice. For example, P4 said, *"I usually used buttons (on the screen) because answering with voice took a longer time. I had to listen to the whole audio of the speaker to use the voice to answer. And the buttons were faster and more accurate."* P10 also mentioned, *"I wanted to complete the answer quickly, but if I speak, speaker audio playing and voice recognition take time. So even if I am lying in bed, when the questionnaire comes up, I use a touch more than 90% of the time because it's faster and more convenient to touch."* Second, speakers often did not recognize their voices properly, leading them to prefer responding via the GUI. P13 noted, *"When I said the number 'one' (for voice response), sometimes it was recognized and sometimes it wasn't. Rather than giving it another try, I just used touch because it was faster."* Finally, some participants preferred using the GUI because it was a familiar interface. For instance, P1 said, *"I preferred the GUI because I used to answer the survey through the GUI."* P8 said, *"I like to read books, so I'm more familiar with text printed or displayed. Due to the intuitive readability, (GUI) was preferred."*

Nevertheless, VUI was notably preferred, particularly when the hands were occupied – performing tasks that required the use of hands. P9 noted, "*When I was doing something that required the use of my hands (like doing laundry), I used the voice.*" and P15 said similarly, *"I used the voice when I was cooking or using the microwave, even though I was mostly using touch."* In addition, P18 mentioned that *"Voice was used for situations that I couldn't touch the screen, such as changing clothes or doing makeup."* Others preferred the VUI because it provided time to think about the question while the voice prompts were playing (P8: *"I think the voice prompts gave me more time to think, so I was able to think more deeply while listening to the voice prompts."*), or it helped them familiarize themselves with the questionnaire content at the beginning of the period (P13: "*I preferred the prompts on the screen, but the voice prompts helped me familiarize myself with the questionnaire content at the beginning.*")

7.2 User Contexts for Modality Preferences

Our interviews revealed a general preference for GUI, but the VUI was preferred in scenarios where users' hands were occupied. Therefore, we explored which previous activities influenced users to respond to ESM with their voices (VUI/MIXED) either solely or in conjunction with GUI. Figure 11 shows users' previous activities before answering the ESM surveys for GUI and VUI/MIXED. Regarding social interactions (pink box), the GUI (4.3%) was more frequent than the VUI/MIXED (0.7%). For doing chores (orange box) and personal hygiene (gray box), the VUI/MIXED (10.3%) was more frequent than GUI (3.0%).

Next, as shown in Table 7, we statistically analyzed whether previous activities significantly affected response modality. For the dependent variable, we included whether participants responded via a GUI (i.e., GUI vs. VUI/MIXED), whereas the participants' previous activities were included as a fixed effect. We examined when GUI is used only versus response modalities that include some voice components. The conditional and marginal R^2 were 0.217 and 0.069, respectively. Our analysis shows that the participants were more likely to respond via VUI/MIXED for doing chores (β=0.95, OR=2.58, $p < .001$) and personal hygiene (β=1.42, OR=4.13, $p < .001$) when compared to using media.

We further explored the proportion of doing chores and personal hygiene across the three modality types. As shown in Figure 10, doing chores appeared more frequently for the VUI (11.7%) and MIXED (9.2%) than for the GUI (4.8%). Personal hygiene did not appear for GUI, whereas it appeared in 8.3% and 11.8% of the cases for VUI and MIXED, respectively. Given that both activities are more likely to require both hands, this suggests that, in home settings, users prefer VUIs or MIXED when their hands are occupied.

7.3 Changes in Modality Preferences over Time

Table 8 shows the percentage of usage of the three response interfaces each week. A repeated-measures ANOVA was conducted to examine the percentage of responses using the GUI over time. The sphericity assumption was tested using Mauchly's test, which was not significant ($p = .486$). The result shows that there was no significant main effect ($F(3, 70) = 1.038$, $p = .381$).

Table 8 shows the percentages of usage of the three response interfaces each week. We analyzed whether modality preferences changed throughout the four-week study. Specifically, we conducted a repeated-measures ANOVA with the percentage of VUI usage as a dependent variable and the number of weeks as an independent variable. We tested its sphericity assumption using Mauchly's test, and it was satisfied ($p = .486$). The result was insignificant ($F(3, 70) = 1.038$, $p = .381$).

Throughout the study, most participants consistently preferred the GUI. Only a few participants used the VUI or MIXED, but eventually, by the fourth week, they shifted to the GUI. This change was influenced by the system errors encountered in the VUI and the convenience offered by the GUI. P16 said, "*In the beginning, when I didn't know the questions, I listened to the whole guidance audio of the speaker and answered by voice a few times. However, later on, I already knew the questions, so I skipped the voice and proceeded to respond by touch.*"

8 DISCUSSION

We designed a mental health self-tracking system that proactively requested ESM surveys via a multimodal smart speaker in a home setting, and evaluated the system with 20 participants with mild depressive symptoms. In this section, we review our major findings and discuss several design considerations for sensor selection, context-adaptation support, engaging interaction design, and system design.

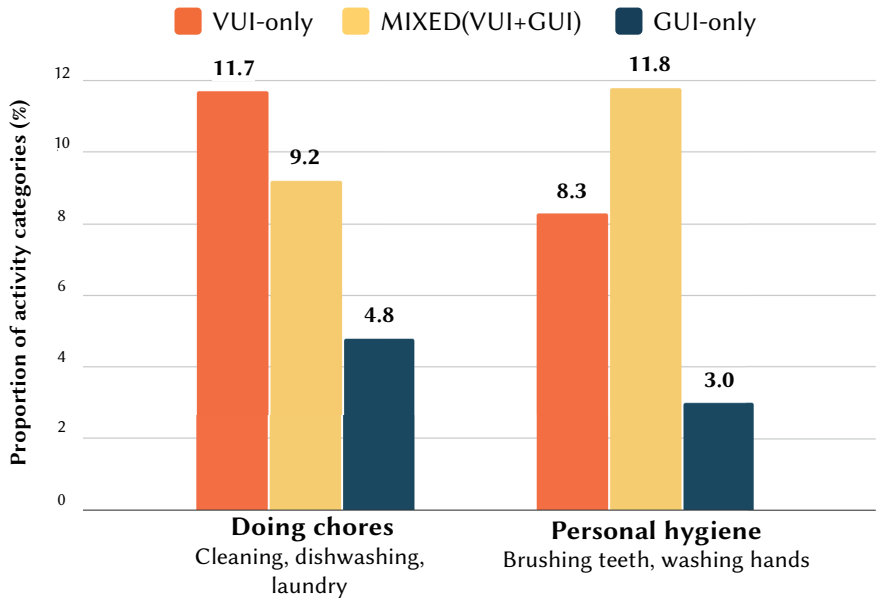


Figure 10: Occurrence of doing chores and personal hygiene activities across response modalities (GUI vs. VUI vs. MIXED).

Figure 11: Previous activities before answering the surveys for each response modality. Overall, the ratios of using media, studying/working, and resting activities appear prominently across all types of response modalities.

Context-Aware ESM in Home Environments: We demonstrate that user compliance (i.e., response rates) for ESM surveys can be improved when users are requested in activity transition contexts. We employed IoT sensors (e.g., CO2, light, noise, and camera sensors) to monitor these contexts. Other IoT sensors such as vibration, motion, and electric current sensors, can also be used to monitor activity transition contexts. IoT sensors effectively detect user presence in homes, bypassing the need for Bluetooth connectivity between smart speakers and phones [57], which may not always be carried by users. Among the IoT sensors, users are mostly concerned with image sensors (i.e., cameras), which are used for human presence detection. Although we informed our users that the cameras were only used for real-time human presence detection without saving any images, they were concerned about potential privacy risks. This aligns with findings from prior studies on smart home privacy risks [9, 61], where users exhibited coping behaviors such as covering up speakers during sensitive moments. To reduce privacy

Predictors	B (SE)	z-statistic	*95% CI for odds ratio*			p-value
			Lower	*Odds ratio*	*Upper*	
(Intercept)	-3.08 (0.25)	-12.29	0.03	0.05	0.08	**<0.001**
Previous Activity Contexts						
Doing Chores	0.95 (0.34)	-1.61	1.32	2.58	5.05	**<0.001**
Eating	0.24 (0.09)	0.23	0.66	1.25	2.37	0.49
Leaving/Returning	-0.02 (0.54)	-0.04	0.34	0.98	2.82	0.97
Miscellaneous	0.91 (0.78)	1.17	0.54	2.48	11.33	0.24
Personal Hygiene	1.42 (0.36)	4.00	2.06	4.13	8.30	**<0.001**
Resting	0.12 (0.35)	0.34	0.57	1.13	2.22	0.73
Self-caring	0.04 (0.63)	0.07	0.31	1.05	3.58	0.94
Sleeping	-0.40 (0.41)	-0.95	0.30	0.67	1.53	0.34
Social Interactions	-1.89 (1.01)	-1.87	0.02	0.15	1.10	0.06
Studying/Working	0.21 (0.29)	0.74	0.70	1.24	2.17	0.46
Working Out	0.35 (1.07)	0.32	0.17	1.41	11.47	0.75

Table 7: Statistical analysis of the relationship between modality and previous context using multilevel logistic regression. Two activities (Doing Chores and Personal Hygiene) were significant predictors of VUI or MIXED.

	Week 1	Week 2	Week 3	Week 4
Number of responses with the GUI interface	565	526	512	462
Number of responses with the MIXED interface	21	18	26	11
Number of responses with the VUI interface	11	19	16	14
Total number of responses in each week	597	563	554	487
GUI response rate (%)	**94.6**	**93.4**	**92.4**	**94.9**
MIXED+VUI response rate (%)	**5.4**	**6.6**	**7.6**	**5.1**

Table 8: Percentages of usage of three response interfaces by each week: Usage ratios of three response interfaces were shown weekly basis.

risks, future systems may use less privacy-invasive sensors for motion detection such as passive infrared sensors or low-resolution time-of-flight range imagers.

Context-Aware Smart Speakers for Multi-User Home Environments: Home inherently functions as a multi-user platform that encompasses diverse individuals who live in or access these spaces. These individuals include not only family members, such as partners, parents, and children, but also roommates, guests, and household employees. Given that this study focused on single-person households, extending it to multi-user home environments could be an interesting direction for future work. Although our speakers focused on ambient sensing, which gathers data about the surroundings of smart speakers using IoT sensors, our context-aware smart speakers can be extended to consider a user's mobile and wearable devices (e.g., the integration of user-device interaction data and sensing information). For example, integrating multimodal wearable sensors can help recognize multi-user activities in home environments [54]. Such considerations not only allow for a more detailed and accurate understanding of user context but also enable a better distinction of individual activities within a multi-user home environment. Furthermore, mobile and wearable device monitoring (e.g., user interaction and motion sensor data) offer new methods for detecting diverse interaction opportunities, as shown in prior work on multi-device-based breakpoint detection (e.g., interaction and physical activity breakpoints) [44]. In summary, this type of conjunction can provide diverse opportunities to interact proactively with users in both single- and multi-user home environments.

Improving Context Awareness for Modality Selection and Adaptation: Proactively requesting sensor-based ESM surveys can increase user compliance with mental health self-tracking. However, we also found that the users' previous contexts determined availability and influenced their interaction modality selection. Thus, it may be helpful to have fine-grained activity recognition in home contexts, beyond simple activity transition detection. Users prefer voice-based responses when multitasking, such as performing household chores or personal hygiene, because they find it difficult to manipulate the screen with their hands. It would be helpful to detect such user contexts using sensors and adaptively select an appropriate response modality. For example, when a multimodal smart speaker detects a scenario that requires using hands (e.g.,

housework or brushing teeth), it should turn up the speaker's voice volume or deliver all survey content by voice such that the user can respond to the survey.

Consideration for Engaging ESM Interaction Design: Self-tracking helped users with self-reflection, and they generally felt comfortable answering their health questions using multimodal agents.

Our users perceived human likeness based on voice dialogues, speaker avatars, and randomized ESM timings. However, a monotonous speaking tone and repetitive content lead to reduced engagement. Users expressed boredom due to the constant repetition of the same questionnaires. One solution to address this issue (e.g., improving user engagement) could be to vary the tone and content of self-tracking ESM surveys, as in context-tailored adaptations [26]. Given the importance of preserving the validity of survey contents in ESM, we can only vary the content of greeting dialogues or response feedback, by randomly selecting them from a database or generating them using large language models. Another strategy is to encourage self-reflection on data. Our users recognized the value of self-reporting in understanding their mental health conditions but also emphasized the need for visualization support for reviewing their self-reported data. Although ESM studies have mostly focused on collecting ecologically valid data, we should enable participants to review their reported data to manage their mental health. This feature can be implemented using a multimodal speaker by allowing users to visually explore their data through a display while interacting with the speaker.

Consideration for Systems Design: Our multi-modal speakers support multimodal interactions, but current interactions are largely GUI-based, mainly because commercial smart speaker platforms such as Google Assistant and Amazon Echo have a default no-speech timeout of 8 s, which cannot be altered. Thus, we decided to incorporate button-touching to address this limitation (e.g., preparation at the beginning or photo description stage). This implies that users can proceed to the next step with speech input before the timeout, but the physical button must be touched after the timeout. Unlike chat-based conversation interactions, voice-based interactions assume *synchronous turn-taking* in dialogue management [47], which must be carefully considered in the ESM conversation design. Context recognition software was implemented via a mobile phone as a hub for connecting and processing various sensor data, and backend storage (e.g., Google Cloud storage) was used to enable communication with conversational agents. For local data processing, we can use the Jetson Nano and Raspberry Pi 4 rather than mobile phones for system building [57, 59]. The context-sensing parts must be *tightly synchronized* with conversational interactions if researchers aim to support timely context-adaptive conversations in ESM.

9 CONCLUSION

Considering the need for mental health self-tracking in home settings, we propose a context-aware self-tracking system with a multimodal speaker that proactively requests ESM surveys by detecting activity transitions using sensors. To evaluate the user experience of the system, we conducted a 4-week field study with 20 participants. The results showed that proactive ESM delivery with sensing facilitated user engagement and compliance and provided positive self-tracking experiences. Touch-based interactions were dominant because of their ease of input; however, voice-based interactions occasionally occurred in multitasking scenarios. Furthermore, we discuss several practical design considerations, such as sensor selection, context adaptation, and interaction design. Our work demonstrates the feasibility of leveraging multimodal smart speakers and home IoT sensing to enable context-aware self-tracking of mental health. We encourage researchers and practitioners to utilize this innovative platform for advancing clinical research and developing new service designs, opening new avenues in mental health.

ACKNOWLEDGMENTS

This research was supported by LGE-KAIST Digital Health Research Center (DHRC) and by the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. 2022M3J6A1063021)

REFERENCES

[1] Marije aan het Rot, Koen Hogenelst, and Robert A Schoevers. 2012. Mood disorders in everyday life: A systematic review of experience sampling and ecological momentary assessment studies. *Clinical psychology review* 32, 6 (2012), 510–523.

[2] Piotr D Adamczyk and Brian P Bailey. 2004. If not now, when? The effects of interruption at different moments within task execution. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 271–278.

[3] Christoph Anderson, Clara Heissler, Sandra Ohly, and Klaus David. 2016. Assessment of social roles for interruption management: a new concept in the field of interruptibility. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct*. 1530–1535.

[4] 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.

[5] 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.

[6] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–24.

[7] Anna L Beukenhorst, Jamie C Sergeant, Max A Little, John McBETH, and William G Dixon. 2018. Consumer Smartwatches for Collecting Self-Report and Sensor Data: App Design and Engagement.. In *MIE*. 291–295.

[8] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis*. American Psychological Association.

[9] AJ Bernheim Brush, Bongshin Lee, Ratul Mahajan, Sharad Agarwal, Stefan Saroiu, and Colin Dixon. 2011. Home automation in the wild: challenges and opportunities. In *proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2115–2124.

[10] Sandra Bucci, Matthias Schwannauer, and Natalie Berry. 2019. The digital revolution and its impact on mental health care. *Psychology and Psychotherapy: Theory, Research and Practice* 92, 2 (2019), 277–297.

[11] 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.

[12] Eun Kyoung Choe, Bongshin Lee, Matthew Kay, Wanda Pratt, and Julie A Kientz. 2015. SleepTight: low-burden, self-monitoring technology for capturing and reflecting on sleep behaviors. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 121–132.

[13] Mihaly Csikszentmihalyi, Reed Larson, and Mihaly Csikszentmihalyi. 2014. The experience sampling method. *Flow and the foundations of positive psychology: The collected works of Mihaly Csikszentmihalyi* (2014), 21–34.

[14] Anind K Dey, Katarzyna Wac, Denzil Ferreira, Kevin Tassini, Jin-Hyuk Hong, and Julian Ramos. 2011. Getting closer: an empirical investigation of the proximity of user to their smart phones. In *Proceedings of the 13th international conference on Ubiquitous computing*. 163–172.

[15] Joel E Fischer, Chris Greenhalgh, and Steve Benford. 2011. Investigating episodes of mobile phone activity as indicators of opportune moments to deliver notifications. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services*. 181–190.

[16] Paul Freihaut, Anja S Göritz, Christoph Rockstroh, and Johannes Blum. 2021. Tracking stress via the computer mouse? Promises and challenges of a potential behavioral stress marker. *Behavior Research Methods* (2021), 1–21.

[17] Javier Hernandez, Daniel McDuff, Christian Infante, Pattie Maes, Karen Quigley, and Rosalind Picard. 2016. Wearable ESM: differences in the experience sampling method across wearable devices. In *Proceedings of the 18th international conference on human-computer interaction with mobile devices and services*. 195–205.

[18] 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*. 909–918.

[19] Institute of Health Metrics and Evaluation. 2019. 2019 Global Burden of Disease (GBD). https://vizhub.healthdata.org/gbd-results/. [Online; accessed 14-May-2022].

[20] Shamsi T Iqbal and Brian P Bailey. 2010. Oasis: A framework for linking notification delivery to the perceptual structure of goal-directed tasks. *ACM Transactions on Computer-Human Interaction (TOCHI)* 17, 4 (2010), 1–28.

[21] Hayeon Jeong, Heepyung Kim, Rihun Kim, Uichin Lee, and Yong Jeong. 2017. Smartwatch wearing behavior analysis: a longitudinal study. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–31.

[22] Haihua Jiang, Bin Hu, Zhenyu Liu, Lihua Yan, Tianyang Wang, Fei Liu, Huanyu Kang, and Xiaoyu Li. 2017. Investigation of different speech types and emotions for detecting depression using different classifiers. *Speech Communication* 90 (2017), 39–46.

[23] Soowon Kang, Cheul Young Park, Auk Kim, Narae Cha, and Uichin Lee. 2022. Understanding Emotion Changes in Mobile Experience Sampling. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 198, 14 pages. https://doi.org/10.1145/3491102.3501944

[24] Soowon Kang, Cheul Young Park, Auk Kim, Narae Cha, and Uichin Lee. 2022. Understanding Emotion Changes in Mobile Experience Sampling. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–14.

[25] Young-Ho Kim, Diana Chou, Bongshin Lee, Margaret Danilovich, Amanda Lazar, David E Conroy, Hernisa Kacorri, and Eun Kyoung Choe. 2022. Mymove: Facilitating older adults to collect in-situ activity labels on a smartwatch with speech. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–21.

[26] Predrag Klasnja, Shawna Smith, Nicholas J Seewald, Andy Lee, Kelly Hall, Brook Luers, Eric B Hekler, and Susan A Murphy. 2019. Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of HeartSteps. *Annals of Behavioral Medicine* 53, 6 (2019), 573–582.

[27] Sanne Koops, Sanne G Brederoo, Janna N de Boer, Femke G Nadema, Alban E Voppel, and Iris E Sommer. 2023. Speech as a Biomarker for Depression. *CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders)* 22, 2 (2023), 152–160.

[28] Kaylee Payne Kruzan, Ada Ng, Colleen Stiles-Shields, Emily G Lattie, David C Mohr, and Madhu Reddy. 2023. The Perceived Utility of Smartphone and Wearable Sensor Data in Digital Self-tracking Technologies for Mental Health. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[29] Danielle C Lavallee, Jenney R Lee, Elizabeth Austin, Richard Bloch, Sarah O Lawrence, Debbe McCall, Sean A Munson, Mara B Nery-Hurwit, and Dagmar Amtmann. 2020. mHealth and patient generated health data: stakeholder perspectives on opportunities and barriers for transforming healthcare. *Mhealth* 6 (2020).

[30] 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.

[31] Kwangyoung Lee and Hwajung Hong. 2017. Designing for self-tracking of emotion and experience with tangible modality. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. 465–475.

[32] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 557–566.

[33] Zhenyu Liu, Changcong Li, Xiang Gao, Gang Wang, and Jing Yang. 2017. Ensemble-based depression detection in speech. In *2017 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE, 975–980.

[34] Hailiang Long, Zhenghao Guo, Xia Wu, Bin Hu, Zhenyu Liu, and Hanshu Cai. 2017. Detecting depression in speech: Comparison and combination between different speech types. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 1052–1058.

[35] LP Information, Inc. 2023. Global Smart Speakers with Display Market Growth 2023-2029. https://www.marketresearch.com/LP-Information-Inc-v4134/Global-Smart-Speakers-Display-Growth-33110157/. [Online; accessed January-2023].

[36] Arnold M Lund. 2001. Measuring usability with the use questionnaire12. *Usability interface* 8, 2 (2001), 3–6.

[37] Deborah Lupton. 2016. *The quantified self*. John Wiley & Sons.

[38] Alberto Machado, Antonio J Herrera, Rocío M de Pablos, Ana María Espinosa-Oliva, Manuel Sarmiento, Antonio Ayala, José Luis Venero, Martiniano Santiago, Ruth F Villarán, María José Delgado-Cortés, et al. 2014. Chronic stress as a risk factor for Alzheimer's disease. *Reviews in the Neurosciences* 25, 6 (2014), 785–804.

[39] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

[40] Kimberly D Mueller, Bruce Hermann, Jonilda Mecollari, and Lyn S Turkstra. 2018. Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. *Journal of clinical and experimental neuropsychology* 40, 9 (2018), 917–939.

[41] Shinichi Nakagawa and Holger Schielzeth. 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4, 2 (2013), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

[42] Christie Napa Scollon, Chu-Kim Prieto, and Ed Diener. 2009. Experience sampling: promises and pitfalls, strength and weaknesses. *Assessing well-being: the collected works of ED Diener* (2009), 157–180.

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

[44] Tadashi Okoshi, Julian Ramos, Hiroki Nozaki, Jin Nakazawa, Anind K Dey, and Hideyuki Tokuda. 2015. Reducing users' perceived mental effort due to interruptive notifications in multi-device mobile environments. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 475–486.

[45] Seung-Jin Park, Hye-Ra Choi, Ji-Hye Choi, Kun-Woo Kim, and Jin-Pyo Hong. 2010. Reliability and validity of the Korean version of the Patient Health Questionnaire-9 (PHQ-9). *Anxiety and mood* 6, 2 (2010), 119–124.

[46] Rashmi Patel, Theodore Lloyd, Richard Jackson, Michael Ball, Hitesh Shetty, Matthew Broadbent, John R Geddes, Robert Stewart, Philip McGuire, and Matthew Taylor. 2015. Mood instability is a common feature of mental health disorders and is associated with poor clinical outcomes. *BMJ open* 5, 5 (2015), e007504.

[47] Cathy Pearl. 2016. *Designing Voice User Interfaces*. O'Reilly.

[48] Michelle M.E. Van Pinxteren, Mark Pluymaekers, and Jos G.A.M. Lemmink. 2022. Human-like Communication in Conversational Agents: A Literature Review and Research Agenda. *Journal of Service Management* 31, 2 (2022), 203–225.

[49] Robert L Spitzer, Kurt Kroenke, Janet BW Williams, Patient Health Questionnaire Primary Care Study Group, Patient Health Questionnaire Primary Care Study Group, et al. 1999. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *Jama* 282, 18 (1999), 1737–1744.

[50] Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of internal medicine* 166, 10 (2006), 1092–1097.

[51] Timothy J Trull and Ulrich W Ebner-Priemer. 2009. Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: introduction to the special section. (2009).

[52] Niels Van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The experience sampling method on mobile devices. *ACM Computing Surveys (CSUR)* 50, 6 (2017), 1–40.

[53] Laura C Walz, Maaike H Nauta, and Marije aan het Rot. 2014. Experience sampling and ecological momentary assessment for studying the daily lives of patients with anxiety disorders: A systematic review. *Journal of anxiety disorders* 28, 8 (2014), 925–937.

[54] Liang Wang, Tao Gu, Xianping Tao, Hanhua Chen, and Jian Lu. 2011. Recognizing multi-user activities using wearable sensors in a smart home. *Pervasive and Mobile Computing* 7, 3 (2011), 287–298.

[55] 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.

[56] Rui Wang, Weichen Wang, Alex DaSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. 2018. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–26.

[57] Jing Wei, Tilman Dingler, and Vassilis Kostakos. 2021. Understanding User Perceptions of Proactive Smart Speakers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–28.

[58] Jing Wei, Weiwei Jiang, Chaofan Wang, Difeng Yu, Jorge Goncalves, Tilman Dingler, and Vassilis Kostakos. 2022. Understanding How to Administer Voice Surveys through Smart Speakers. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–32.

[59] Jing Wei, Benjamin Tag, Johanne R Trippas, Tilman Dingler, and Vassilis Kostakos. 2022. What Could Possibly Go Wrong When Interacting with Proactive Smart

Speakers? A Case Study Using an ESM Application. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.

[60] Peter West, Max Van Kleek, Richard Giordano, Mark J Weal, and Nigel Shadbolt. 2018. Common barriers to the use of patient-generated data across clinical settings. In *proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[61] Eric Zeng, Shrirang Mare, and Franziska Roesner. 2017. End user security and privacy concerns with smart homes. In *Symposium on Usable Privacy and Security (SOUPS)*, Vol. 220.

[62] Fred RH Zijlstra, Robert A Roe, Anna B Leonora, and Irene Krediet. 1999. Temporal factors in mental work: Effects of interrupted activities. *Journal of Occupational and Organizational Psychology* 72, 2 (1999), 163–185.