# AnnoSense: A Framework for Physiological Emotion Data Collection in Everyday Settings for AI

PRAGYA SINGH, IIIT-Delhi, India
ANKUSH GUPTA, IIIT-Delhi, India
MOHAN KUMAR, RIT, New York, US
PUSHPENDRA SINGH, IIIT-Delhi, India

Emotional and mental well-being are vital components of quality of life, and with the rise of smart devices like smartphones, wearables, and artificial intelligence (AI), new opportunities for monitoring emotions in everyday settings have emerged. However, for AI algorithms to be effective, they require high-quality data and accurate annotations. As the focus shifts towards collecting emotion data in real-world environments to capture more authentic emotional experiences, the process of gathering emotion annotations has become increasingly complex. This work explores the challenges of everyday emotion data collection from the perspectives of key stakeholders. We collected 75 survey responses, performed 32 interviews with the public, and 3 focus group discussions (FGDs) with 12 mental health professionals. The insights gained from a total of 119 stakeholders informed the development of our framework, *AnnoSense*, designed to support everyday emotion data collection for AI. This framework was then evaluated by 25 emotion AI experts for its clarity, usefulness, and adaptability. Lastly, we discuss the potential next steps and implications of *AnnoSense* for future research in emotion AI, highlighting its potential to enhance the collection and analysis of emotion data in real-world contexts.

CCS Concepts: • **Human-centered computing → Empirical studies in ubiquitous and mobile computing**; **Empirical studies in HCI**; • **Computing methodologies → Machine learning**.

Additional Key Words and Phrases: Mental Health, Emotion AI, Passive Sensing, Data Work, Wearable Devices, HCI, Emotion Recognition, Physiological Signals, Data-centric AI, Behavioral Sensing, Well-being, AI, Smartphones, Stress-tracking, Stressor-logging, Visualizations, Stress Intervention, Behavioral Change, Affective Computing, Wearable Sensors

**ACM Reference Format:**
Pragya Singh, Ankush Gupta, Mohan Kumar, and Pushpendra Singh. 2025. AnnoSense: A Framework for Physiological Emotion Data Collection in Everyday Settings for AI. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3, Article 131 (September 2025), 47 pages. https://doi.org/10.1145/3749519

## 1 Introduction

In today's fast-paced world, maintaining good mental health has become increasingly important. The challenges associated with monitoring emotional burnout and stress often lead to significant mental health issues and a diminished quality of life. Recent advancements in ubiquitous computing, wearable devices and mobile phones equipped with sensors for tracking physiological or behavioral changes and artificial intelligence (AI) algorithms are creating new opportunities for monitoring mental well-being [162, 171]. Several wearable and mobile phone-based interventions have been designed to monitor stress, sleep, mood, habits, and emotions [51, 109, 152]. These bio-signal data-driven technologies have the potential to support continuous monitoring,

Authors' Contact Information: Pragya Singh, IIIT-Delhi, New Delhi, India, pragyas@iiitd.ac.in; Ankush Gupta, IIIT-Delhi, New Delhi, India, ankush21232@iiitd.ac.in; Mohan Kumar, RIT, Rochester, New York, US, mjkvcs@rit.edu; Pushpendra Singh, IIIT-Delhi, New Delhi, India, psingh@iiitd.ac.in.


ACM 2474-9567/2025/9-ART131
https://doi.org/10.1145/3749519

providing capabilities for early diagnosis and enabling data-driven insights for mental health professionals [153]. Recent commercial developments highlight the increasing integration of physiological sensors into mainstream wearable devices. Consumer products such as smartwatches and smart rings now commonly include sensors for Photoplethysmography (PPG), Electrodermal Activity (EDA), and Skin Temperature (SKT), in addition to standard accelerometer and gyroscope components. These combined features enable more comprehensive emotion and stress assessments. For example, the Oura Ring categorizes stress into four states—stressed, engaged, relaxed, and restored—to help users monitor, understand, and manage daily stressors [11, 98, 99, 103]. Major wearable brands such as Apple [12], Samsung [128], Fitbit [40], and Garmin [43] have incorporated stress and mental well-being tracking into their devices and companion apps. Products like the Oura Ring [105], Apple Watch [12], WHOOP [168], and Samsung wearables [128] now also feature emotion journaling, reflecting a growing emphasis on physiological approaches to mental health monitoring. Moreover a new category of devices physiological-sensing devices called *Earables* are also emerging [120]. In addition to wearables, mobile phone applications are playing an increasingly important role in tracking emotions and mental well-being. Notable examples include Wysa [5], Woebot [4], Calm [1], Daylio [2], and Headspace [3], which provide tools for emotion monitoring, cognitive support, and mindfulness.

Physiological signal based-emotion tracking still remains in its early stages, especially when it comes to continuous and reliable monitoring in everyday life [51]. While there is growing commercial interest and increasing integration of physiological sensors in consumer wearables, current systems are limited to stress tracking and are lacking the sensitivity, personalization, and contextual awareness needed to fully capture the complexity of human emotions in real-world settings. AI-powered solutions that combine wearable devices with mobile phone applications present a promising approach to addressing these challenges, especially given recent advancements and the growing adoption of AI in tackling complex, real-world problems. However, a major limitation in developing such models is their heavy reliance on high-quality, labeled emotion data that accurately reflects the nuances of human emotional experiences [91, 126, 138, 169]. Most existing emotion datasets are collected in controlled laboratory environments or through short-term studies using self-report methods or expert-annotations. These approaches frequently fall short of capturing the dynamic, context-dependent, and multi-layered nature of emotions as they naturally unfold [33, 41, 64, 138, 139]. Consequently, models trained on such data often struggle to generalize to real-world scenarios, limiting their effectiveness and reliability for end users [41, 97, 126, 169], suggesting a need for methods to accurately capture emotional annotations and contextual information in real-life settings. Prior work in real-life settings has primarily relied on self-assessment approaches, such as the Experience Sampling Method (ESM)—which prompts users to report their emotions at regular or random intervals throughout the day—and the Day Reconstruction Method (DRM), where participants provide retrospective reports of their emotional states via structured questionnaires at the end of the day [45, 72, 130, 144, 157]. While useful for capturing momentary or reflective self-reports, these methods often provide labels using predefined emotional scales (such as, Self-Assessment Manikin (SAM) [25] or scales based on six basic emotions [37]), offering limited insights on contextual information about emotional experiences. Additionally, previous studies have highlighted further challenges; for instance, the act of annotation itself may influence the user's emotional state, introducing measurement bias [61]. Moreover, participants often experience annotation fatigue, leading to low motivation and engagement over time [175]. Furthermore, prior research has also emphasized the need for ecologically valid, human-centric data collection approaches that reflect how emotions are naturally experienced in everyday contexts [41, 126, 138] These insights suggests a significant gap in studies that investigate in existing methodologies from participant perspectives, who are not only the sources of data but also its annotators [41, 138].

To explore the challenges and identify opportunities in everyday emotion annotation methods for physiological signal-based emotion AI (referred to as "emotion AI" in this paper) research using wearables and mobile phone data, we designed this study. Our approach centers on examining the perspectives of diverse stakeholders,

including both users and non-users of emotion-tracking technologies, as well as mental health professionals. Through this lens, we aim to understand the human side of emotion data collection and its implications for designing more effective and user-centric emotion AI systems. Specifically, we address the following research questions:

**RQ1:** What are participants' perspectives on the challenges and opportunities for annotating (identifying and labeling) and tracking emotions in their daily lives?

**RQ2:** What are the perspectives of mental health professionals on the challenges and opportunities in physiological emotion data collection for developing emotion AI interventions?

**RQ3:** How can participants' and domain experts' perspectives be integrated to develop a holistic methodology for collecting physiological emotion data in real-life settings?

We employed a qualitative research method, including surveys (n = 75) and interviews (n = 32) with members of the public (with and without experience in therapy or counseling, as well as those who have used wearables and mobile phone applications for tracking stress and emotions, or participated in emotion data collection studies), as well as focus group discussions (n = 3) involving 12 mental health professionals. Following our methodology, our study evaluated the needs of participants from diverse perspectives, covering a total sample size of 119 participants. In this study, *emotion* is defined in line with the "Theory of Constructed Emotion" as proposed by Lisa Barrett [15], which views emotions - "as individualized, context-dependent experiences constructed by the brain through the interpretation of bodily sensations (e.g., heart rate, arousal) in relation to past experiences, situational context, and learned emotional concepts, not as fixed biological responses". In contrast with the prior research where *emotions* are typically modeled as changes in physiological response and behavioral reaction, using physiological signals (e.g., heart rate, skin temperature, electrodermal activity), behavioral patterns, and self-reported data [110]. By adopting Barrett's perspective, this study emphasizes the context-dependent and dynamic nature of emotions, enabling a more comprehensive exploration of emotional experiences. Also, for this study, we have referred to structured methods, such as scales like PANAS, SAM, or Likert scales, as objective methods, while unstructured methods, such as providing an option to write, audio record, or add images, are referred to as subjective methods. This human-centric view of *emotions* helps us to study the subjectivity and variability of emotional experiences, challenging the assumption that specific physiological or behavioral patterns can map directly to discrete or dimensional emotion categories [139]. Our paper makes the following key contributions:

- **Annosense Framework:** We introduce Annosense, a novel framework comprising 15 actionable guidelines for collecting well-annotated wearable and mobile-based emotion data in everyday settings. These guidelines are derived from an in-depth analysis of user experiences, contextual challenges, and common challenges in emotion data collection.
- **Expert Evaluation**: We evaluate the Annosense framework through feedback from 25 emotion AI experts with professional and academic experience, making this the first work to present evaluated guidelines tailored specifically for real-world emotion data collection.
- **Potential Implementation**: We identify next steps for designing participant-aware systems in practice, informed by the Annosense framework and a review of current tools, technologies, and applications.
- **Design Implications**: Through our findings and expert discussions, we offer design recommendations for future emotion data collection practices and AI algorithm development.

Finally, it is crucial to note that this work introduces a new paradigm of designing emotion data-collection studies from participants' and domain experts' perspectives. These contributions are significant for the Ubiquitous Computing (UbiComp), Human Computer Interaction (HCI), and affective computing communities as they address a critical gap in how emotional data is collected and validated in real-world, everyday settings, moving beyond controlled lab environments. By offering a systematically developed and expert-evaluated framework, this work equips researchers and practitioners with practical guidelines that are grounded in user context,

data-centric AI practices, and ethical considerations. This not only advances methodological rigor in affective computing but also aligns with HCI's emphasis on human-centered design, participatory development, and context-sensitive technologies. Moreover, by outlining pathways for implementation, the work bridges theory and practice, supporting the development of emotion-aware technologies that are both technically sound and socially responsible.

## 2 Related Work

### 2.1 Understanding Emotions - From Emotion Theories to Frameworks

The definition of Emotions has been a widely debated topic among researchers across various domains, including philosophy, psychology, and neuroscience. It has evolved over time alongside the development of different theoretical perspectives on what emotions are and how they function [150]. Early theories of emotions have defined emotions as an outcome of evolution [112], where theorists have classified emotions into distinct sets of basic emotions that are universal. In Ekman's Basic Emotion Theory [37], emotions have been classified into six basic emotions - anger, fear, disgust, happiness, surprise, and sadness - that are biologically hardwired and are universally recognized. In contrast, Appraisal theories, pioneered by psychologists like Richard Lazarus [81], defined emotions as the result of individual cognitive appraisals of events, accounting for subjectivity in emotional responses, and suggesting emotions as a response to how a person interprets a situation. More recently, Lisa Feldman Barrett's "Theory of Constructed Emotion" [15] proposed emotions as an active construction by the brain based on a combination of sensory input, past experiences, and cultural learning, and not something universal. Within computational approaches, emotions are often defined as complex psychological states with key components such as subjective experiences and behavioral and physiological responses. This lack of theoretical consensus has significant implications for physiological signal-based emotion AI. Unlike fields such as computer vision and NLP, where data modalities (e.g., images, videos, text) are well-defined, and labeling methods are standardized as per the task, such as classification or segmentation. Data in Emotion AI vary widely both in terms of the signal modalities collected, ranging from ECG, PPG, EDA, and heart rate variability to EEG, EMG, and fMRI, and the frameworks applied for emotion annotation (e.g., discrete categories, dimensional models like valence-arousal, hybrid or custom approaches). As a result, generalization across models is limited, reproducibility suffers, and the lack of interoperability between datasets complicates benchmarking and progress [38, 96, 114, 143].

Despite these challenges, there are growing efforts to impose structure through methodological guidelines, annotation standards, and sensing toolkits. Researchers in computer science have borrowed standardized self-reporting scales and questionnaires from psychology for emotion annotations. Some of the most widely used scales include the Positive and Negative Affect Schedule (PANAS), which asks participants to rate how they experience different emotions [166], and the Self-Assessment Manikin (SAM), which helps users to rate their feelings along dimensions like valence (pleasant–unpleasant), arousal (excited–calm), and dominance [25]. These tools provide a structured way for researchers to capture emotional states. Further, researchers have also developed frameworks for keeping the interdisciplinary nature of emotion research in mind. For instance, frameworks like the HUMAINE project [35] and Emotion Annotation and Representation Language (EARL) [131] were developed to offer a common format for labeling emotions across different modalities such as speech, facial expression, and physiological signals, helping standardize how emotions are defined and interpreted in computational systems. Additionally, these frameworks provide guidelines for collecting emotion annotations in different use cases. Beyond these initial frameworks, for real-world emotion data collection, researchers have developed digital phenotyping and mobile sensing tools for context-aware ecological momentary assessments (EMAs). Popular tools includes, MindLamp [155], Beiwe [104], AWARE [39], AWARE-Light [156], PACO [46], Sensingkit [63], mEMA [56], Experiencesampler [148], and MobileQ [89]. These tools provide a shared structure for sensor integration, self-reports, and temporal alignment of signal data. However, prior research has still emphasized

the challenge in grounding emotion data for machine learning algorithms, pointing out the need for adding participants' context to emotion data [64, 138]. Further researchers have also cautioned about the biases that emotion data contains and its impact on data quality [33]. Moreover, the need to enhance the methodological aspects of emotion data collection was also highlighted for improving the overall data quality [41]. This points to the significance of further improving on the nuances of current approaches by integrating experts' opinions and participants ' specific factors into the data collection methodology. However, an in-depth exploration of how both participants and domain experts perceive these methods remains lacking, highlighting the need for further work in bridging the gap between structured approaches and meaningful participant engagement.

## 2.2 Emotion Data Practices

Prior research in emotion and affect recognition has utilized a range of computational methods to infer emotional states, often relying on various data proxies [110]. Within the fields of ubiquitous computing and artificial intelligence, a variety of proxy data modalities have been explored to infer emotions, including physiological signals (e.g., heart rate, electrodermal activity), facial expressions, vocal characteristics, mobile sensor data, and textual data [106, 164]. This study specifically focuses on physiological signals and mobile sensing data, as these modalities offer a valuable combination of being continuously collectible and non-invasive. Their unobtrusive nature makes them particularly suitable for long-term, real-world emotion monitoring, enabling them to capture emotions with minimal participant disruption. In this section, we will discuss in detail how emotion data is collected and annotated in various settings.

**Lab-Based Emotion Datasets:** The primary reason for collecting emotion data in lab settings is the amount of control it provides over stimulus presentation and participant conditions [129, 133, 146]. As shown in Table 1, a variety of emotion elicitation techniques and annotation approaches have been explored by researchers within lab settings, including elicitation methods like images, videos, audio, virtual reality (VR), cognitive tasks, and physical activities. While these datasets provide rich multimodal records of emotions, their methodological choices vary widely. Moreover, many of these methods rely on rigid self-reporting methods such as PANAS [165], SAM [25], or scales/emotion categories based on evolutionary theories of emotions like Ekman's Basic Emotions [37], Kazemzadeh's 20 categories [66]). Further, there is minimal opportunity for participants to provide context on their emotional states [138].

**Real-life and Constrained Emotion Datasets:** Emotion data collection in real-life and constrained settings offers a balance between ecological validity and experimental control, contributing to a deeper understanding of emotional states in diverse contexts as shown in Table 2. Constrained task-based studies such as ForDigitStress [50], NURSE [53], and G-REx [22] often involve structured environments like job interviews, healthcare shifts during COVID-19, or prolonged exposure to emotion-eliciting media, where emotions are annotated through self-reports, physiological markers, or external ratings. These studies benefit from higher control over task and timing but face challenges such as context-driven label bias (like, Nurse dataset contains mostly negative emotion data due to the collection setting) and limited participant context in the annotation. On the other hand, semi-naturalistic datasets collected within constraint environments like colleges or workplaces such as StudentLife [162], Laureate [76], GLOBEM [171], DiversityOne [27], TILES [173], and the SWEET Study [141] rely on Ecological Momentary Assessment (EMA) for self-reports alongside other survey information and sensor data to capture context. Further contributions, such as DAPPER [135] and K-EmoPhone [60], explore daily-life emotional states through intensive prompting or periodic logging.

While these approaches yield high ecological realism, they also introduce challenges such as lower response rates and reactivity to prompts. Additionally, self-reports collected within natural settings often lack enough information for reliably contextualizing the collected self-reports [41, 42]. Despite methodological innovations, a key limitation across both types of data collection is the reliance on approaches designed to simply collect data

| Dataset | Elicitation Method | Annotation Approach |
|---|---|---|
| **WESAD** [129] | Video Clips, Public speaking, mental arithmetic, and Meditation | PANAS, SAM, State-Trait Anxiety Inventory (STAI), Short Stress State Questionnaire (SSSQ), physiological signals |
| **ASCERTAIN** [146] | Video Clips | Valence-Arousal, Engagement, Liking, Familiarity, Personality Traits |
| **CASE** [133] | Video Clips | Continuous Valence-Arousal Annotations |
| **Neurological Status** [21] | Physical Activities | Task-based Labels |
| **CLAS** [87] | Video Clips, Images, Math, Stroop, Logic tasks | Arousal-Valence, Task-based Labels |
| **VREED** [147] | VR Video Clips | SAM, PANAS |
| **POPANE** [18] | Speech preparation, Anticipation task, Interpersonal communication, Affective Images, and Video Clips | Discrete Emotion Categories, SAM, Avoidance Approach Motivation |
| **EMOGNITION** [124] | Audio-visual stimuli | Discrete Emotion Categories, SAM, Avoidance Approach Motivation |
| **StressID** [28] | Cognitive load tasks | SAM, Custom Perceived Stress Assessment |
| **BIRAFFE2** [74] | Games, Affective Music, and Images | SAM, Game Experience Questionnaire (GEQ) |
| **EEVR** [137] | VR Video Clips | Textual Descriptions, SAM, PANAS, Familiarity, Liking, Personality Traits |
| **KEMOCON** [108] | 10-minute-long debate on social issues | Self-report Valence-Arousal, Discreet Emotion Category, Partner Annotations, and Expert Annotation |
| **AMIGOS** [94] | Video clips (Long and Short) | SAM, PANAS, Personality traits |
| **RECOLA** [119] | Collaborative task (video chat) | SAM, PANAS |

Table 1. Lab-based emotion datasets: Elicitation methods, annotation strategies, and labeling approaches

without keeping participants' factors in mind [33, 138], which often led to quality issues in datasets. Recently, methods based on appraisal theories and constructive theories [159] have also been explored for labeling emotion data to capture more context-dependent labels [58, 78]. However, it remains in an initial phase, suggesting the need for more research in designing data collection methods within everyday settings.

| Dataset | Context/Task | Annotation Method |
|---|---|---|
| **ForDigitStress** [50] | Job interview tasks simulating time pressure | Custom Stress Scale and Saliva Cortisol |
| **NURSE** [53] | Healthcare workers during COVID-19 | Custom Stress Questionnaire |
| **G-REx** [22] | Long movie viewing sessions | Post-Hoc SAM Scale Based Tool |
| **Laureate** [76] | University setting with student academic routines | Custom EMA (PANAVA-KS, physical activity, breakfast ingestion, caffeine intake, study-time and sleep quality) |
| **StudentLife** [162] | University campus life over multiple weeks | Photographic Affect Meter (PAM) EMA, Single-item Stress EMA |
| **GLOBEM** [171] | Naturalistic daily experiences across diverse locations | EMA Survey (PHQ-4, PSS-4, PANAS), and Pre-Post Survey |
| **TILES** [173, 173] | Workplace monitoring in hospital environment | Single-item Stress EMA, Survey on daily stressors, work behaviors, and sleep |
| **DAPPER** [135] | Daily life across varied settings (field study) | 20-Item ESM (Information about daily events, Participants' openness to sharing emotion, TIPI-C, PANAS), DRM with Open-ended Question |
| **K-EmoPhone** [60] | Daily life across varied settings (field study) | Custom Questionnaire (Valence, Arousal, Attention, Stress, Emotion Duration, Task Disturbance, Emotion Change) |
| **SWEET Study** [141] | Office workers' daily routines in real-life settings | EMA (Stress, Activity, Food and Beverage Consumption, Sleep Quality, and Gastro-intestinal Symptoms) |

Table 2. Tasks and Annotation Methods in Semi-Controlled Emotion Datasets.

## 2.3 Ubiquitous Interventions for Emotion Annotations and Self-Reporting

Recently, ubiquitous computing and related communities have begun exploring interactive approaches for labeling emotional data [44, 52, 116, 160]. Notable works on emotional annotation include "Find the Bot" [172], which uses a web-based gaming platform to collect emotion annotations for machine learning algorithms, Reconexp [68], where participants were provided with a both mobile and web-based interface, mirrorU [161], which supports reflective writing through memory-based cues. Similarly, several other emotion measurement tools have been developed to support self-reporting through structured formats. These include the Affect Grid [123], the Differential Emotions Scale [24], and interactive tools such as Premo [34] and the Photographic Affect Meter (PAM) [113], where users select images that best represent their emotional state. Researchers have recently designed an interactive mobile version for the Geneva emotion wheel to support emotion self-reporting [136]. Further, prior works

like PResUP [14], a framework that probes users to self-report emotions opportunistically, and Mirror Ritual [117], which uses facial recognition to detect participants' emotions and generate poems to encourage emotional reflections, are also explored. Further techniques, like Diurnal Rhythms of Emotions, based on circadian rhythms [144], Technology-Assisted Reconstruction (TAR) [62] where passively collected data is used in assisting the later annotations at the end of the day [135], and Mirror Hearts [29], where AI-powered third-person view is leveraged for self-reporting emotions were also explored. These approaches aim to provide more accurate representations of human emotions, moving beyond the limitations of reductionist models and scale-based labels. Recently, LLM-based self-reporting and in-context journaling have also been explored for behavioral monitoring in everyday settings, including Dairyhelper [82], Mindshift [170], and Mindscape [101]. Despite ongoing work, emotion annotations in everyday settings remain challenging, as most of these methods are not translated for data collection methodology in noisy real-life settings. This paper aims to address this translation gap from stakeholders' (users and mental health professionals) perspectives to design novel participant-centric methods for emotion data collection in everyday settings.

### 2.4 Emotion Monitoring and Stakeholders' Perspective

Emotions have been studied for decades to improve human-machine or human-human interactions. Emotion recognition to support mental well-being [73, 149], employee well-being, productivity [122], individual monitoring, behavior tracking, and emotion regulation [13, 32, 140] have been studied in the past. Prior works have also explored data collection practices within AI from multiple stakeholders' points of view [127], highlighting the assumptions about data being something that is readily available to be used [111, 127]. However, this attitude of considering data as something readily available has often led to poor data quality and degraded the performance of AI algorithms. In the past, mental health monitoring solutions were scrutinized from users' perspectives [67, 163, 175]. Kelley et al., who work on students' mental health, highlighted the challenges of self-tracking [67, 175] and found motivation to be a major challenge due to factors, such as fear towards tracking negative emotion data. Further, Zhang et al. [175] studied the experiences of users suffering from depression and anxiety in using mental health tracking applications and highlighted that the use of customization (such as creating checklists for daily progress) is correlated with symptom severity. Another common theme among prior works on users' attitudes was the doubt towards the authenticity of digital tools as compared to humans, which can provide real-world care and social support [23, 79, 121]. Further prior work on the perceived utility of wearables for mental well-being [73] highlighted users' attitudes towards tracking as per needs, for instance, tracking for maintenance, for people with few symptoms, versus tracking for active symptom management, for people with more symptoms. Studies have also highlighted the users' views on the benefits of using mental health applications, such as support in identifying patterns, better emotional awareness, and emotional regulation [23]. However, challenges remain around potential drawbacks or adverse effects that "misdiagnosis" and "failure or error in delivering important messages" can have on users [23]. Further emotion recognition on social media, work environments, and daily life emotional tracking [10, 23, 23, 31, 122] has also been perceived as invasive, frightening, and associated with a loss of autonomy or control, suggesting the importance of considering the sensitivity of emotion data. Moreover, Singh et al. [138], Swain et al. [33], and Gao et al. [41] have highlighted the influence that participants' self-reporting can have on the data quality. Prior work has also explored machine learning models for prompting users to annotate emotions as per their physiological data [36]. However, the performance of the pre-trained model remained at a subpar level due to the unavailability of quality data that is representative of real-life settings. The challenges around the usability of mental health tracking and data quality suggest the importance of careful maneuvering for emotion data collection methods. It is crucial for several reasons, as annotations play an important role in overall data quality. In the case of emotion data, it remains further significant due to the absence of a global gold-standard definition of emotions [139, 143]. Thus, in this

work, we aim to explore these gaps in-depth to get a holistic view of stakeholders' perspectives on emotion data collection for AI.

## 3 Methodology

We employed a qualitative research approach, utilizing surveys, interviews, and focus group discussions to gather diverse perspectives from our stakeholders. This study received approval from the Institutional Review Board (IRB) at IIIT-Delhi, India, to ensure ethical compliance. Participation (Survey, Interview, FGD, and Guideline Evaluation) was entirely voluntary, and no compensation was offered to participants. The following subsections provide a detailed explanation of our employed methods.

### 3.1 Survey

To answer our research questions, we surveyed individuals aged 18 and above, including people with or without experience with emotion tracking technologies. Our survey was developed as an exploratory, mixed-methods tool to investigate how users perceive, interpret, and prefer to annotate emotional experiences in their daily lives. It was designed with reference to established emotion theories [15, 16] and user-centered design principles [17, 80], including a mix of quantitative and open-ended questions to capture both structured responses and rich personal narratives (see appendix C for detailed questionnaire). It comprised 27 questions organized into three main sections: **1) Demographic Details**: This section collected basic demographic information about our study participants, **2) Understanding Emotional Awareness**: This section was designed to explore how participants perceive, differentiate, and articulate their emotional experiences and was grounded in the concepts of emotional awareness, emotional vocabulary, and emotional granularity [15, 16]. To assess emotional awareness, participants were asked questions such as, "How often do you take time to reflect on your emotions?", "When experiencing a strong emotion, how easily can you identify what emotion you are feeling?" and "How often do you feel mixed emotions?" Further, they were asked to reflect on which emotions they find easier or harder to identify and to list five positive and/or negative emotions they commonly experience, along with their impact on daily life. These responses helped us understand each participant's emotional vocabulary and how they articulate emotional states. To further assess emotional granularity— the ability to distinguish between similar emotions— participants were asked whether they could tell emotions like sadness and disappointment or anger and frustration apart, and to explain their reasoning. This provided insight into their ability to make fine-grained emotional distinctions linked to more effective emotional regulation and self-awareness. In addition, participants were prompted to describe any recent situations involving strong emotions and to identify the emotions they experienced, to evaluate further their ability to identify and express emotions linguistically [83]. We also included questions to assess the conceptual understanding of important terms like emotions and emotional intensity. To assess conceptual understanding, we included targeted items such as a multiple-choice question asking participants to define "emotion" (e.g., as a bodily sensation, mental state, or response to external events). Further, we also asked them to define "emotional intensity" in their own words. Further, participants were asked about their previous experience with emotion management and use of tools or real-life techniques, such as wearables, emotion-tracking applications, mindfulness practices, and journaling, that assist them in identifying, labeling, and regulating emotions. **3) Attitudes Toward Daily Emotion Annotation**: This section examined how users feel about incorporating emotion annotation into their daily routines. It included questions assessing the willingness to annotate positive and negative emotions, preferred annotation methods, and perceived barriers. Example questions included: "Would you like to annotate your emotions daily?", "What factors are most important to consider when labeling emotions?", and "How easy do you find it to annotate your emotions daily?" Participants were also asked to express their annotation preferences (e.g., emoji, voice input, or descriptive text) and their preferred frequency to annotate emotions daily.

To maintain the quality of our survey responses, we included several consistency checks in our questionnaire, where users were prompted to explain their choices qualitatively for multiple-choice and Likert-type questions. Further, our survey contained 14 open-ended questions, which also enhanced the depth and authenticity of our survey data. Additionally, to validate our survey design for question clarity, logical flow, and completion time, we conducted a pilot with 6 participants before our data collection. Moreover, to evaluate the quality of the responses to our open-ended survey questions, we calculated completion rates to assess participant engagement, distinguishing between required and optional questions. Additionally, we examined word count statistics for each open-ended question, including range, mean, and standard deviation, as proxies for response depth and variation (see Table 14). There were 14 open-ended questions in the survey, eight of which were mandatory. Overall, the completion rate for open-ended questions was high, with an average of 85% for required questions and 82.9% for non-mandatory ones. This indicates strong participant engagement, even when responses were optional. The quality of responses varied across questions, with some eliciting brief answers and others generating in-depth, detailed feedback. Following designing and testing, our survey was distributed digitally using Google Forms. Participants were recruited using convenience sampling methods [145], using social media platforms like WhatsApp and an email call within our institute. Before filling out the survey, we provided our participants with brief information about our study's aim, potential risks, benefits, and confidentiality policy, followed by an informed consent form. Our survey did not collect any identifiable information to maintain anonymity. We got 77 responses to our survey, out of which 75 participants filled out the complete form; the demographic details of our participants are provided in Table 3. All the valid survey responses were exported into Google Sheets for analysis. We analyzed the closed-ended question (n=13) using descriptive analysis, such as calculating percentages, cross-tabulation, and visualizations. For open-ended questions (n=14), we performed thematic analysis [30] and generated codes such as *"Self-reflective practices"*, *"Challenges in Emotion Identification"*, and *"Emotional Literacy"*. Examples of themes we identified were *"Emotional Awareness and Regulation"* and *"Language and Emotional Expression"*.

| Category | Details and Count |
|---|---|
| **Age** | Range: **18 - 41**, Mean = **24.9**, SD = **3.54** |
| **Gender** | Males = **46**, Females = **28**, Prefer not to say = **1** |
| **Education** | Bachelors = **42**, Masters = **21**, Doctorate = **7**, Senior High School = **4**, Vocational Diploma = **1** |
| **Occupation** | Students = **22**, Professionals/Business = **24**, N/A = **29** |
| **Prior Experience** | No Experience = 38, With Experience = 37 (Journaling, Mindfulness, Self-reflection/Introspection, Apps and Wearables) |

Table 3. Summary of Survey Participants' Demographics. Prior experience included details about participants' experience using tools and techniques for emotion tracking or management. Note: Participants mentioned more than one technique, and no experience means people do not actively track or manage emotions in their daily lives.

## 3.2 Interviews

We conducted our formative semi-structured interviews with 32 participants. To guide our semi-structured interviews, we adopted the 5W1H framework [174]. This framework was particularly well-suited for our study since it was an early-stage design study, which aimed at exploring how individuals perceive, approach, and reflect on the act of annotating or self-reporting emotions in their daily lives. As emotion tracking is a deeply personal and context-dependent practice, we needed a method that could surface not only what participants do, but also why and how they do it, within the broader landscape of their routines, motivations, and challenges. Prior to

conducting our participants' interviews, we did pilots with 5 participants to understand the flow of our interview design and the relevance of questions. Our interview design was further guided by prior qualitative research on emotions [8, 100, 101]. Below is an explanation of our interview design: **1) "WHO**- are they?": Focused on understanding participants' emotional awareness, experiences, and familiarity with technology for emotion tracking. Questions explored how they perceive and manage emotions, their prior experience with emotion data collection and logging, and the perceived psychological impact of the process on their lifestyle. **2) "WHAT**- would they annotate?": Examined the types of emotions or emotional events participants considered worth annotating. Questions included privacy concerns and whether they would share detailed information about their emotions. **3) "WHEN**- would they annotate?": Addressed the timing and frequency of emotion annotation. Participants were asked about their preferences for real-time versus retrospective annotation, the contexts or scenarios where they felt annotation was most appropriate, what kind of prompts, and at what frequency they might prefer to be notified for annotating. **4) "WHERE**- would they annotate?": Explored the environments where participants would feel comfortable annotating their emotions, as well as locations they might avoid. **5) "WHY**- would they annotate?": Investigated participants' motivations for annotating emotions, including perceived benefits and potential challenges or barriers. We asked them to discuss the benefits they see in annotating both positive and negative emotions. **6) "HOW**- would they annotate?": Delved into preferred tools and methods for annotation, the time participants were willing to dedicate, and their expectations for simplifying or improving the annotation process. A detailed description of our interview questions is provided in Appendix A.

Our participant pool was well-educated, technology-friendly individuals aged 18 and above, with and without any experience of emotion tracking technology, recruited through convenience sampling [145], using social media platforms like WhatsApp, as well as an email call. We received interest from 32 individuals for the interviews. Before conducting the interviews, we obtained digital consent from each participant through Google Forms sent via email. Along with their consent, we also collected information on their age, gender, education, current occupation, mental health conditions, prior experience with therapy/counseling, and wearables/emotion tracking/emotion data collection studies. Sixteen of our participants had prior experience with either participating in emotion data collection studies or using wearables for stress detection and emotion/mood tracking applications. The rest of our participants did not use technology-based mediums to understand their emotions and mostly relied on techniques such as self-introspection, meditation, exercises, communication with other people, or other mindfulness or coping techniques to deal with emotions. Details about our interview participants are summarized in Table 4.

| Category | Details and Count |
|---|---|
| **Age** | Range: **19 - 43**, Mean = **26.96**, SD = **7.15** |
| **Gender** | Males = **15**, Females = **17** |
| **Education** | High School/Diploma = **3**, Bachelors = **17**, Masters = **6**, Doctorate = **3**, Postdoctoral = **3** |
| **Occupation** | Students = **19**, Professionals/Business = **13** |
| **Attended Therapy** | Yes = **13**, No = **19** |
| **Diagnosis** | Yes = **2**, No = **30** |
| **Prior Experience** | No Experience = **16**, With Experience = **16** |

Table 4. Summary of Interview Participants Demographics. Diagnosis includes details about participants' Mental Health Diagnosis. Prior experience included details about participants' experience of using tools and techniques for emotion/mood tracking and participating in emotion-data collection studies. No experience includes participants who don't actively use any technology to manage their emotions. Note: Participants mentioned more than one technique.

The interviews were conducted in English, either online or offline, based on the participant's preference. Each session was recorded using Zoom Pro, following verbal consent. The interviews began with a brief introduction to the study and familiarization with terms such as "emotions," "emotion annotations," and "emotion intensity" to ensure participants understood the terminology and process of emotion data collection. The definition used to explain emotion annotation to our participant is *"The process of identifying, labeling, and documenting emotional experiences, often to capture emotional data for research, self-reflection, or technological applications. It involves assigning labels (e.g., specific emotions like happiness, anger, or sadness) to emotional events using methods such as written records, mood-tracking apps, emojis, or voice recordings."* Each interview lasted approximately 30 minutes. The sessions were transcribed using Zoom's built-in audio transcription feature. The transcribed documents were then exported to Google Docs and manually reviewed by the first and second authors for grammatical and transcription errors using the original voice recordings. Following transcription, we performed the inductive thematic analysis [26]. We began with authors 1 and 2 reading and re-reading the interview to familiarize themselves with all the data individually. Following this, they individually generated the initial codes for all the data, which included codes like *"Avoidance and denial as coping mechanism," "Understanding of basic emotions"*, and *"Challenges with using static Likert scales"*. Later, authors 1 and 2 grouped similar codes together to form potential themes. Following this, all the authors reviewed the themes together by reviewing the data within each theme to ensure it accurately reflected the data. The high-level themes included *"Emotional literacy and awareness," "Technology and concerns"*, *"Emotional Regulation and management methods"*, and *"Barriers to Identifying and Annotating Emotions"*. All authors reviewed and refined the themes iteratively to ensure they were coherent and distinct until saturation. The themes formulated in the process helped us to structure our findings (see section 4).

| **Category** | **Details and Count** |
|---|---|
| Professional Title | Psychologist = **1**, Clinical Psychologist = **2**, Psychiatrist = **6**, Peer Counselor = **3** |
| Year of Experience | Less than 1 year = **2**, 1-3 years = **4**, 4-7 years = **1**, 8-10 years = **3**, More than 10 years = **2** |
| Experience with AI and Wearable Technology | No = **8**, Yes = **4** |

Table 5. Summary of Focus Group Discussion Participants Demographics.

### 3.3 Focus Group Discussion

To conduct our focus group discussions (FGDs), we utilized purposive sampling [151] to recruit mental health professionals. We reached out to our collaborators, including doctors and NGOs, who helped disseminate our call for participation along with an interest form. From the 18 responses we received, 12 professionals were available for the scheduled time slots. We conducted three separate FGDs: FGD1 included 4 professionals (1 Psychiatrist and 3 Peer Counselors), FGD2 included 3 professionals (2 Clinical Psychologists and 1 Psychiatrist), and FGD3 included 5 professionals (1 Psychologist and 4 Psychiatrists). All participants in the focus groups were from the same country and shared a common cultural background as interview and survey participants, minimizing variability due to cross-cultural differences. Our FGD was designed in line with the prior qualitative studies done with domain experts [8, 20, 127]. Prior to the FGDs, we obtained digital consent from each participant through Google Forms sent via email. In addition to consent, we collected information on their professional titles, years of experience in the mental health field, and familiarity with AI or wearable technology. Details about the participants are summarized in Table 5. All our FGDs were conducted online via Zoom Pro, with a single moderator leading each session following verbal consent to record the meeting. Each FGD began with

a brief introduction of all participants within the discussion, followed by introduction slides presented by the moderator to outline the role of AI in healthcare, the definition of emotion AI, and a brief description of the current practices in AI for emotion data collection and labeling. This overview ensured that all participants had a shared understanding before the discussions began. Following this introduction, the discussion was organized into three main segments: 1) Current Practices for Assessing Emotional States, 2) Attitudes Towards Data and AI, and 3) Challenges and Opportunities in Emotion Data Collection and Recognition. In the first segment, professionals discussed their current methods for assessing the emotional states of patients and clients. The second segment focused on their initial impressions of using AI to understand and monitor emotions, including potential benefits and drawbacks in clinical settings. In the final segment, participants reviewed the current practices of AI data collection, as described in the introduction, and provided their perspectives, recommendations, and insights based on their own practices for the future. A more detailed description of our FGD is provided in the appendix B. Each FGD lasted approximately 1 hour and 15 minutes and was conducted in English. The sessions were transcribed using Zoom's built-in audio transcription feature. The transcribed documents were then exported to Google Docs and manually reviewed by authors for grammatical and transcription errors using the original voice recordings. The first two authors then completed the familiarization, where they thoroughly read all the transcripts. Next, initial codes are generated by reading all the data systematically, highlighting data segments, and assigning brief labels that capture their essence, following inductive thematic analysis [26]. The initial codes included *"Positive attitude towards AI"* and *"Emotional labeling is a mix of subjective and objective labels"*. The authors 1 and 2 searched for themes by grouping similar codes, forming broader patterns. Following this, all the authors jointly reviewed and refined the themes to generate coherent and distinct themes for structuring the findings (see section 4). A few examples of our identified themes are *"Parallel source of information"* and *"Emotional ground truth"*.

### 3.4 Development of Guidelines

To develop our guidelines, we analyzed the data from each source (surveys, interviews, and focus groups) independently to identify recurring themes and patterns. Next, we grouped similar themes and organized them iteratively [26] under the three guideline stages *"Pre-data collection," "During data collection," and "Post-data collection"*, ensuring logical flow and coherence. We then cross-referenced our identified themes with methodologies and recommendations from prior studies to validate and expand our understanding [60, 67, 77, 85, 126, 135]. Finally, we synthesized the insights from participant data and literature to create an end-to-end framework for everyday emotion data collection that is practical, evidence-based, and user-centered. This process resulted in 15 guidelines (named G#) divided into three data-collection stages, as presented in Table 9, 11, and 12. Further, we evaluated our guidelines for their validity in emotion AI research (see section 5).

## 4 Designing *AnnoSense* - An Everyday Emotion Data Collection Framework for AI

This section introduces *AnnoSense*, a framework comprising 15 guidelines designed to support robust emotion data collection in everyday contexts, enabling the development of AI models applicable to real-life scenarios. *AnnoSense* is structured into three phases: pre-data collection, during-data collection, and post-data collection. Within each subsection, we will present findings from our data surveys, interviews, and FGDs to demonstrate the data-driven origins of each guideline. To ensure clarity and ease of navigation, we begin by presenting our data observations, structured into thematic subsections. These are followed by a dedicated subsection—Derived Guidelines—which outlines design guidelines that are directly informed by and grounded in the preceding observations.

| Survey Question | Response | Count | Percentage |
|---|---|---|---|
| How often do you take time to reflect on your emotions? | Never | 2 | 2.7% |
| | Rarely | 11 | 14.7% |
| | **Sometimes** | **27** | **36.0%** |
| | Often | 25 | 33.3% |
| | Always | 10 | 13.3% |
| When experiencing a strong emotion, how easily can you identify the emotion? | Very difficult | 2 | 2.7% |
| | Difficult | 12 | 16.0% |
| | Neutral | 9 | 12.0% |
| | **Easy** | **40** | **53.3%** |
| | Very easy | 12 | 16.0% |
| How often do you feel mixed emotions (experiencing multiple emotions at once)? | Never | 1 | 1.3% |
| | Rarely | 18 | 24.0% |
| | **Sometimes** | **30** | **40.0%** |
| | Often | 24 | 32.0% |
| | Always | 2 | 2.7% |

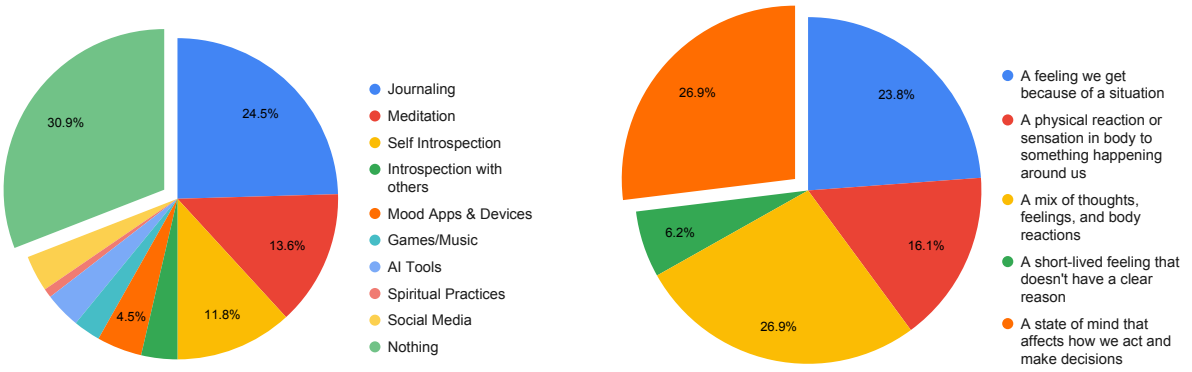
Table 6. Participant responses on emotional awareness and reflection (N = 75)

## 4.1 *"Two-way Communication"*: Pre-Data Collection Phase (G1-G6)

*4.1.1* ***Data Observations:*** To design our pre-study guidelines, we analyzed data from surveys, participant interviews, and focus group discussions to understand participants' specific needs prior to data collection.

**1) Need for Prior Preparation and Training:** Our survey findings indicate that participants generally believed that they possess moderate to high emotional awareness, with 69.3% reporting that they can easily identify their emotions during intense emotional experiences. However, their emotional reflection habits vary considerably, 46.6% reported to engage consistently in self-reflection, while a notable portion rarely or never does (more details in Table 6). Although many participants acknowledge experiencing mixed emotions, suggesting an awareness of emotional complexity, fewer than half use structured methods such as journaling, meditation, or introspection to process these feelings (see Figure 1a). A significant number (34 participants) reported using nothing at all or using suppression or avoidance techniques such as social media and video games, implying that emotional insight for many relies on instinct rather than intentional strategies. This trend was further observed in our interview participants. Participants reported using distraction techniques like watching movies, playing games, or avoiding emotions as a common method for dealing with emotions (mostly negative). As expressed by **(P8, Interview)** - *"I usually sleep. I usually watch television. I usually watch a web series. Nothing else means I can do anything, or I just play some video game. That is the only way of dealing with these emotions, like, sometimes when I'm too stressed"*. Moreover, participants also mentioned not talking about or reflecting on deeper negative emotions due to the stigma of sharing or acknowledging emotions. Participants have used statements like - *"Emotions make me feel weak"*, *"It's better to keep emotions inside"*, or *"Why should we track emotions? It is for people with mental disorders"* suggesting the deep-rooted stigma towards expressing, processing, or tracking emotions.

Furthermore, in our survey, when we asked participants to *recall a recent situation in which they experienced a strong emotion and describe both the context and the emotion identified*, to explore participants' emotional awareness, articulation, and the types of emotional experiences they tend to recall. A majority of participants (60% ) were able to identify specific emotions tied to their experiences. Among these, anger, sadness, and anxiety were the most commonly reported emotions, suggesting negative emotions are commonly recalled by people. Mixed emotions were a notable part of the responses (9%), reflecting the complex nature of human emotions.




(a) Survey results on emotion tracking practices.

(b) Participants' understanding of the definition of emotion.

Fig. 1. Survey results for emotion awareness and management practices among our participants.

Lastly, 21% of participants displayed uncertainty in expressing or identifying their emotions, with some responses showing emotional ambiguity or no clear emotion at all. This reflected the differences in participants' recall behaviors, where a majority of participants recalled negative events or were uncertain about expressing their emotions, possibly due to subconscious stigma or lack of vocabulary or awareness.

Further, in our survey data, we found that a large number of participants are only aware of basic emotions such as happiness (46), sadness (28), joy (26), and anger (27), as shown in Table 7. Interestingly, these primary emotions, such as anger (20), happiness (18), and sadness (13), were also frequently reported as easily identifiable (see Table 8). We also observed that several emotions appeared in both "easy" and "hard" to identify categories such as, anger (20 vs. 10), sadness (13 vs. 8), anxiety (4 vs. 8), and happiness (18 vs. 4). This contradiction suggests the presence of distinct subgroups with varying levels of emotional literacy within our sample. Additionally, our survey data also revealed varying understanding among our participants about what they consider *emotions*, as shown in Figure 1b.

**2) Understanding the Participant Profile:** Extending our investigation into emotional literacy, analysis of our survey question *"Can you differentiate between similar emotions (e.g., sadness vs. disappointment)?"* revealed significant variations in participants' emotional granularity capabilities. Results showed that 44.0% of participants explicitly reported difficulty differentiating between similar emotions, while only 20.9% indicated confidence in their ability to distinguish nuanced emotional states. The remaining 35.2% provided responses that were difficult to categorize definitively. Further, in our data, we observed several recurring themes: 1) sadness was characterized as a broader mood and disappointment as a more targeted emotion, 2) disappointment was frequently framed as a response to unmet expectations, and 3) they were differentiated based on perceived control, emotional intensity, and temporal duration. These findings further reflected the varying emotional abilities among our participants, suggesting that a one-size-fits-all solution to emotion data collection might not be sufficient for collecting quality data. Further, our discussions with experts also reconfirmed the *varying emotional literacy* as explained by an expert **(P1, Psychologist, FGD3)**, *"People tend to feel only 4-5 basic emotions and lack a vocabulary to explain their emotions and must be taught...A therapist tries to teach people about emotional awareness to improve vocabulary as it helps them identify emotions more clearly, along with their professional methods."* To overcome these challenges, domain experts within our FGDs emphasized efficient history-taking to understand emotion data reliably. Further experts also emphasized the importance of assessing various emotional aspects such as *emotional vocabulary*, *emotional range* (the spectrum of emotions a person can experience, express, and recognize), *emotional congruency*

| Positive Emotion | Frequency | Negative Emotion | Frequency |
|---|---|---|---|
| Happy | 46 | Sadness | 28 |
| Gratitude | 38 | Anger | 27 |
| Joy | 26 | Anxiety | 16 |
| Hope | 21 | Frustration | 10 |
| Love | 18 | **Motivation**** | 9 |
| Peace | 16 | Fear | 9 |
| Excitement | 14 | Loneliness | 8 |
| Satisfaction | 11 | Guilt | 6 |
| Confidence | 10 | Jealousy | 6 |
| Motivation | 9 | Stress | 4 |
| Calm | 9 | Irritation | 4 |

Table 7. Frequency of Top 10 Positive and Negative Emotions in Daily Life as Reported in our Survey. For positive emotions, the mean frequency of responses was 3.88 with a standard deviation of 1.8, while for negative emotions, the mean frequency was 2.48 with a standard deviation of 1.9. ** Note: **Motivation** is contextually a positive emotion but was mentioned in the negative list—possibly reflecting low or lack of motivation.

| Emotions Easy to Identify | | Emotions Hard to Identify | |
|---|---|---|---|
| **Emotion** | **Count** | **Emotion** | **Count** |
| Anger | 20 | Anger | 10 |
| Happiness | 18 | Sadness | 8 |
| Sadness | 13 | Anxiety | 8 |
| Frustration | 6 | Satisfaction | 4 |
| Joy | 6 | Fear | 4 |
| Anxiety | 4 | Happiness | 4 |
| Love | 4 | Depression | 4 |
| Loneliness | 3 | Positive | 3 |
| Disappointment | 3 | Jealousy | 3 |
| Hope | 3 | Guilt | 3 |

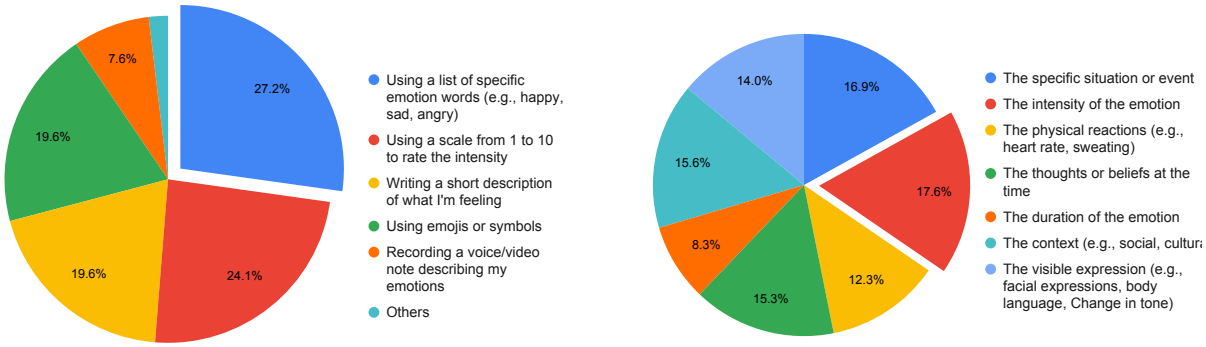
Table 8. Comparison of top 10 emotions based on ease of identification as per our survey response.

(the consistency between inner feelings and outward expressions), *emotional intensity* (degree of an emotional experience), and *emotional reactivity* (the intensity and speed of an individual's emotional response to a stimulus) to understand emotional data better. Finally, experts highlighted the importance of screening for conditions like alexithymia, which affects an individual's ability to identify and describe emotions. Finally, our participants' data and experts' discussions also revealed that emotions are deeply personal, and participants will find it challenging to share emotional details without assurance of privacy.

*4.1.2* ***Derived Guidelines:*** As reflected in our data, there were differences in participants' emotional awareness and attitude towards emotion management. This inspired our guidelines **G3, G4** on participant training and initial calibrations to ensure data collection methods are accessible and relevant to a diverse population. This is further crucial for collecting richer data. Discussions with domain experts further reinforced the need for participant training, and prior research has also shown that the varying ability in identifying and articulating emotions can

| Guideline | Description |
|---|---|
| **G1** | **Selecting Participants**:<br>**1.** Clearly document the inclusion and exclusion criteria based on the study's objective and data requirements.<br>**2.** Recruit individuals from diverse demographic groups (age, gender, culture, education, and occupation) in line with the study's inclusion criteria and data-requirements.<br>**3.** Screen participants for alexithymia (difficulty identifying and expressing emotions) using standardized screening tools such as the Toronto Alexithymia Scale [48] or Perth Alexithymia Questionnaire [115], neurological disorders (e.g., cognitive impairments), and health conditions (e.g., cardiovascular issues or chronic illnesses) that might impact physiological signals, ability to identify and express emotions, and in line with the study's exclusion criteria. |
| **G2** | **Obtaining Informed Consent**:<br>**1.** Clearly explain the purpose, benefits, potential risks, compensation, voluntary participation, time commitment, and key procedures of the study in simple, accessible language. Provide enough information to the participants without revealing details that could compromise the integrity of the study design.<br>**2.** Clearly outline ethical approval and privacy measures, such as how participant data will be anonymized (e.g., removal of personal identifiers), compliance with relevant data protection laws (e.g., GDPR, HIPAA), secure data storage practices, and data sharing in the consent document. |
| **G3** | **Conduct Initial Calibration**:<br>**1.** Conduct a baseline session or calibration trials (as per study requirements and resources) to familiarize participants with the devices being used in the study.<br>**2.** Provide clear instructions on how to correctly wear the devices and ensure they are functioning accurately during data collection. |
| **G4** | **Provide Participant Training**:<br>**1.** Organize practice sessions where participants label their emotions (e.g., joy, anger, sadness), identify subtle distinctions (e.g., frustration vs. irritation), and document contextual factors such as environment, social interactions, and cultural norms in real-time or respond to controlled stimuli, enabling researchers to clarify doubts and improve annotation accuracy.<br>**2.** Educate participants about the broader impacts of emotion annotations and the data privacy measures in place to build initial trust and engagement.<br>**3.** Offer expert-verified resources, including video clips, audio recordings, or books, to improve participants' emotional literacy and understanding of what is meant by emotion annotations. |
| **G5** | **Perform Detailed Psycho-Social Profiling of Participants**:<br>**1.** Collaborate with domain experts to collect detailed histories on emotional characteristics such as emotional range (the spectrum of emotions a person can experience, express, and recognize), emotional congruency (the consistency between inner feelings and outward expressions), emotional intensity (degree of an emotional experience), and emotional reactivity (the intensity and speed of an individual's emotional response to a stimulus), and emotional vocabulary.<br>**2.** Collect contextual details such as past traumatic experiences, daily routines, work-life balance, family dynamics, emotional awareness, regulation habits, and potential stigma using standardized questionnaires or with the assistance of domain experts. |
| **G6** | **Collect Comprehensive Demographic and Medical Data**:<br>**1.** Gather detailed demographic information (e.g., age, gender, education, socio-economic status, personality traits, and medical information) based on the specific needs of your research questions.<br>**2.** Ensure that demographic data is relevant to the study objectives, is ethically approved, and captures any additional factors that may influence emotional responses. |

Table 9. Guidelines for Pre-Data Collection Phase




(a) Preference of annotation method.

(b) Factors that can impact annotations.

Fig. 2. Survey results for attitude towards emotion annotation in daily life.

adversely impact the quality of emotion data [41, 42]. Further to mitigate these impacts, we have added guidelines **G5, G6**. These guidelines are necessary for collecting detailed additional information to effectively contextualize emotion data [101]. Further, keeping in mind the wide range of hypotheses that inspire emotion data collection, we have added **G1**, which also includes exclusion-inclusion and screening guidelines inspired by data-centric AI and prior emotion data collection [8, 59, 126]. Diversity, necessary screening, alongside a comprehensive understanding of factors influencing emotional data, is crucial for ensuring the reliability of emotion assessments. Lastly, as expressed **(P3, Psychiatrist, FGD2)**, *"...the issue will be regarding the privacy part, how the data is being stored by the AI ... And you know who has access to it and how it is being used by the 3rd party. So overall, there are a lot of privacy-related challenges because there will be a lot of sensitive information. How we are tackling this will be an important point. And it should be communicated early on."*, we have added **G2**. Elaborate informed consent was necessary alongside training and contextualization because stigma and privacy concerns, as visible in our data, can deeply influence the self-reporting behaviors. Consequently, together with our insights from data, alongside prior practices to collect quality data and emotion data collection methodologies, have informed our pre-dataset collection guidelines as provided in Table 9.

## 4.2 Understanding the needs of "*Data Source*": During Data Collection Phase (G7-G11)

*4.2.1* ***Data Observations:*** Within everyday settings, participants are typically prompted to annotate their emotions based on random, fixed time, or event-based triggers in response to changes in physiological or activity data [60, 61, 135]. These prompts often ask participants to fill out surveys or questionnaires based on pre-defined scales (as discussed in section 2.2). However, these predefined surveys offer limited flexibility for participants to share additional context or express emotions as per their intensities.

**1) Need for Adaptable Design**: In contrast to these rigid methods, our survey data highlighted the diverse preferences participants have when it comes to emotion annotation (see Figure 2a). While 27.2% of participants preferred using an emotion list, 19.6% expressed a desire for an open-ended option to write about their emotions. Further, on examining participants' motivations behind their preferred annotation methods. *Ease of use* emerged as the primary consideration (14.8%), closely followed by *expressiveness* (12.6%) - the ability to fully convey emotional experiences. Participants also mentioned *clarity* (8.9% ) of methods and their ability to capture *emotional complexity* (8.9% ) as important factors. These findings suggest that participants are seeking annotation methods that balance accessibility with expressiveness, allowing them to capture nuanced emotional states. Moreover, the relatively

even distribution across annotation preferences points to significant individual variation. Further, in our interview data, we found similar patterns that highlighted the need for annotation methods that consider the transient nature of emotions. As one participant **(P30 - Interview)** explained- *"I would say the objective scales (likert, SAM or PANAS) will be easier use daily, but you should always give an option that if I am feeling extreme emotions — say if I'm extremely happy, extremely sad, or extremely angry —then there should be an option to write down or something."* Additionally, participants noted that during intense emotional moments, writing about the situation or their reactions would be easier rather than trying to identify and label specific emotions.

This finding suggested that participants may struggle to articulate intense emotional experiences, highlighting the need for structured guidance to help them navigate and understand complex emotional states. Mental health experts reinforced this insight, recommending adaptable annotation methods modeled after diary writing approaches. They specifically suggested incorporating probing questions about triggers, situational contexts, and emotional reactions. Such structured frameworks can significantly reduce the cognitive burden of subjective emotion annotation, particularly for individuals experiencing complex or overwhelming emotional states. Further, our interview participants noted that emotions with visible cues are easier to identify. However, identifying and articulating complex emotions (such as co-occurring, mixed, layered, or ambiguous emotions) is challenging. These emotions—like anxiety combined with fear or frustration intertwined with anger—were described as harder to pinpoint. These insights highlight the importance of designing interfaces that can facilitate the expression of both simple and complex emotional experiences. Such an interface should provide participants with the required support, emotional vocabulary hints, reflective prompts, guided questions, and relatable analogies.

A few participants also suggested using more abstract and expressive methods for annotating emotions. They felt that predefined scales or specific words often limited how they could express their feelings. Instead, they proposed alternatives like sharing the songs they were listening to, quotes that reflected their mood, or photos of their environment. Some also mentioned sketches or free-form journaling. These methods, as explained by participants, allowed for a more personal and authentic expression of emotions, reflecting the need for adaptable design. In addition to this, our survey data also revealed varying factors that can influence the identification of emotions (see Figure 2b). The intensity of emotional experience as felt by a participant emerged as the most frequently mentioned factor (53 mentions), closely followed by the specific situational context triggering the emotion (51 mentions). Contextual elements, including social or cultural factors, thoughts present during emotional episodes, physical sensations, and visible expressions (facial expressions, body language, vocal changes) are also mentioned as crucial. This even distribution of factors further reinforced that participants recognize emotion as a multifaceted phenomenon requiring multidimensional annotation approaches.

**2) Participant's Agency, Learning and Participant-Aware Sampling:** Our analysis of emotion annotation preferences and practices survey data (see Table 10) revealed significant resistance to daily emotion tracking, with 62.7% of respondents indicating reluctance compared to 37.3% expressing interest. This reluctance corresponds with perceived difficulty, as 36.0% found emotion annotation difficult, while only 25.3% considered it easy. Frequency data further reinforced these patterns, with only 35.9% of participants willing to annotate emotions daily, while 37.3% preferred weekly or less frequently. Further analysis of open-ended data and interviews revealed that participants wanted an annotation method that would prompt them according to their emotional intensities and pace. They also mentioned that the method should provide them feedback, insights, and an opportunity to learn from their data. Further, they mentioned that methods that only collect data without any learning engagements might not motivate them to annotate frequently. Participants also highlighted the importance of well-timed prompting methods per their personalized schedules. As mentioned by a participant (P26 - Interview) who uses an emotion logging application, the app frequently sends notifications when he begins working, distracting him. As a result, although he is willing to use the tracking technology, but he often does not annotate.

This suggests that the varying needs of participants and assumptions, such as prompting users when they are not in motion, might not hold for everyone. For instance, while some may find such prompts helpful during idle

| Question | Response | Count |
|---|---|---|
| Would you like to annotate your emotions daily? | Yes | 28 |
| | **No** | **47** |
| How easy do you find it to annotate your emotions daily? | Very difficult | 5 |
| | Difficult | 22 |
| | **Neutral** | **29** |
| | Easy | 16 |
| | Very easy | 3 |
| How frequently can you annotate your emotions? | Multiple times a day | 17 |
| | Once a day | 10 |
| | **Few times a week** | **20** |
| | Once a week | 11 |
| | Less than once a week | 9 |
| | Never | 8 |
| If you are going through a negative emotion, will you annotate? | **Yes** | **25** |
| | No | 14 |
| | Not Answered | 36 |
| If you are going through a positive emotion, will you annotate? | **Yes** | **21** |
| | No | 20 |
| | Not Answered | 34 |
| Do cultural or societal factors influence how you perceive emotions? | **Yes** | **44** |
| | No | 31 |

Table 10. Survey Results on Emotion Annotation Practices and Perceptions

moments, others—like P26—may perceive them as intrusive, especially when they coincide with focused work sessions. Our interviews further highlighted that the timing and context of annotation must align with users' emotional states and willingness to engage alongside other contextual data such as activity levels, behavioral cues, and physiological changes. Further, our data also revealed that many participants preferred non-digital alternatives, viewing digital tools as requiring extra effort and time. Participants highlighted the availability of real-life alternatives (such as writing with pen and paper, sketching, sitting in silence, playing sports, or talking to friends) as the reason behind their preferences. This highlights the importance of participant-aware interventions incorporating user-agency in design and adapting to individual routines and preferences, rather than relying on one-size-fits-all strategies. Further, our survey data also revealed a significant component of cultural and societal influence on emotions. On deeper analysis, we found that these influences are shaped by negative connotations about sharing or expressing emotions, or stigma, and can hinder unbiased and balanced annotations. This suggested a need for an emotional literacy component in data collection methods.

**3) Multi-perspective Assessments:** Finally, our discussions with experts emphasized the importance of collecting emotional data from multiple sources, specifically for people with mental disorders or significant life events. They recommended combining self-reports with family members' input and regular evaluations by psychologists or psychiatrists. Emotional assessment is complex —even for professionals— so relying on a single

| Guideline | Description |
|---|---|
| **G7** | **Focus on Participant's Agency**:<br>**1.** Use lightweight, non-intrusive wearable devices to avoid disrupting daily activities.<br>**2.** Allow users to adjust the annotation frequency based on their preferences or schedules while ensuring a minimum frequency that balances data accuracy with preventing fatigue and disengagement.<br>**3.** Set realistic expectations for emotional changes as per the research objective, for example conditions like depression don't show significant daily fluctuations, so daily recordings may not be necessary. |
| **G8** | **Develop Participant-Aware Sampling**:<br>**1.** Trigger annotations by corroborating information on participants' characteristics (gathered in G5) such as, daily schedules, activity levels, physiological changes, and emotional profile while keeping G7 and research objectives in mind. |
| **G9** | **Design Adaptable Annotation Methods**:<br>**1.** Offer participants the choice between structured annotation methods (e.g., SAM, PANAS) that use scales and unstructured subjective annotation methods (e.g., text, audio, images) as per their emotional intensity.<br>**2.** For subjective annotations, adopt structured frameworks like the ABC model (Activating Event, Belief, and Consequence) to guide participants' responses. Alternatively, design LLM-based prompts [101, 170] customized to align with participants' unique emotional traits, as identified in steps G5, to provide tailored guidance.<br>**3.** Provide participants with support in understanding complex emotions by offering tools such as emotion vocabulary lists, options to select multiple emotions simultaneously, visual aids like emotion wheels, reflective prompts, guided questions, and relatable scenarios or activity list to foster emotional clarity. |
| **G10** | **Incorporate Multi-Perspective Assessments**:<br>**1.** Collect assessments not only from the participants themselves but also from trusted individuals in their support system, such as family members, peers, or mental health professionals, based on the participant cohort and study requirements. For example, clinical populations may require multiple assessments, whereas healthy individuals might need fewer.<br>**2.** Integrate additional data streams, such as location, social media activity, phone usage, sleep patterns, and calendar events.<br>**3.** Allow participants to select who and what data streams can contribute to their data based on their comfort and preferences. |
| **G11** | **Focus on Participant Engagement, Learning and Support**:<br>**1.** Periodically reach out to participants to address any concerns, clarify expectations, motivate, and support.<br>**2.** Use UI designs and prompts to encourage reflection, show growth, and provide supportive feedback to maintain engagement.<br>**3.** Integrate interventions within the study that help participants enhance their emotional literacy over time.<br>**4.** Integrate prompts that encourage reflection on positive outcomes or gratitude to offset the potential negative impact of recording difficult emotions. Additionally, provide access to mental health resources or emotional support tools for participants who may experience distress from self-reporting. |

Table 11. Guidelines for During Data Collection Phase

source may lead to unreliable results. Experts also highlighted the value of integrating additional data streams. These included ecological activity data, social media behaviors, physiological signals, and other automated, objective measures. These sources can help complement and contextualize subjective self-reports. Experts also stressed the need for emotion assessment methods tailored to different groups. For the general population, tools should focus on promoting wellness and addressing everyday stressors. In contrast, more in-depth emotional assessments and professional evaluations are critical for clinical populations or those facing significant life events to ensure accurate and meaningful insights. Further, our interview data also highlighted a set of participants who were skeptical about using technology for managing emotions and emphasized the need for a human touch. This finding reinforced the importance of incorporating multi-source assessment approaches, involvement of trusted people, and thoughtful data sharing mechanisms into emotion data collection methodologies. By integrating these elements, emotion tracking systems can complement rather than replace human interaction.

*4.2.2* ***Derived Guidelines:*** Our analysis revealed a preference for a balanced approach to emotion annotation. Participants valued having the flexibility to choose between structured, scale-based methods and unstructured, journal-writing methods based on the intensity of their emotions. The suggested need for this flexibility in our data collection approaches guided our addition of guideline **G9**. Within our data, it was also evident that participants frequently linked their emotions to specific environmental or situational cues. For example, loneliness was associated with the absence of companionship, stress with workload, fear with significant life events, and joy with time spent with loved ones. This underscores the need for tools that allow participants to articulate emotions by connecting them to contextual factors **(G9.3)**. Further, the need for user-agency to personalize the prompts per their schedules and emotional spectrum is also highlighted. Thus, designing methods with interfaces that could balance user-agency and participant-burden with data needs would be essential, guiding our inclusion of **G7**. Our data also highlighted a need to move beyond the context-aware sampling [95, 100], and adding a layer of participants' persona, cultural knowledge [158] to sampling strategies [101], as included in guideline **G8**. In addition to it, our discussion with experts and participants' interviews highlighted a need for adding multiple-perspective assessments and the option to multi-source data contribution [154] for improving the data quality. This supported our addition of guideline **G10**. Finally, our data observations suggested a need for better participant support and components for improving emotion literacy over time in our data collection strategies for better participant engagement, guiding the inclusion of **G11**. Our detailed during data collection guidelines are presented in Table 11.

## 4.3 Learning from *Dynamic* Data: Post-Data Collection Phase (G12-G15)

*4.3.1* ***Data Observations:*** Following data collection, the post-processing stage involves several critical steps to ensure data usability and integrity. Our data highlighted several observations for the post-data collection stage.

**1) Consistent Best Practices:** The post-processing stage typically includes quality checks, consistent structuring, and preparation for data sharing to enable reproducibility and collaborative research [59]. While prior work highlights these best practices, data sharing remains inconsistent. For example, datasets such as WESAD [129], EEVR [137], and ASCERTAIN [146] provide not only the data but also baseline experiments to support emotion recognition research. In contrast, datasets like EMOGNITION [125] and GReX [22] focus solely on data release with quality checks, without offering baseline evaluations. While valuable, the absence of standardized benchmarks increases friction for downstream use and hinders fair comparisons across studies.

**2) Handling Dynamic Data:** Further analysis, as discussed in Section 4.2, revealed that participants have diverse needs and preferences regarding the sharing of their emotion data. Recognizing and addressing these needs is critical for fostering participant engagement and trust. Incorporating such preferences into data collection practices can enrich the resulting datasets, enabling the integration of information from multiple sources and annotation methods, including both structured and unstructured formats. Effectively managing this dynamic

emotion data requires the establishment of a standardized data pipeline. This pipeline should include procedures for identifying and handling missing values, inconsistencies, and artifacts that could compromise data quality. Given the heterogeneous nature of emotion data, validation must also extend to the annotation layer. This involves assessing the reliability of labels through cross-validation across various sources—such as physiological signals, self-reports, behavioral observations, and expert annotations where applicable. Such practices enhance the robustness and accuracy of the dataset, ultimately offering a more comprehensive and trustworthy representation of participants' emotional experiences. Normalization is another critical pre-processing step, particularly because individual differences, such as physiological signal ranges, emotional reactivity, or even environmental factors like temperature, can significantly influence emotional data. It is also important to document when and why normalization is applied to ensure transparency in the data processing steps. Normalization is crucial because it helps reduce bias and variation that could lead to inaccurate conclusions.

**3) Grounding Emotion data:** Our participant data highlighted that adopting more dynamic annotation approaches during data collection will likely produce annotations that differ from the standardized, fixed-scale labels typically used. These annotations will be more nuanced, context-dependent, and reflect participants' real-time emotional experiences. Domain experts supported this view. They recommended using both structured and unstructured data when labeling emotions. They emphasized that relying on just one type of data could miss essential nuances. When discussing emotional "ground truth," experts pointed out that exact accuracy is less critical than generating actionable insights. They stressed that aligning different data sources is a key indicator of accurate emotion labeling. Although these dynamic annotation approaches will capture richer and more authentic emotional experiences, they pose challenges for applying traditional AI models. This highlights the need to design newer systematic approaches to ground the emotion data.

**4) Secure Data Handling:** As discussed in Section 4.1, our participants have expressed a preference for keeping their emotions private or sharing them only with trusted individuals, such as family members or mental health professionals. Sharing emotional information with others or through technology was not the first choice for many participants unless it significantly impacted their lives. Moreover, our survey (see Table 10), 62.7% of participants expressed reluctance to track their emotions using digital tools. Alongside time constraints and concerns about overthinking, emotional privacy emerged as a key reason for avoiding such tools. This highlights a strong stigma around tracking or sharing emotional data, as expressed by **(P21, Interview)**: *"If I am in real distress and I really want to get myself treated or understand the depth of my emotions, then I might provide access to my journal.* This suggests the need to maintain data security post-data collection.

*4.3.2* ***Derived Guidelines:*** To address the privacy concerns, it is crucial that data collectors ensure participants feel confident in the secure handling of their data. Participants must also have the option to review or request deletion of their data, as outlined in **G12**. This guideline is essential because ensuring participants' trust is the foundation of ethical data collection, particularly when dealing with sensitive emotional information. By offering data review and deletion options, we respect participants' autonomy and privacy, addressing stigma and data misuse concerns. Subsequently, after data collection, validation of data quality [59] is an essential step, guiding the addition of **G13**. This includes support for heterogeneous data formats, standardized metadata schemas, and robust pre-processing pipelines that handle noise, missing values, and temporal inconsistencies. Further, the need for contextually grounding the collected dynamic data informed our guideline **G14**, which emphasizes the importance of holistically analyzing and grounding data. This guideline ensures that emotion labels reflect the complexity of participants' lived experiences rather than oversimplifying them for AI processing. Grounding the data involves assessing the reliability and relevance of its sources.

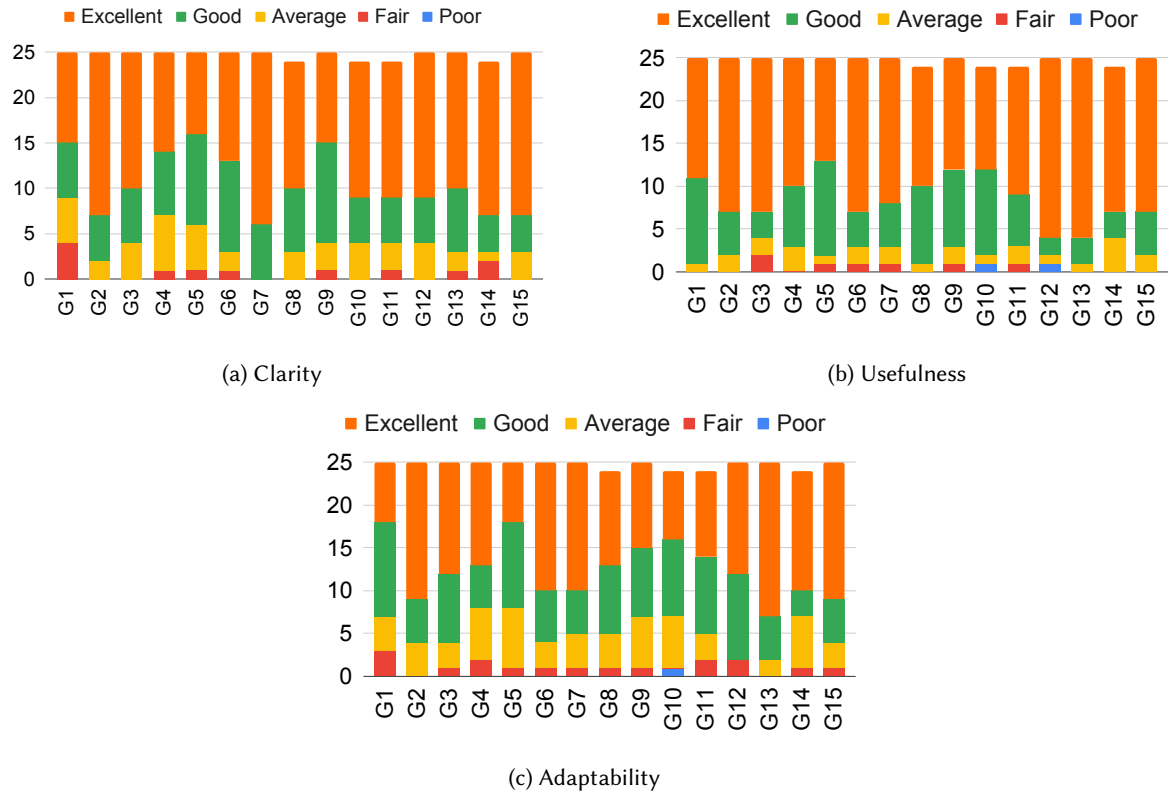
Recognizing psychosocial individual differences through standardized tools [75, 102] or through expert interpretations to contextualize emotional labels more accurately. Additionally, combining unstructured subjective responses (by quantifying them using either psycholinguistic analysis [132] or expert feedback) with structured,

| Guideline | Description |
|---|---|
| **G12** | **Secure Data Handling**:<br>**1.** Store data securely using encryption and anonymization techniques, adhering to ethical guidelines (e.g., GDPR, HIPAA).<br>**2.** Allow participants to request data reviews within a specified timeframe (e.g., 30 days), with researchers providing an overview instead of direct access to raw data to avoid misinterpretation and better confidentiality. Authorized researchers should handle deletion to maintain data security if deletion is requested.<br>**3.** Clearly communicate any limitations on data review or deletion, such as once data has been anonymized or aggregated for analysis. |
| **G13** | **Data Quality Validation and Pre-processing**:<br>**1.** Review datasets for missing values, artifacts, or inconsistencies. Depending on the study's goals and the extent of missing values, methods such as imputation, removal, or flagging of problematic data can be used to handle these issues.<br>**2.** Cross-validate multiple data-sources (if available, G9 and G10) to improve reliability.<br>**3.** Normalize data where individual differences (e.g., physiological ranges, personality traits) or environmental factors (e.g., time of day, activity) significantly affect results. Document when normalization is applied and why. |
| **G14** | **Holistically Analyzing and Grounding the Data**:<br>**1.** Combine qualitative insights (e.g., text-based descriptions) and quantitative data (e.g., scale-based measures) to create multi-dimensional emotion labels that accurately capture the emotional experience within its context.<br>**2.** Ground data on the reliability and relevance of the source (if G10 is applicable), such as expert assessments for emotional dysregulation, peer feedback for social interactions, and self-reports for subjective experiences.<br>**3.** Combine psychosocial details (e.g., emotional traits, past experiences, daily routines) with emotion data to create a context-rich foundation for analysis and labeling, and document how these psychosocial factors impact emotion labeling to enhance transparency.<br>**4.** Collaborate with domain experts to review and ensure the accuracy and consistency of grounded emotion labels. |
| **G15** | **Share Findings, Best Practices, Data Limitations and Usability**:<br>**1.** Present key findings and any challenges faced during data collection, such as participant engagement issues, device inaccuracies, or contextual variability. Describe the study protocol in detail to ensure reproducibility.<br>**2.** Highlight data limitations such as device reliability, data quality concerns, participant biases, or issues with ecological validity.<br>**3.** Specify the intended AI applications for the data, like emotion detection, disorder diagnosis, or longitudinal tracking of emotional changes. Then, evaluate the data's suitability for each of these specific use cases. |

Table 12. Guidelines for Post-Data Collection Phase

scale-based data provides a more nuanced and actionable grounding approach. This ensures that emotion annotations are participant-centered, addressing individual emotional experiences while also enabling the extraction of meaningful and useful emotion labels. Lastly, **G15** outlines the necessity of transparently presenting key findings alongside any challenges encountered during data collection, such as participant engagement issues,

device inaccuracies, or contextual variability. By detailing these challenges, researchers offer transparency into the reliability and scope of the data, which is critical for ensuring the reproducibility and validity of the findings. Additionally, specifying the intended applications for the collected data—whether for emotion detection, disorder diagnosis, or tracking emotional changes over time—is crucial for guiding its use. Further, benchmarking and evaluating the data's suitability for the downstream task is equally essential for the future applicability of the dataset. Finally, our guidelines for the post-data collection phase are presented in Table 12.




(a) Clarity

(b) Usefulness

(c) Adaptability

Fig. 3. The evaluator's scores for our guidelines.

## 5 Evaluation of Guidelines

The 15 guidelines were first internally evaluated by all the authors and fellow researchers/colleagues for clarity, usefulness, and adaptability. Following internal evaluations, we conducted an external evaluation involving 25 expert evaluators. These evaluators possessed expertise in emotion AI, physiological data collection, affective and ubiquitous computing research (detailed demographics presented in Table 13). This evaluation aimed to gather expert feedback on the clarity, usefulness, and adaptability of our guidelines. We employed purposive sampling [107] to select these experts, who were then contacted via email and social media platforms such as WhatsApp, Twitter, and LinkedIn. Each expert received a survey form encompassing informed consent, demographic questions (area of expertise and years of experience), and the 15 guidelines themselves, organized into three sections: *"Pre-data collection," "During data collection,"* and *"Post-data collection"*. For each guideline, we

asked the experts to provide their ratings for clarity (description and communication), usefulness (practicality and goal achievement), and adaptability (real-world applicability across diverse contexts) on a 5-point Likert scale (1=poor to 5=excellent). Additionally, we asked the experts to provide qualitative feedback in the form of comments/suggestions for further refining the guidelines. Our method draws inspiration from Amershi et al.'s modified heuristic evaluation [9], where we adapted the principles of discount usability testing to evaluate our guidelines. Furthermore, to refine the guidelines, we conducted a descriptive analysis [84] of evaluator ratings as illustrated in figure 3.

Descriptive analysis of the expert evaluations reveals a strongly positive overall reception of the guidelines across all assessed criteria - clarity, usefulness, and adaptability. The guidelines were consistently rated highly for their clarity and usefulness, with the vast majority of evaluators rating them as "Good" or "Excellent" in these domains. While Adaptability also received predominantly positive ratings, a slightly higher proportion of "Average" and "Fair" scores in this category suggests a potential for refinements to enhance their perceived applicability across diverse research contexts. Crucially, the consistent absence of "Poor" ratings across all guidelines and criteria indicated a robust framework without major perceived weaknesses. In addition to descriptive analysis, open-ended feedback was subjected to inductive thematic analysis [26], performed by the first author, to identify key suggestions for improvement. These suggestions, derived from expert feedback, primarily focused on enhancing clarity and comprehensiveness. Evaluators recommended adding more detailed explanations, illustrative examples, and definitions to make the guidelines more accessible. Furthermore, they emphasized the need to acknowledge the context-dependent nature of the guidelines, noting that their application may vary based on specific research objectives. In response to this feedback, we iteratively revised and rephrased the guidelines where needed to increase their adaptability to a wider everyday emotion research context. For example, Guidelines *#G1.1* and *#G1.2* were refined to explicitly state the importance of diverse recruitment while remaining aligned with specific study objectives. For *#G1.3*, to improve accessibility for interdisciplinary audiences, we incorporated references to screening tools like the Toronto Alexithymia Scale and added a definition of alexithymia. Similarly, Guidelines *#G3*, *#G4*, and *#G5* were revised to include examples and definitions, enhancing their overall clarity and broadening their applicability. Finally, all the authors then revisited and finalized the guidelines internally as presented in table 9, 11, and 12.

| **Category** | **Details and Count** |
|---|---|
| Gender | Male = **13**, Female = **12** |
| Year of Experience | 0-5 years = **12**, 5-10 years = **8**, 10+ years = **5** |
| Role | Researcher (Emotion AI/ Affective Computing/ HCI) = **23**, Data Scientist (Emotion AI) = **1** Researcher (Ubiquitous Computing/AI) = **1** |

Table 13. Summary of Guidelines Evaluators.

## 6 Discussion

This section discusses how future research can leverage *AnnoSense* framework for designing participant-centric methodologies. First, we will discuss how to prototype tools based on our guidelines in section 6.1. Then, we will discuss the implications of pre, during, and post-data collection guidelines for future work.

### 6.1 Implementing *AnnoSense*: Designing for Participants

Building on the *AnnoSense* framework, in this section, we envision potential directions for prototyping new tools that can further support both real-life emotion data collection and the advancement of wearable and mobile-based

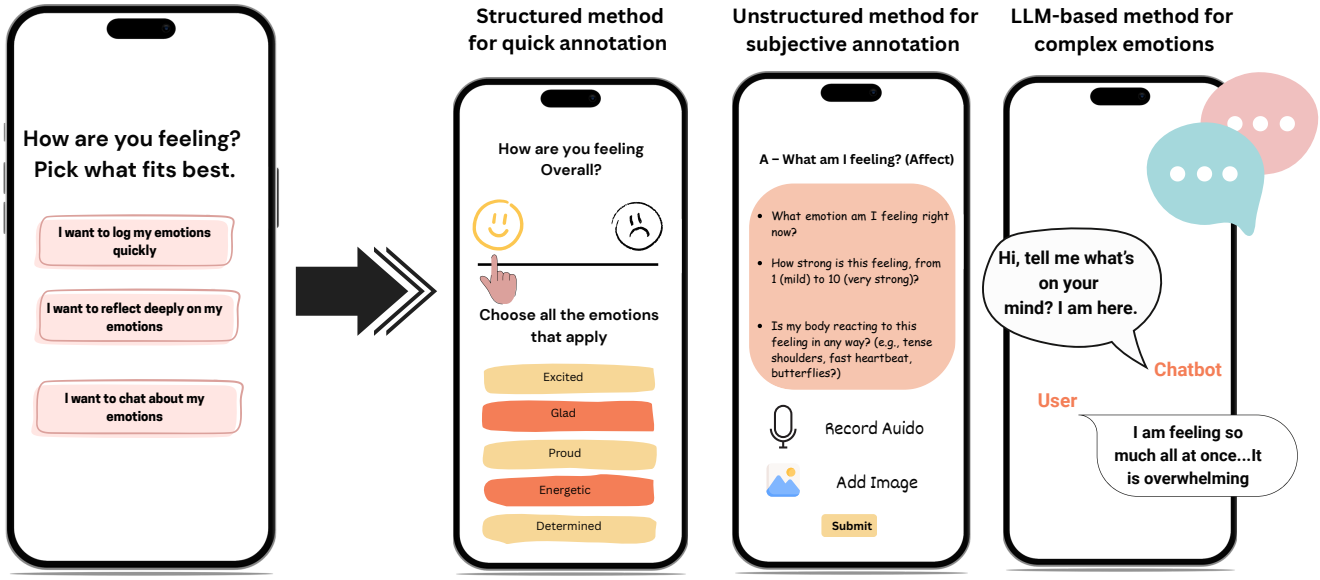



Fig. 4. Visualization of the Participant-Centric Adaptable Annotation Approach. This approach offers flexible annotation options tailored to participants' emotional intensity and time availability. For quick annotations, a structured method using predefined emotion scales and lists is provided. In intense emotional experiences, participants can opt for a subjective, open-ended annotation guided by reflective questions. Additionally, large language model (LLM)-based support can facilitate meaningful annotation for users with lower emotional literacy.

AI solutions. To inform the design of potential prototypes, we reviewed the designs of currently available mobile applications and wearable technologies that support mood tracking, mental health monitoring, and emotion-related interventions. This included mood and journaling apps such as MoodPrism [118], MetricWire [92], Daylio [2], and MindLamp 2 [155]; well-being and mindfulness platforms like Headspace [3], Happify [47, 55], and Calm [1]; as well as wearable ecosystems including Apple Health [12], Samsung Health [128], Fitbit [40], Oura Ring [105], and WHOOP [168]. We also examined AI-supported mental health applications such as Wysa [5] and Woebot [4]. Additionally, we also reviewed well-know EMA frameworks like MindLamp [155], Beiwe [104], AWARE [39], PACO [46], Sensingkit [63], mEMA [56], Experiencesampler [148], and MobileQ [89]. Our review identified several opportunities for designing future emotion annotation tools. Based on our review and the AnnoSense guidelines, we propose a set of prototype interfaces to accommodate users' needs.

**1) Adaptable Annotation Interface:** We propose a prototype for adaptive annotation interfaces that provides users with an opportunity to select an annotation method according to their emotional intensity or available time, as illustrated in Figure 4. This approach, in contrast to traditional ESM methods, provides users with an option to select between quick scale-based annotations, detailed subjective annotations [69, 71, 82, 101], and chatbot-supported approaches [100, 101, 170]. It also offers users appropriate guidance through curated emotion lists based on their selected overall feelings during quick annotations. Diary-inspired reflection prompts based on psychological frameworks, such as the ABC model (A - Activating Event, B - Beliefs, C- Consequence) [86] or Cognitive Behavioral Therapy (CBT) thought record model (Triggering event, Automatic thoughts, Emotions, Evidence supporting, and Evidence against) [70, 82, 100, 167]. And an empathetic chatbot interface to support emotion annotations. Additionally, it supports a diverse range of users by offering options to record audio and upload images as part of their emotion annotations. Furthermore, these interfaces can be designed with an added

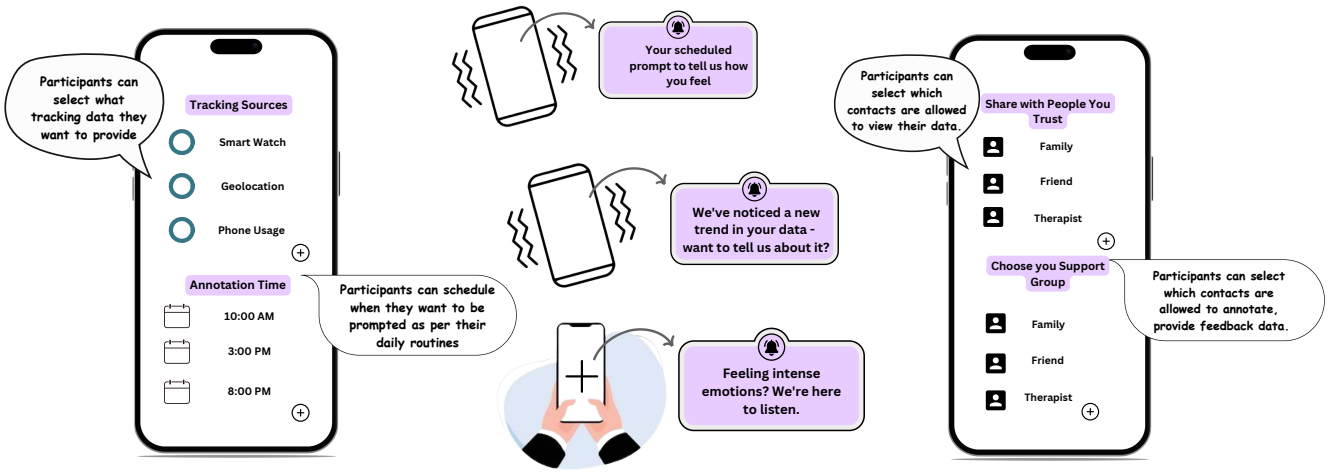



Fig. 5. Visualization of Integrating Participant Agency into the Design Process. The first screen illustrates how participants can exercise agency by selecting preferred data sources and specifying suitable time slots for receiving prompts based on their individual schedules. The second screen presents three prompting strategies: (1) prompts delivered at user-specified times, (2) context-aware prompts triggered by physiological or behavioral indicators, and (3) user-initiated annotations during emotionally salient moments. To further support multi-perspective reflection, participants are also given the option to include input from trusted members of their support network.

redundancy layer to enhance data collection consistency. This redundancy can be implemented by adding a quick-annotation option in each annotation mode, ensuring a baseline level of information, followed by options to add deeper reflections. This structure helps maintain a consistent data format while supporting varying user engagement levels. Moreover, the interface can also contain a curated list of activities that users can select to suggest the situational context of their data.

**2) User-Powered Interface for Prompting and Multi-Source Assessments:** We propose an emotion co-annotation platform where users can choose what data to share, set personalized prompting schedules, and invite trusted individuals to contribute their perspectives, as illustrated in Figure 5. These tools can allow users to set personalized prompting conditions (e.g., time-based, data-triggered [8, 62, 65, 90], or self-initiated) and control the granularity of the emotion data they wish to share. Overall, such systems can enhance user agency. Moreover, an additional Likert scale to provide feedback on the confidence of emotional assessments can also be incorporated [131] for users and other sources. This additional confidence assessment from various sources can help recognize the validity of data and provide users with an additional layer of reflection on their annotations.

**3) Interface to Accommodate Learning and Support:** We propose adding data insights, verified sources for emotional well-being and awareness guidance, and LLM-supported guidance systems to the data-collection applications. These design elements can enhance user engagement by motivating them to understand themselves better, as illustrated in Figure 6.

Lastly, future work can explore a range of mood tracking and self-reflection applications that prioritize adaptability, user privacy, data sharing, and personalized emotional insight. For example, iMoodJournal [57] allows users to select their mood from an extensive list of emotions and supplement entries with journal notes, images, and location tags. It supports mood log sharing while maintaining a strong emphasis on data privacy. Similarly, Apple Health [12] offers a “State of Mind” mood logging feature, which prompts users to first categorize

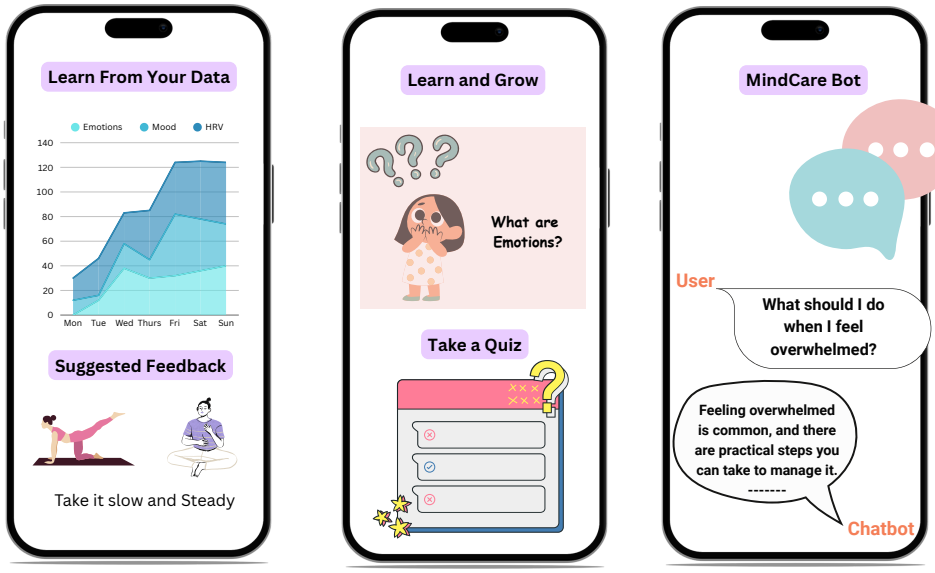



Fig. 6. Visualization of Participant Engagement, Learning, and Support Elements Integrated into the Design. The first screen displays personalized data insights derived from participants' inputs to foster self-reflection. The second screen offers curated, trustworthy information aimed at enhancing emotional awareness and literacy. The third screen illustrates how LLM-supported systems can be incorporated to provide contextually relevant guidance and emotional support.

their mood as positive or negative, then select specific emotions with the option for multiple selections. Users can also add contextual notes and identify potential causes of mood changes, such as activities and relationships, while choosing between real-time and daily summary logging. Additional examples of feedback-oriented platforms include Mindsera [93], an AI-powered journaling app that provides emotion analysis and personalized suggestions to guide self-reflection and promote mental fitness. Another example is Daylio [2], which offers a quick and streamlined interface for logging moods and activities multiple times throughout the day. Overall, platforms that integrate mood tracking with everyday lifestyle have the potential to generate more accurate, actionable emotion data and offer valuable insights for both research and personal well-being.

## 6.2 Understanding the Implications of Pre-Study Guidelines

> *"I am made of little rooms full of thoughts, emotions & memories. You cannot define me by listening to me once. I'm too complex."* [Source: Unknown, Credit: Pinterest]

Prior research has often relied on one-way communication for everyday emotion data collection, borrowing heavily from traditional lab-based methodologies. However, everyday settings differ significantly from lab environments, as they lack the level of control typically available in laboratories. This lack of control introduces challenges in ensuring the quality and reliability of the collected data. To overcome these challenges, it is crucial to prepare a data collection pipeline that is robust to noise and bias in the real world. Beyond the lack of control, our findings also highlighted the diversity in participants' attitudes toward collecting and sharing emotional data, influenced by varying levels of emotional literacy. This diversity has been shown to impact emotion tracking among participants in previous studies [23, 67, 79, 121, 175]. Drawing from data-centric AI guidelines [59], ethical considerations for emotion AI [143], and past literature on emotion tracking, our findings emphasize the need for careful pre-preparation which involves: 1) **careful selection of participants** (G1), by clearly defining

the inclusion and exclusion criteria, selecting participants from diverse backgrounds [88], and screening for possible conditions that could impact the data, 2) **preparing the participants** by elaborately informing (G2, G12), and training (G3–G4) them about how to interact efficiently with devices and annotation methods involved in data-collection methodology and its benefits, 3) **understanding the participants emotional profile** by performing elaborate psycho-social profiling (G5) and comprehensive demographic data collection (G6). Including these steps in the data collection pipeline can enhance participant engagement but also minimize errors, reduce hesitations, and foster a sense of collaboration between researchers and participants, which was often missing in prior methods [54, 60, 135, 142]. This careful pre-preparation will ensure an inclusive experience for participants of varying levels of emotional awareness. Furthermore, gathering psycho-social profiles and comprehensive demographic data will allow researchers to collect broader context about emotional responses missing in prior context collection that was limited to activity levels, basic demographics, and personality traits [60, 146] and will further help researchers to tailor the data collection process to the participants' emotional traits and lifestyles.

### 6.3 *Go with the Flow*: Understanding the Implications of During Data-Collection Guidelines

Prior research on emotion data collection has typically focused on two approaches for collecting emotion data in real-life settings. 1) Designing real-life emotional scenarios, such as work-related stress [54, 142], group entertainment [22], or driving stress [49]. 2) Complete in-situ settings- where data collectors rely completely on participants' willingness to complete ecological momentary assessments (EMAs) or emotion questionnaires [60, 135]. These approaches often lead to data of a specific emotional scenario or incomplete data with limited contextual information. However, our findings emphasized the **inherent diversity in participants' attitudes** towards tracking and sharing emotion data and suggested designing sampling strategies that can accommodate this diversity. To address the challenges posed by the diverse engagement styles, we propose designing annotation methods that are tailored to the specific needs and capabilities of different individuals while also being flexible enough to support a broad range of participants. We recommend designing adaptable methods that can accommodate the varying needs of people, as recommended in section 6.1. Additionally, for participants with lower emotional literacy, researchers can frame emotional tools as practical aids rather than self-reflective interventions (e.g., stress relief or productivity enhancers) to increase engagement. Subsequently, for clinical participants or participants dealing with emotional trauma or other life-changing events, researchers can investigate a multiple-assessment approach [154]. Furthermore, our findings show that participants preferred using objective methods (such as scales) and lower frequencies on neutral days, while subjective methods (like written descriptions) and higher frequencies were favored during periods of intense emotions. However, present approaches such as ESM (In-the-moment annotation) and DRM (after-the-fact annotation) [45, 72, 130, 144, 157] often overlook this fluidity in emotional experiences. These methods use either a fixed scale (e.g., SAM, Likert Scale) or questionnaires (e.g., STAI, PHQ-9) to capture emotion ratings within fixed or random time periods. This often doesn't provide users with an opportunity to label as per emotion intensity, thus leading to datasets that fail to capture the dynamic and contextual aspects of emotional experiences, instead treating emotions as discrete snapshots, to overcome these challenges we recommend designing systems that can adapt to users changing emotional landscapes (G9), as shown in Figure 4. Moreover, the timing of the annotation prompt can significantly affect the precision of annotations (G7, G8). For instance, in-the-moment annotation requires participants to assess and record their emotions as they occur, which can capture more immediate and authentic emotional states. However, this method can be cognitively demanding, as participants need to be aware of their emotions while balancing other activities in their environment [45, 72, 157]. In contrast, after-the-fact annotation allows participants to reflect on their emotional experiences once they have passed, which can provide a more thorough and considered response. However, this retrospective approach comes with its own cognitive challenges: memory bias and difficulty in recalling the intensity or nuances of past emotions accurately [130, 144]. This can lead to

data that may not fully reflect the emotional state experienced at the time, impacting the validity of the data for training AI systems. We recommend future works to design participant-aware sampling techniques and adaptable annotation methods that combine closed-end and open-ended questions (G9) as shown in Figures 4 and 5. Further, an interface for adding contextual metadata, such as associated events or environmental factors, alongside emotional labels [19, 142], should also be added. Finally, our data also shed light on the **psychological influences that self-reporting emotions** can have on participants' daily lives. While self-reporting can foster self-awareness and provide emotional patterns, it can also influence the user's emotions in unintended ways. For instance, users shared that recording subtle negative emotions can amplify overthinking. Conversely, documenting positive emotions can foster a sense of gratitude. To overcome potential influences, we recommend designing supportive and non-judgmental annotation techniques (G9, G11). For example, adding reflective prompts to encourage users to frame negative emotions constructively, like "What can this feeling teach me?". Further, LLM-based structured journaling activities, guidance for mindfulness or relaxation exercises, and references to resources during distressing periods can be added to the applications [101, 134]. Further tools can integrate features that allow users to record emotions without immediate analysis and then review entries after a period of detachment, or ask participants to note a small positive event or something they feel grateful for (G11), can be added, as shown in the prototype figure 6.

### 6.4 Moving beyond the Traditional Data Modeling: Implications on Post-Data Collection Approaches

Our findings emphasize that the concept of **"emotional ground truth"** extends far beyond the survey responses typically gathered through standard questionnaires. However, current datasets often assume a universal definition of emotions and one-to-one mappings between emotion data and filled surveys [15, 41, 139], to label emotion data. Moreover, prior work on developing models for physiological emotion data often applies simplistic labeling approaches like categorizing emotions into discrete groups based on objective labels, such as predefined emotion categories (e.g., happy, angry) or scales (e.g., 1 to 5). These methods often do not use additional contextual data [126, 129, 133, 146], while modeling the AI algorithms leading to the development of models that cannot be adapted in real-life [51] or clinical settings [6, 7]. The continued use of such approaches can be attributed to several factors: 1) Simplifying emotion categorization reduces the complexity of emotion recognition models, making them easier to develop, train, and implement. 2) Discrete emotion categories are easier for participants or experts to label, lowering the annotation burden. 3) Standardizing datasets based on these categories facilitates generalization across various AI applications, such as sentiment analysis and video emotion recognition. 4) The influence of early psychological theories, such as basic emotion theory, has strongly shaped these practices. However, our findings based on interviews with domain experts challenge these assumptions. Experts argue that actionable insights and contextually relevant data should take precedence over overly generic labeling (G13, G14). They suggested that emotional ground truth is not a simple, one-to-one mapping from data to labels. In fact, it's a composite representation that varies according to the user profile.

Insights from both participants and experts have shaped our guidelines for collecting emotion data that is dynamic, layered, and actionable. Unlike traditional methods, our approach captures emotions in real-time and across varying contexts, resulting in data that is fundamentally different in structure and complexity. This shift highlights the need for new labeling and validation techniques that can accommodate the richness and variability of the collected data. Traditional emotion datasets often collect inputs, such as single-point self-reports, task-based annotations, or expert labels, resulting in relatively uniform data structures that are easy to label. These inputs are then reduced to binary or discrete categories (e.g., "happy" or "stressed") by binning self-reported or task-driven labels to fit downstream tasks. For example, in the WESAD dataset [129], emotional states are classified into stress versus no-stress categories based solely on experimental stimuli, without incorporating participant self-reports. Similarly, GLOBEM [171] focuses on depression detection as a downstream task, and ASCERTAIN [146] performs

arousal-valence classification based on self-reports. In contrast, our approach captures emotion as a dynamic, evolving state, influenced by contextual, physiological, and subjective factors as discussed in section 4. This results in data that is more variable and multidimensional. For instance, each emotional annotation in our system may include a combination of quick scale ratings, emotion labels, option text/audio/image data, confidence scores, and contextual metadata. The structure and depth of these annotations can vary based on the user's engagement and the intensity of the emotional experience. Such variability introduces both opportunities and challenges: while the data offers a more accurate and holistic view of emotional states, it also complicates traditional labeling and validation pipelines, which typically assume uniform input formats.

To effectively handle our dynamic data, we propose a set of new validation schemes that go beyond traditional practices. **1) Triangulated Validation**: In this technique, we can assign a final emotion score or label by combining information from multiple sources, such as scale-based self-reports, emotion-list, physiological signals, AI-generated or text annotations, images/audio annotations (optional), and contextual, peer, or expert feedback. Each of these sources can first be evaluated for coherence and reliability in a given context, and a confidence score can be assigned to them. For example, if a user provides a confident self-report, it might carry a higher weight of 0.9, while physiological signals with strong indicators could be weighted at 0.8, and AI-based reflections with uncertain text data might be assigned a lower weight, such as 0.5. All emotion representations are then aligned into a common format, such as a valence-arousal score or a set of discrete emotion categories. Finally, the final label can be computed using a weighted aggregation, such as a weighted average for numerical scores or a confidence-weighted majority vote for categorical labels. This ensures that more reliable sources contribute more to the outcome. Overall, this approach allows for a more robust and context-aware emotional label, addressing the limitations of relying on any single data source. **2) Semantic Validation**: Given the redundant nature of information collected through multiple methods, such as scale-based self-reports, emotion names from the list, or optional text, should be validated for semantics. This means making sure that elements like emotion labels, confidence scores, and multimedia content (such as text, images, or audio) align coherently. For example, if an annotation includes the emotion label "joy," but the accompanying text expresses sadness or the image shows someone crying, a mismatch may need to be addressed. Similarly, if users rate their emotional intensity as very high but give a very low confidence score, that inconsistency could indicate confusion or noise in the data. This form of validation can add a layer of reliability in collected data, which is often missing in traditional data. **3) Contextual validation**: This involves checking whether the data fits logically within the context in which it was collected. This validation technique is similar to traditional approaches of checking the data contextually. **4) Annotation Agreement Validation and Co-Development:** This validation focuses on assessing the consistency of emotion annotation when multiple annotators (participants, experts, and peer groups) are involved in labeling the same content. Since emotions are highly personal and subjective, it's common for different users to interpret the same situation differently. This step helps identify how much agreement or disagreement exists among annotators. Techniques like inter-annotator agreement metrics (e.g., Cohen's Kappa or Krippendorff's Alpha) can be used by future works to quantify the level of consistency across annotations. This approach also provides a framework for emotion data collectors, mental health professionals, and emotion AI experts to co-develop and evaluate new tools and methodologies with users. By bringing together multiple stakeholders in the data validation process, the resulting systems can benefit from diverse expertise: users' lived experiences, clinicians' domain knowledge, and technical experts' implementation capabilities. This collaborative approach ensures data collection tools are not only technically sound but also clinically relevant and ethically implemented. Consequently, these validation techniques can validate emotion data more robustly, and they also align with domain experts' guided strategies of finding *congruence* in emotional assessments. Further, they provide a platform to add clinical and participant insights to traditional data, thus adding an opportunity for co-development with experts while keeping participants in the loop. Lastly, our findings underscore the critical need to design AI algorithms that prioritize actionable outcomes [8], such as identifying meaningful patterns—like recurring emotional states—over simplistic

emotional categorizations [23, 73]. This shift is essential for developing systems that align more closely with real-world applications. For example, in therapeutic settings, recognizing patterns in emotional states over time can help identify triggers or trends in mental health, providing valuable insights for personalized interventions. However, for clinical settings, tracking changes in symptoms can be targeted over nuanced emotional changes. By focusing on actionable outcomes, algorithms can also provide deeper insights for decision-making, enabling stakeholders to address the underlying causes of emotional responses rather than just classifying emotions into predefined categories. This approach moves away from a rigid framework of labeling emotions, embracing a more dynamic and context-sensitive model of emotion tracking.

## 7 Limitations

This study aimed to explore people's attitudes and preferences toward tracking and monitoring emotions in everyday settings for emotion AI data collection and interventions, and provide a set of guidelines for future emotion data collection methods. A limitation of our work was that most of our participants were well-educated, tech-savvy individuals familiar with AI, emotion monitoring, and wearable technologies. Thus, our findings might not generalize well to people with low literacy and less experience with emotion tracking. We also recognize that our findings might be influenced by the participants' demographics, group composition, and backgrounds, since all our participants belonged to the same cultural background and country. To address this, we have included a diverse audience of users and non-users of emotion-tracking technology with varying levels of technological familiarity, emotional awareness, and demographic profiles (age, gender, occupation, education). Moreover, it is also important to recognize that different research objectives may encounter unique challenges when adapting these guidelines. Thus, we recommend that future work customize these guidelines to their specific needs for better adaptability. For instance, participant training and psycho-social profiling can be significantly more challenging when working with clinical populations compared to undiagnosed, healthy counterparts. Individuals with diagnosed mental disorders may require tailored approaches to ensure ethical considerations, comfort, and engagement throughout the data collection process. To address these complexities, we recommend engagement with mental health professionals to navigate the sensitivities associated with clinical populations. This is particularly crucial given that many emotion monitoring interventions are designed to target individuals with diagnosed mental health conditions. By including professional supervision, researchers can better align their methodologies with the needs of clinical audiences, creating a more inclusive, ethical, and effective data collection process [8] tailored to diverse participant groups.

## 8 Conclusion

Our study investigated the perspectives of key stakeholders (public and mental health professionals) on annotating emotion data as part of everyday life. Previously, emotion data collection relied on approaches based on objective scales and questionnaires, both in laboratory settings and real-world environments. However, the impact of user-specific factors and the influence of everyday contexts on the quality of emotion annotations has been largely understudied. Our analysis reveals that factors such as the fluidity of emotional experiences, stigma, and emotional literacy significantly affect the accuracy of these annotations. By examining these factors, our study provides a comprehensive understanding of their implications on data collection in everyday settings. Based on these insights, we offer a framework *AnnoSense* for future works to develop more holistic approaches for emotion data collection. In the future, we aim to expand these design guidelines into practical solutions and algorithms suitable for daily life settings, exploring their effectiveness in future solutions for wearable and mobile-based emotion AI systems.

## 9 Acknowledgements

We gratefully acknowledge Dr. Koushik Sinha Deb from the Department of Psychiatry at AIIMS New Delhi for his valuable support in recruiting focus group participants. We also extend our sincere thanks to all participants and evaluators for their time, openness, and for sharing their insights throughout the study. And acknowledge the support of the iHub-Anubhuti-IIITD Foundation, established under the NM-ICPS scheme of the DST at IIIT-Delhi.

## References

[1] 2024. Calm - Meditation and Sleep App. https://www.calm.com.
[2] 2024. Daylio - Mood Tracker and Diary. https://daylio.net.
[3] 2024. Headspace - Mindfulness and Meditation App. https://www.headspace.com.
[4] 2024. Woebot - Your Self-Care Expert. https://woebothealth.com.
[5] 2024. Wysa - Mental Health Support. https://www.wysa.io.
[6] Alaa Abd-Alrazaq, Rawan AlSaad, Manale Harfouche, Sarah Aziz, Arfan Ahmed, Rafat Damseh, and Javaid Sheikh. 2023. Wearable artificial intelligence for detecting anxiety: systematic review and meta-analysis. *Journal of medical Internet research* 25 (2023), e48754.
[7] Alaa Abd-Alrazaq, Rawan AlSaad, Farag Shuweihdi, Arfan Ahmed, Sarah Aziz, and Javaid Sheikh. 2023. Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression. *NPJ Digital Medicine* 6, 1 (2023), 84.
[8] Daniel A. Adler, Yuewen Yang, Thalia Viranda, Xuhai Xu, David C. Mohr, Anna R. Van Meter, Julia C. Tartaglia, Nicholas C. Jacobson, Fei Wang, Deborah Estrin, and Tanzeem Choudhury. 2024. Beyond Detection: Towards Actionable Sensing Research in Clinical Mental Healthcare. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 160 (Nov. 2024), 33 pages. doi:10.1145/3699755
[9] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.
[10] Nazanin Andalibi and Justin Buss. 2020. The Human in Emotion Recognition on Social Media: Attitudes, Outcomes, Risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3313831.3376680
[11] Emmi Antikainen, Anna Iashina, Iman Alikhani, and Mari Karsikas. 2024. How acute stress affects sleep: large-scale observations from continuous smart ring measurements in free-living conditions. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 1–4.
[12] Apple Inc. 2024. Apple Watch. Retrieved 2025-04-12 from https://www.apple.com/watch/
[13] David Bakker and Nikki Rickard. 2018. Engagement in mobile phone app for self-monitoring of emotional wellbeing predicts changes in mental health: MoodPrism. *Journal of affective disorders* 227 (2018), 432–442.
[14] Swarnali Banik, Sougata Sen, Snehanshu Saha, and Surjya Ghosh. 2024. Towards Reducing Continuous Emotion Annotation Effort During Video Consumption: A Physiological Response Profiling Approach. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 91 (Sept. 2024), 32 pages. doi:10.1145/3678569
[15] Lisa Feldman Barrett. 2017. *How emotions are made: The secret life of the brain.* Pan Macmillan.
[16] Lisa Feldman Barrett, James Gross, Tamlin Conner Christensen, and Michael Benvenuto. 2001. Knowing what you're feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation. *Cognition & Emotion* 15, 6 (2001), 713–724.
[17] Kathy Baxter, Catherine Courage, and Kelly Caine. 2015. *Understanding your users: a practical guide to user research methods.* Morgan Kaufmann.
[18] Maciej Behnke, Mikołaj Buchwald, Adam Bykowski, Szymon Kupiński, and Lukasz D Kaczmarek. 2022. Psychophysiology of positive and negative emotions, dataset of 1157 cases and 8 biosignals. *Scientific Data* 9, 1, 10.
[19] Ananya Bhattacharjee, Dana Kulzhabayeva, Mohi Reza, Harsh Kumar, Eunchae Seong, Xuening Wu, Mohammad Rashidujjaman Rifat, Robert Bowman, Rachel Kornfield, Alex Mariakakis, Syed Ishtiaque Ahmed, Munmun De Choudhury, Gavin Doherty, Mary P Czerwinski, and Joseph Jay Williams. 2023. Integrating Individual and Social Contexts into Self-Reflection Technologies. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 356, 6 pages. doi:10.1145/3544549.3573803
[20] Aditya Bhattacharya, Simone Stumpf, and Katrien Verbert. 2024. Representation Debiasing of Generated Data Involving Domain Experts. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (Cagliari, Italy) *(UMAP Adjunct '24)*. Association for Computing Machinery, New York, NY, USA, 516–522. doi:10.1145/3631700.3664910
[21] Javad Birjandtalab, Diana Cogan, Maziyar Baran Pouyan, and Mehrdad Nourani. 2016. A Non-EEG Biosignals Dataset for Assessment and Visualization of Neurological Status. In *2016 IEEE International Workshop on Signal Processing Systems (SiPS)*. 110–114. doi:10.1109/

SiPS.2016.27
[22] Patrícia Bota, Joana Brito, Ana Fred, Pablo Cesar, and Hugo Silva. 2024. A real-world dataset of group emotion experiences based on physiological data. *Scientific Data* 11, 1 (2024), 116.
[23] Dionne Bowie-DaBreo, Corina Sas, Heather Iles-Smith, and Sandra Sünram-Lea. 2022. User Perspectives and Ethical Experiences of Apps for Depression: A Qualitative Analysis of User Reviews. 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 21, 24 pages. doi:10.1145/3491102.3517498
[24] Gregory J Boyle. 1984. Reliability and validity of Izard's differential emotions scale. *Personality and individual Differences* 5, 6 (1984), 747–750.
[25] 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, 49–59.
[26] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
[27] Matteo Busso, Andrea Bontempelli, Leonardo Javier Malcotti, Lakmal Meegahapola, Peter Kun, Shyam Diwakar, Chaitanya Nutakki, Marcelo Dario Rodas Britez, Hao Xu, Donglei Song, Salvador Ruiz Correa, Andrea-Rebeca Mendoza-Lara, George Gaskell, Sally Stares, Miriam Bidoglia, Amarsanaa Ganbold, Altangerel Chagnaa, Luca Cernuzzi, Alethia Hume, Ronald Chenu-Abente, Roy Alia Asiku, Ivan Kayongo, Daniel Gatica-Perez, Amalia de Götzen, Ivano Bison, and Fausto Giunchiglia. 2025. DiversityOne: A Multi-Country Smartphone Sensor Dataset for Everyday Life Behavior Modeling. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 1, Article 1 (March 2025), 49 pages. doi:10.1145/3712289
[28] Hava Chaptoukaev, Valeriya Strizhkova, Michele Panariello, Bianca Dalpaos, Aglind Reka, Valeria Manera, Susanne Thümmler, Esma ISMAILOVA, Nicholas W., francois bremond, Massimiliano Todisco, Maria A Zuluaga, and Laura M. Ferrari. 2023. StressID: a Multimodal Dataset for Stress Identification. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 29798–29811. https://proceedings.neurips.cc/paper_files/paper/2023/file/5f09bfe6730e9627a9f800d01a8ad5cd-Paper-Datasets_and_Benchmarks.pdf
[29] Si Chen, Haocong Cheng, Jason Situ, and Yun Huang. 2023. Mirror Hearts: Exploring the (Mis-)Alignment between AI-Recognized and Self-Reported Emotions. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 206, 7 pages. doi:10.1145/3544549.3585607
[30] Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The journal of positive psychology* 12, 3 (2017), 297–298.
[31] Shanley Corvite, Kat Roemmich, Tillie Ilana Rosenberg, and Nazanin Andalibi. 2023. Data Subjects' Perspectives on Emotion Artificial Intelligence Use in the Workplace: A Relational Ethics Lens. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 124 (apr 2023), 38 pages. doi:10.1145/3579600
[32] Jean Costa, François Guimbretière, Malte F Jung, and Tanzeem Choudhury. 2019. Boostmeup: Improving cognitive performance in the moment by unobtrusively regulating emotions with a smartwatch. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–23.
[33] Vedant Das Swain, Victor Chen, Shrija Mishra, Stephen M Mattingly, Gregory D Abowd, and Munmun De Choudhury. 2022. Semantic gap in predicting mental wellbeing through passive sensing. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–16.
[34] PMA Desmet. 2003. Measuring emotion. *M. Blythe, A Monk, K. Overbeeke, & P* (2003).
[35] Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, et al. 2007. The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. In *Affective Computing and Intelligent Interaction: Second International Conference, ACII 2007 Lisbon, Portugal, September 12-14, 2007 Proceedings 2*. Springer, 488–500.
[36] Maciej Dzieżyc, Joanna Komoszyńska, Stanisław Saganowski, Magda Boruch, Jakub Dziwiński, Katarzyna Jabłońska, Dominika Kunc, and Przemysław Kazienko. 2021. How to catch them all? Enhanced data collection for emotion recognition in the field. In *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. 348–351. doi:10.1109/PerComWorkshops51409.2021.9431143
[37] Paul Ekman. 1992. Are there basic emotions? (1992).
[38] Felix Engel, Alphonsus Keary, Kevin Berwind, Marco Xaver Bornschlegl, and Matthias Hemmje. 2017. The role of reproducibility in affective computing. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2008–2014. doi:10.1109/BIBM.2017.8217969
[39] Denzil Ferreira, Vassilis Kostakos, and Anind K Dey. 2015. AWARE: mobile context instrumentation framework. *Frontiers in ICT* 2 (2015), 6.
[40] Fitbit, Inc. 2024. Fitbit Wearables. Retrieved 2025-04-12 from https://www.fitbit.com/global/us/products
[41] Nan Gao, Soundariya Ananthan, Chun Yu, Yuntao Wang, and Flora D Salim. 2023. Critiquing Self-report Practices for Human Mental and Wellbeing Computing at Ubicomp. *arXiv preprint arXiv:2311.15496* (2023).

[42] Nan Gao, Mohammad Saiedur Rahaman, Wei Shao, and Flora D Salim. 2021. Investigating the Reliability of Self-report Data in the Wild: The Quest for Ground Truth. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers* (Virtual, USA) *(UbiComp/ISWC '21 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 237–242. doi:10.1145/3460418.3479338
[43] Garmin Ltd. 2024. Garmin Smartwatches. Retrieved 2025-04-12 from https://www.garmin.com/en-US/c/wearables/
[44] Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2021. Designing an Experience Sampling Method for Smartphone Based Emotion Detection. *IEEE Transactions on Affective Computing* 12, 4 (2021), 913–927. doi:10.1109/TAFFC.2019.2905561
[45] Surjya Ghosh, Bivas Mitra, and Pradipta De. 2020. Towards Improving Emotion Self-report Collection using Self-reflection. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3334480.3383019
[46] Google Inc. 2015. Paco: Personal Analytics Companion. https://code.google.com/archive/p/paco/
[47] Happify Research. 2025. Happify. https://www.happify.com/research/ Mobile application.
[48] Mark G Haviland. 1996. Structure of the twenty-item Toronto Alexithymia Scale. *Journal of personality assessment* 66, 1 (1996), 116–125.
[49] Jennifer A Healey and Rosalind W Picard. 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems* 6, 2 (2005), 156–166.
[50] Alexander Heimerl, Pooja Prajod, Silvan Mertes, Tobias Baur, Matthias Kraus, Ailin Liu, Helen Risack, Nicolas Rohleder, Elisabeth André, and Linda Becker. 2023. ForDigitStress: A multi-modal stress dataset employing a digital job interview scenario. *arXiv preprint arXiv:2303.07742*.
[51] Blake Anthony Hickey, Taryn Chalmers, Phillip Newton, Chin-Teng Lin, David Sibbritt, Craig S McLachlan, Roderick Clifton-Bligh, John Morley, and Sara Lal. 2021. Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review. *Sensors* 21, 10 (2021), 3461.
[52] Kristina Höök. 2009. Affective loop experiences: designing for interactional embodiment. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1535 (2009), 3585–3595.
[53] Seyedmajid Hosseini, Raju Gottumukkala, Satya Katragadda, Ravi Teja Bhupatiraju, Ziad Ashkar, Christoph W Borst, and Kenneth Cochran. 2022. A multimodal sensor dataset for continuous stress detection of nurses in a hospital. *Scientific Data* 9, 1, 255.
[54] Seyedmajid Hosseini, Raju Gottumukkala, Satya Katragadda, Ravi Teja Bhupatiraju, Ziad Ashkar, Christoph W Borst, and Kenneth Cochran. 2022. A multimodal sensor dataset for continuous stress detection of nurses in a hospital. *Scientific Data* 9, 1 (2022), 255.
[55] Annika Howells, Itai Ivtzan, and Francisco Jose Eiroa-Orosa. 2016. Putting the 'app'in happiness: a randomised controlled trial of a smartphone-based mindfulness intervention to enhance wellbeing. *Journal of happiness studies* 17 (2016), 163–185.
[56] ilumivu. 2024. Ecological Momentary Assessment (mEMA) App. Retrieved 2025-07-15 from https://ilumivu.com/solutions/ecological-momentary-assessment-app/
[57] Inexika Inc. 2025. iMoodJournal – Mood Tracking Mobile Application. https://www.imoodjournal.com/.
[58] Laura SF Israel and Felix D Schönbrodt. 2019. Emotion prediction with weighted appraisal models–Towards validating a psychological theory of affect. *IEEE Transactions on Affective Computing* 13, 2 (2019), 604–615.
[59] Mohammad Hossein Jarrahi, Ali Memariani, and Shion Guha. 2023. The Principles of Data-Centric AI. *Commun. ACM* 66, 8 (July 2023), 84–92. doi:10.1145/3571724
[60] Soowon Kang, Woohyeok Choi, Cheul Young Park, Narae Cha, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Heepyung Kim, Yong Jeong, and Uichin Lee. 2023. K-emophone: A mobile and wearable dataset with in-situ emotion, stress, and attention labels. *Scientific data* 10, 1, 351.
[61] 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. doi:10.1145/3491102.3501944
[62] Evangelos Karapanos. 2012. Beyond Experience Sampling: Evaluating Personal Informatics with Technology-Assisted Reconstruction. *arXiv preprint arXiv:1207.1821* (2012).
[63] Kleomenis Katevas, Hamed Haddadi, and Laurissa Tokarchuk. 2016. Sensingkit: Evaluating the sensor power consumption in ios devices. In *2016 12th International conference on intelligent environments (IE)*. IEEE, 222–225.
[64] Harmanpreet Kaur, Daniel McDuff, Alex C. Williams, Jaime Teevan, and Shamsi T. Iqbal. 2022. "I Didn't Know I Looked Angry": Characterizing Observed Emotion and Reported Affect at Work. 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 199, 18 pages. doi:10.1145/3491102.3517453
[65] Jane Kaye, Edgar A Whitley, David Lund, Michael Morrison, Harriet Teare, and Karen Melham. 2015. Dynamic consent: a patient interface for twenty-first century research networks. *European journal of human genetics* 23, 2 (2015), 141–146.
[66] Abe Kazemzadeh, Sungbok Lee, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2011. Emotion twenty questions: Toward a crowd-sourced theory of emotions. In *International conference on affective computing and intelligent interaction*. Springer, 1–10.

[67] Christina Kelley, Bongshin Lee, and Lauren Wilcox. 2017. Self-tracking for Mental Wellness: Understanding Expert Perspectives and Student Experiences. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 629–641. doi:10.1145/3025453.3025750

[68] Vassilis-Javed Khan, Panos Markopoulos, Berry Eggen, Wijnand IJsselsteijn, and Boris de Ruyter. 2008. Reconexp: a way to reduce the data loss of the experiencing sampling method. In *Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services* (Amsterdam, The Netherlands) *(MobileHCI '08)*. Association for Computing Machinery, New York, NY, USA, 471–476. doi:10.1145/1409240.1409316

[69] Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-Woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2024. MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 701, 20 pages. doi:10.1145/3613904.3642937

[70] Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-Woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2024. MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24, Vol. 55)*. ACM, 1–20. doi:10.1145/3613904.3642937

[71] Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-LLM: Large Language Models for Health Prediction via Wearable Sensor Data. arXiv:2401.06866 [cs.CL] https://arxiv.org/abs/2401.06866

[72] Thomas Kosch, Mariam Hassib, Robin Reutter, and Florian Alt. 2020. Emotions on the go: Mobile emotion assessment in real-time using facial expressions. In *Proceedings of the International Conference on Advanced Visual Interfaces*. 1–9.

[73] 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 88, 16 pages. doi:10.1145/3544548.3581209

[74] Krzysztof Kutt, Dominika Drążyk, Laura Żuchowska, Maciej Szelążek, Szymon Bobek, and Grzegorz J Nalepa. 2022. BIRAFFE2, a multimodal dataset for emotion-based personalization in rich affective game environments. *Scientific data* 9, 1, 274.

[75] Richard D Lane, Donald M Quinlan, Gary E Schwartz, Pamela A Walker, and Sharon B Zeitlin. 1990. The Levels of Emotional Awareness Scale: A cognitive-developmental measure of emotion. *Journal of personality assessment* 55, 1-2 (1990), 124–134.

[76] Matias Laporte, Martin Gjoreski, and Marc Langheinrich. 2023. LAUREATE: A Dataset for Supporting Research in Affective Computing and Human Memory Augmentation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 106, 41 pages. doi:10.1145/3610892

[77] Fanny Larradet, Radoslaw Niewiadomski, Giacinto Barresi, Darwin G Caldwell, and Leonardo S Mattos. 2020. Toward emotion recognition from physiological signals in the wild: approaching the methodological issues in real-life data collection. *Frontiers in psychology* 11 (2020), 1111.

[78] Fanny Larradet, Radoslaw Niewiadomski, Giacinto Barresi, and Leonardo S Mattos. 2019. Appraisal theory-based mobile app for physiological data collection and labelling in the wild. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 752–756.

[79] Emily G. Lattie, Rachel Kornfield, Kathryn E. Ringland, Renwen Zhang, Nathan Winquist, and Madhu Reddy. 2020. Designing Mental Health Technologies that Support the Social Ecosystem of College Students. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3313831.3376362

[80] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction*. Morgan Kaufmann.

[81] Richard S Lazarus. 1991. Progress on a cognitive-motivational-relational theory of emotion. *American psychologist* 46, 8 (1991), 819.

[82] Junze Li, Changyang He, Jiaxiong Hu, Boyang Jia, Alon Halevy, and Xiaojuan Ma. 2024. DiaryHelper: Exploring the Use of an Automatic Contextual Information Recording Agent for Elicitation Diary Study. arXiv:2404.19738 [cs.HC] https://arxiv.org/abs/2404.19738

[83] Kristen A Lindquist, Maria Gendron, Lisa Feldman Barrett, and Bradford C Dickerson. 2014. Emotion perception, but not affect perception, is impaired with semantic memory loss. *Emotion* 14, 2 (2014), 375.

[84] Susanna Loeb, Susan Dynarski, Daniel McFarland, Pamela Morris, Sean Reardon, and Sarah Reber. 2017. Descriptive Analysis in Education: A Guide for Researchers. NCEE 2017-4023. *National Center for Education Evaluation and Regional Assistance* (2017).

[85] Bhargavi Mahesh, Teena Hassan, Erwin Prassler, and Jens-Uwe Garbas. 2019. Requirements for a Reference Dataset for Multimodal Human Stress Detection. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. 492–498. doi:10.1109/PERCOMW.2019.8730884

[86] Ruth Malkinson. 2010. Cognitive-behavioral grief therapy: The ABC model of rational-emotion behavior therapy. *Psihologijske teme* 19, 2 (2010), 289–305.

[87] Valentina Markova, Todor Ganchev, and Kalin Kalinkov. 2019. CLAS: A Database for Cognitive Load, Affect and Stress Recognition. In *2019 International Conference on Biomedical Innovations and Applications (BIA)*. 1–4. doi:10.1109/BIA48344.2019.8967457

[88] Lakmal Meegahapola, William Droz, Peter Kun, Amalia de Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, George Gaskell, Altangerel Chagnaa, Amarsanaa Ganbold, Tsolmon Zundui, Carlo Caprini, Daniele

Miorandi, Alethia Hume, Jose Luis Zarza, Luca Cernuzzi, Ivano Bison, Marcelo Rodas Britez, Matteo Busso, Ronald Chenu-Abente, Can Günel, Fausto Giunchiglia, Laura Schelenz, and Daniel Gatica-Perez. 2023. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 176 (Jan. 2023), 32 pages. doi:10.1145/3569483

[89] Kristof Meers, Egon Dejonckheere, Elise K Kalokerinos, Koen Rummens, and Peter Kuppens. 2020. mobileQ: A free user-friendly application for collecting experience sampling data. *Behavior Research Methods* 52 (2020), 1510–1515.

[90] Mike A. Merrill, Akshay Paruchuri, Naghmeh Rezaei, Geza Kovacs, Javier Perez, Yun Liu, Erik Schenck, Nova Hammerquist, Jake Sunshine, Shyam Tailor, Kumar Ayush, Hao-Wei Su, Qian He, Cory Y. McLean, Mark Malhotra, Shwetak Patel, Jiening Zhan, Tim Althoff, Daniel McDuff, and Xin Liu. 2024. Transforming Wearable Data into Health Insights using Large Language Model Agents. arXiv:2406.06464 [cs.AI] https://arxiv.org/abs/2406.06464

[91] Angeliki Metallinou and Shrikanth Narayanan. 2013. Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 1–8. doi:10.1109/FG.2013.6553804

[92] MetricWire Inc. 2025. MetricWire. Retrieved 2025-04-12 from https://www.metricwire.com

[93] Mindsera. 2025. Mindsera: AI-powered journal for mental fitness. https://www.mindsera.com/.

[94] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. 2021. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. *IEEE Transactions on Affective Computing* 12, 2 (April 2021), 479–493. doi:10.1109/TAFFC.2018.2884461

[95] Varun Mishra, Byron Lowens, Sarah Lord, Kelly Caine, and David Kotz. 2017. Investigating contextual cues as indicators for EMA delivery. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers* (Maui, Hawaii) *(UbiComp '17)*. Association for Computing Machinery, New York, NY, USA, 935–940. doi:10.1145/3123024.3124571

[96] Varun Mishra, Sougata Sen, Grace Chen, Tian Hao, Jeffrey Rogers, Ching-Hua Chen, and David Kotz. 2020. Evaluating the Reproducibility of Physiological Stress Detection Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 147 (Dec. 2020), 29 pages. doi:10.1145/3432220

[97] Vivian Genaro Motti. 2019. Assistive wearables: opportunities and challenges. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (London, United Kingdom) *(UbiComp/ISWC '19 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 1040–1043. doi:10.1145/3341162.3349573

[98] Sujay Nagaraj, Sarah Goodday, Thomas Hartvigsen, Adrien Boch, Kopal Garg, Sindhu Gowda, Luca Foschini, Marzyeh Ghassemi, Stephen Friend, and Anna Goldenberg. 2023. Dissecting the heterogeneity of "in the wild" stress from multimodal sensor data. *NPJ Digital Medicine* 6, 1 (2023), 237.

[99] Peter Neigel, Andrew Vargo, Benjamin Tag, and Koichi Kise. 2024. Using Wearables to Unobtrusively Identify Periods of Stress in a Real University Environment. In *Proceedings of the 2024 ACM International Symposium on Wearable Computers*. 17–24.

[100] Subigya Nepal, Arvind Pillai, William Campbell, Talie Massachi, Eunsol Soul Choi, Xuhai Xu, Joanna Kuc, Jeremy F Huckins, Jason Holden, Colin Depp, et al. 2024. Contextual AI Journaling: Integrating LLM and Time Series Behavioral Sensing Technology to Promote Self-Reflection and Well-being using the MindScape App. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[101] Subigya Nepal, Arvind Pillai, William Campbell, Talie Massachi, Michael V. Heinz, Ashmita Kunwar, Eunsol Soul Choi, Xuhai Xu, Joanna Kuc, Jeremy F. Huckins, Jason Holden, Sarah M. Preum, Colin Depp, Nicholas Jacobson, Mary P. Czerwinski, Eric Granholm, and Andrew T. Campbell. 2024. MindScape Study: Integrating LLM and Behavioral Sensing for Personalized AI-Driven Journaling Experiences. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 186 (Nov. 2024), 44 pages. doi:10.1145/3699761

[102] Matthew K Nock, Michelle M Wedig, Elizabeth B Holmberg, and Jill M Hooley. 2008. The emotion reactivity scale: development, evaluation, and relation to self-injurious thoughts and behaviors. *Behavior therapy* 39, 2 (2008), 107–116.

[103] Ju Lynn Ong, TeYang Lau, Mari Karsikas, Hannu Kinnunen, and Michael WL Chee. 2021. A longitudinal analysis of COVID-19 lockdown stringency on sleep and resting heart rate measures across 20 countries. *Scientific Reports* 11, 1 (2021), 14413.

[104] Jukka-Pekka Onnela, Caleb Dixon, Keary Griffin, Tucker Jaenicke, Leila Minowada, Sean Esterkin, Alvin Siu, Josh Zagorsky, and Eli Jones. 2021. Beiwe: A data collection platform for high-throughput digital phenotyping. *Journal of Open Source Software* 6, 68 (2021), 3417.

[105] Oura Health Oy. 2024. Oura Ring. Retrieved 2025-04-12 from https://ouraring.com

[106] Shantanu Pal, Subhas Mukhopadhyay, and Nagender Suryadevara. 2021. Development and Progress in Sensors and Technologies for Human Emotion Recognition, In Sensor Technology for Improving Human Movements and Postures. *Sensors* 21, 16. doi:10.3390/s21165554

[107] Lawrence A Palinkas, Sarah M Horwitz, Carla A Green, Jennifer P Wisdom, Naihua Duan, and Kimberly Hoagwood. 2015. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and policy in mental*

*health and mental health services research* 42 (2015), 533–544.

[108] Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. 2020. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data* 7, 1 (2020), 293.

[109] Ignacio Perez-Pozuelo, Dimitris Spathis, Emma AD Clifton, and Cecilia Mascolo. 2021. Wearables, smartphones, and artificial intelligence for digital phenotyping and health. In *Digital health*. Elsevier, 33–54.

[110] Rosalind Picard. 1997. W.,(1997). Affective Computing.

[111] Kathleen Pine, Claus Bossen, Naja Holten Møller, Milagros Miceli, Alex Jiahong Lu, Yunan Chen, Leah Horgan, Zhaoyuan Su, Gina Neff, and Melissa Mazmanian. 2022. Investigating Data Work Across Domains: New Perspectives on the Work of Creating Data. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 87, 6 pages. doi:10.1145/3491101.3503724

[112] Robert Plutchik. 1982. A psychoevolutionary theory of emotions.

[113] 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* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 725–734. doi:10.1145/1978942.1979047

[114] Pooja Prajod, Bhargavi Mahesh, and Elisabeth André. 2024. Stressor Type Matters! – Exploring Factors Influencing Cross-Dataset Generalizability of Physiological Stress Detection. arXiv:2405.09563 [eess.SP] https://arxiv.org/abs/2405.09563

[115] David Preece, Rodrigo Becerra, Ken Robinson, Justine Dandy, and Alfred Allan. 2018. The psychometric assessment of alexithymia: Development and validation of the Perth Alexithymia Questionnaire. *Personality and Individual Differences* 132 (2018), 32–44.

[116] Nina Rajcic and Jon McCormack. 2020. Mirror ritual: An affective interface for emotional self-reflection. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[117] Nina Rajcic and Jon McCormack. 2020. Mirror Ritual: An Affective Interface for Emotional Self-Reflection. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376625

[118] Nikki Rickard, Hussain-Abdulah Arjmand, David Bakker, Elizabeth Seabrook, et al. 2016. Development of a mobile phone app to support self-monitoring of emotional well-being: a mental health digital innovation. *JMIR mental health* 3, 4 (2016), e6202.

[119] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 1–8. doi:10.1109/FG.2013.6553805

[120] Tobias Röddiger, Christopher Clarke, Paula Breitling, Tim Schneegans, Haibin Zhao, Hans Gellersen, and Michael Beigl. 2022. Sensing with Earables: A Systematic Literature Review and Taxonomy of Phenomena. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 135 (Sept. 2022), 57 pages. doi:10.1145/3550314

[121] Kat Roemmich and Nazanin Andalibi. 2021. Data Subjects' Conceptualizations of and Attitudes Toward Automatic Emotion Recognition-Enabled Wellbeing Interventions on Social Media. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 308 (oct 2021), 34 pages. doi:10.1145/3476049

[122] Kat Roemmich, Florian Schaub, and Nazanin Andalibi. 2023. Emotion AI at Work: Implications for Workplace Surveillance, Emotional Labor, and Emotional Privacy. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 588, 20 pages. doi:10.1145/3544548.3580950

[123] James A Russell, Anna Weiss, and Gerald A Mendelsohn. 1989. Affect grid: a single-item scale of pleasure and arousal. *Journal of personality and social psychology* 57, 3 (1989), 493.

[124] Stanisław Saganowski, Joanna Komoszyńska, Maciej Behnke, Bartosz Perz, Dominika Kunc, Bartłomiej Klich, Łukasz D Kaczmarek, and Przemysław Kazienko. 2022. Emognition dataset: emotion recognition with self-reports, facial expressions, and physiology using wearables. *Scientific data* 9, 1, 158.

[125] Stanisław Saganowski, Joanna Komoszyńska, Maciej Behnke, Bartosz Perz, Dominika Kunc, Bartłomiej Klich, Łukasz D Kaczmarek, and Przemysław Kazienko. 2022. Emognition dataset: emotion recognition with self-reports, facial expressions, and physiology using wearables. *Scientific data* 9, 1 (2022), 158.

[126] Stanisław Saganowski, Bartosz Perz, Adam G. Polak, and Przemysław Kazienko. 2023. Emotion Recognition for Everyday Life Using Physiological Signals From Wearables: A Systematic Literature Review. *IEEE Transactions on Affective Computing* 14, 3 (2023), 1876–1897. doi:10.1109/TAFFC.2022.3176135

[127] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 39, 15 pages. doi:10.1145/3411764.3445518

[128] Samsung Electronics. 2024. Samsung Galaxy Watch. Retrieved 2025-04-12 from https://www.samsung.com/global/galaxy/galaxy-watch/

[129] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 400–408. doi:10.1145/3242969.3242985

[130] Stefan Schneider, Doerte U Junghaenel, Tania Gutsche, Hio Wa Mak, and Arthur A Stone. 2020. Comparability of emotion dynamics derived from ecological momentary assessments, daily diaries, and the day reconstruction method: Observational study. *Journal of Medical Internet Research* 22, 9 (2020), e19201.

[131] Marc Schröder, Hannes Pirker, and Myriam Lamolle. 2006. First suggestions for an emotion annotation and representation language. In *Proceedings of LREC*, Vol. 6. 88–92.

[132] Bahar Sert and Selami Varol Ülker. 2023. A Review of LIWC and Machine Learning Approaches On Mental Health Diagnosis. *Social Review of Technology and Change* 1, 2 (2023), 71–92.

[133] Karan Sharma, Claudio Castellini, Egon L Van Den Broek, Alin Albu-Schaeffer, and Friedhelm Schwenker. 2019. A dataset of continuous affect annotations and physiological signals for emotion analysis. *Scientific data* 6, 1, 196.

[134] Donghoon Shin, Subeen Park, Esther Hehsun Kim, Soomin Kim, Jinwook Seo, and Hwajung Hong. 2022. Exploring the Effects of AI-assisted Emotional Support Processes in Online Mental Health Community. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 300, 7 pages. doi:10.1145/3491101.3519854

[135] Xinyu Shui, Mi Zhang, Zhuoran Li, Xin Hu, Fei Wang, and Dan Zhang. 2021. A dataset of daily ambulatory psychological and physiological recording for emotion research. *Scientific data* 8, 1, 161.

[136] Nicolas Simonazzi, Jean-Marc Salotti, Marie Morelle, Caroline Dubois, and Philippe Le Goff. 2021. The Geneva Emotion Wheel Mobile Interface: an Instrument to Report Emotions on Android Devices. In *ERGO'IA 2021-De l'Interaction Homme-Machine à la Relation Homme-Machine, comment concevoir des systèmes performants et éthiques.*

[137] Pragya Singh, Ritvik Budhiraja, Ankush Gupta, Anshul Goswami, Mohan Kumar, and Pushpendra Singh. 2024. EEVR: A Dataset of Paired Physiological Signals and Textual Descriptions for Joint Emotion Representation Learning. *Advances in Neural Information Processing Systems* 37 (2024), 15765–15778.

[138] Pragya Singh, Ritvik Budhiraja, Pankaj Jalote, Mohan Kumar, and Pushpendra Singh. 2025. Translating Emotions to Annotations: A Participant's Perspective of Physiological Emotion Data Collection. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW195 (May 2025), 30 pages. doi:10.1145/3711093

[139] Pragya Singh, Mohan Kumar, and Pushpendra Singh. 2024. Can we say a cat is a cat? Understanding the challenges in annotating physiological signal-based emotion data. arXiv:2406.14908 [cs.HC] https://arxiv.org/abs/2406.14908

[140] Petr Slovak, Alissa Antle, Nikki Theofanopoulou, Claudia Daudén Roquet, James Gross, and Katherine Isbister. 2023. Designing for emotion regulation interventions: an agenda for HCI theory and research. *ACM Transactions on Computer-Human Interaction* 30, 1 (2023), 1–51.

[141] Elena Smets, Emmanuel Rios Velazquez, Giuseppina Schiavone, Imen Chakroun, Ellie D'Hondt, Walter De Raedt, Jan Cornelis, Olivier Janssens, Sofie Van Hoecke, Stephan Claes, et al. 2018. Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *NPJ digital medicine* 1, 1, 67.

[142] Elena Smets, Emmanuel Rios Velazquez, Giuseppina Schiavone, Imen Chakroun, Ellie D'Hondt, Walter De Raedt, Jan Cornelis, Olivier Janssens, Sofie Van Hoecke, Stephan Claes, et al. 2018. Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *NPJ digital medicine* 1, 1 (2018), 67.

[143] Luke Stark and Jesse Hoey. 2021. The Ethics of Emotion in Artificial Intelligence Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 782–793. doi:10.1145/3442188.3445939

[144] Arthur A Stone, Joseph E Schwartz, David Schkade, Norbert Schwarz, Alan Krueger, and Daniel Kahneman. 2006. A population approach to the study of emotion: diurnal rhythms of a working day examined with the Day Reconstruction Method. *Emotion* 6, 1 (2006), 139.

[145] Samuel J Stratton. 2021. Population research: convenience sampling strategies. *Prehospital and disaster Medicine* 36, 4 (2021), 373–374.

[146] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L. Vieriu, Stefan Winkler, and Nicu Sebe. 2018. ASCERTAIN: Emotion and Personality Recognition Using Commercial Sensors. *IEEE Transactions on Affective Computing* 9, 2, 147–160. doi:10.1109/TAFFC.2016.2625250

[147] Luma Tabbaa, Ryan Searle, Saber Mirzaee Bafti, Md Moinul Hossain, Jittrapol Intarasisrisawat, Maxine Glancy, and Chee Siang Ang. 2021. Vreed: Virtual reality emotion recognition dataset using eye tracking & physiological measures. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5, 4 (2021), 1–20.

[148] Sabrina Thai and Elizabeth Page-Gould. 2018. ExperienceSampler: An open-source scaffold for building smartphone apps for experience sampling. *Psychological Methods* 23, 4 (2018), 729.

[149] Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*

27, 5 (2020), 1–53.
[150] Leimin Tian, Sharon Oviatt, Michal Muszynski, Brent Chamberlain, Jennifer Healey, and Akane Sano. 2022. Applied Affective Computing. Morgan & Claypool.
[151] Maria Dolores C Tongco. 2007. Purposive sampling as a tool for informant selection. (2007).
[152] Peter Traunmuller, Anice Jahanjoo, Soheil Khooyooz, Amin Aminifar, and Nima TaheriNejad. 2024. Wearable Healthcare Devices for Monitoring Stress and Attention Level in Workplace Environments. arXiv:2406.05813 [cs.HC] https://arxiv.org/abs/2406.05813
[153] Charalampos Tsirmpas, Dimitrios Andrikopoulos, Panagiotis Fatouros, Georgios Eleftheriou, Joaquin A Anguera, Konstantinos Kontoangelos, and Charalabos Papageorgiou. 2022. Feasibility, engagement, and preliminary clinical outcomes of a digital biodata-driven intervention for anxiety and depression. *Frontiers in digital health* 4 (2022), 868970.
[154] Anupriya Tuli, Pushpendra Singh, Mamta Sood, Koushik Sinha Deb, Siddharth Jain, Abhishek Jain, Manan Wason, Rakesh Chadda, and Rohit Verma. 2016. Harmony: close knitted mhealth assistance for patients, caregivers and doctors for managing SMIs. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 1144–1152. doi:10.1145/2968219.2968301
[155] Aditya Vaidyam, John Halamka, John Torous, et al. 2022. Enabling research and clinical use of patient-generated health data (the mindLAMP Platform): digital phenotyping study. *JMIR mHealth and uHealth* 10, 1 (2022), e30557.
[156] Niels van Berkel, Simon D'Alfonso, Rio Kurnia Susanto, Denzil Ferreira, and Vassilis Kostakos. 2023. AWARE-Light: A smartphone tool for experience sampling and digital phenotyping. *Personal and Ubiquitous Computing* 27, 2 (2023), 435–445.
[157] 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. *Int. J. Hum.-Comput. Stud.* 125, C (may 2019), 118–128. doi:10.1016/j.ijhcs.2018.12.002
[158] Veniamin Veselovsky, Berke Argin, Benedikt Stroebl, Chris Wendler, Robert West, James Evans, Thomas L. Griffiths, and Arvind Narayanan. 2025. Localized Cultural Knowledge is Conserved and Controllable in Large Language Models. arXiv:2504.10191 [cs.CL] https://arxiv.org/abs/2504.10191
[159] Alexander Viola, Vladimir Pavlovic, and Sejong Yoon. 2021. Constructivist approaches for computational emotions: A systematic survey. In *AAAI Fall Symposium*. Springer, 30–50.
[160] Liuping Wang, Xiangmin Fan, Feng Tian, Lingjia Deng, Shuai Ma, Jin Huang, and Hongan Wang. 2018. mirrorU: scaffolding emotional reflection via in-situ assessment and interactive feedback. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.
[161] Liuping Wang, Xiangmin Fan, Feng Tian, Lingjia Deng, Shuai Ma, Jin Huang, and Hongan Wang. 2018. mirrorU: Scaffolding Emotional Reflection via In-Situ Assessment and Interactive Feedback. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3170427.3188517
[162] 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.
[163] Asra Sakeen Wani, Ishika Joshi, Nadia Ishfaq Nahvi, and Pushpendra Singh. 2024. " Unrest and trauma stays with you!": Navigating mental health and professional service-seeking in Kashmir. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
[164] Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and Eliathamby Ambikairajah. 2021. A Comprehensive Review of Speech Emotion Recognition Systems, In IEEE Access. *IEEE Access* 9, 47795–47814. doi:10.1109/ACCESS.2021.3068045
[165] David Watson and Lee Anna Clark. 1994. The PANAS-X: Manual for the positive and negative affect schedule-expanded form. *Unpublished manuscript, University of Iowa* (1994).
[166] 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, 1063.
[167] Amy Wenzel. 2017. Basic strategies of cognitive behavioral therapy. *Psychiatric Clinics* 40, 4 (2017), 597–609.
[168] Whoop, Inc. 2024. WHOOP Wearable. Retrieved 2025-04-12 from https://www.whoop.com
[169] Hamidan Z. Wijasena, Ridi Ferdiana, and Sunu Wibirama. 2021. A Survey of Emotion Recognition using Physiological Signal in Wearable Devices. In *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*. 1–6. doi:10.1109/AIMS52415.2021.9466092
[170] Ruolan Wu, Chun Yu, Xiaole Pan, Yujia Liu, Ningning Zhang, Yue Fu, Yuhan Wang, Zhi Zheng, Li Chen, Qiaolei Jiang, Xuhai Xu, and Yuanchun Shi. 2024. MindShift: Leveraging Large Language Models for Mental-States-Based Problematic Smartphone Use Intervention. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 248, 24 pages. doi:10.1145/3613904.3642790
[171] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff.

2022. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (Dec. 2022), 1–34. doi:10.1145/3569485

[172] Yeonsun Yang, Ahyeon Shin, Nayoung Kim, Huidam Woo, John Joon Young Chung, and Jean Y Song. 2024. Find the Bot!: Gamifying Facial Emotion Recognition for Both Human Training and Machine Learning Data Collection. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 364, 20 pages. doi:10.1145/3613904.3642880

[173] Joanna C. Yau, Benjamin Girault, Tiantian Feng, Karel Mundnich, Amrutha Nadarajan, Brandon M. Booth, Emilio Ferrara, Kristina Lerman, Eric Hsieh, and Shrikanth Narayanan. 2022. TILES-2019, A longitudinal physiologic and behavioral data set of medical residents in an intensive care unit. *Sci Data* 9, 536 (2022). doi:10.1038/s41597-022-01636-4

[174] Yonghai Yu and Yun Bi. 2010. A study on "5W1H" user analysis on interaction design of interface. In *2010 IEEE 11th International Conference on Computer-Aided Industrial Design & Conceptual Design 1*, Vol. 1. IEEE, 329–332.

[175] Renwen Zhang, Kathryn E. Ringland, Melina Paan, David C. Mohr, and Madhu Reddy. 2021. Designing for Emotional Well-being: Integrating Persuasion and Customization into Mental Health Technologies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 542, 13 pages. doi:10.1145/3411764.3445771

## A Semi-Structured Interview Guide

This appendix presents the semi-structured interview guide used to explore participants' experiences, perceptions, and preferences related to emotion annotation. Given the semi-structured nature of the interviews, the questions were adapted as needed to ensure clarity and comprehensibility for participants. The guide is organized according to the *Who, What, When, Where, Why,* and *How* framework, followed by additional probing questions.

**WHO: Participant Background and Emotional Self-Reflection**

**Self-Reflection on Emotions**

- Could you tell me a bit about yourself, particularly in terms of how you experience and relate to emotions?
- How would you describe your emotional landscape and the role emotions play in your daily life?
- How do you perceive your ability to manage or process emotions?
- Would you describe yourself as more emotionally expressive or emotionally reserved?
- How do you typically respond to emotional experiences?
- To what extent would you consider yourself emotionally self-aware?

**Familiarity with Technology**

- How familiar are you with using digital technologies, such as mobile applications or wearable devices, for tracking or annotating emotions?

**Experience with Emotion Annotation**

- Have you had any prior experience with tracking or annotating your emotions?
- Have you used any specific tools or methods—such as journaling, mood-tracking apps (e.g., Likert scales, emojis), or verbal/voice recordings—to annotate emotions?

**Psychological Impact of Annotation**

- How does the process of annotating emotions affect you psychologically?
- Do you find it therapeutic, stressful, or something else?
- In what ways does it influence your emotional awareness and understanding?

*Note to participants: "Emotion annotation refers to the practice of labeling or recording emotional states, often to support self-reflection, research, or the training of AI systems."*

**WHAT: Content and Scope of Annotation**

- What kinds of emotions do you think should be annotated?
- Which emotional states or types of experiences do you believe are most important to capture?
- Can you provide examples of specific situations or emotional experiences that you would consider annotating?
- Do you have any privacy concerns regarding emotion annotation? Would you feel comfortable annotating deeply personal emotions in detail?

**WHEN: Timing of Emotion Annotation**

- When do you think is the most appropriate time to annotate emotions?
- (For participants with prior experience) When do you typically annotate your emotions, and in what kinds of scenarios?
- Would you prefer to annotate emotions in real-time (immediately after experiencing them), or retrospectively (e.g., summarizing emotions at the end of the day)? Why?
- How frequently do you believe emotional annotation should occur?

**WHERE: Context and Environment for Annotation**

- In what types of environments would you feel most comfortable annotating your emotions?
- Would you prefer to annotate emotions at home, in the workplace, or in another setting? Why?
- Are there any places or contexts where you would feel uncomfortable annotating emotions?
  - If participant responds "alone," follow up with: "If you are unable to be alone—e.g., at work or in a public space—would you still feel comfortable annotating?"
- Can you describe a scenario in which annotating emotions would be particularly difficult?
  - What factors would contribute to that difficulty?
  - How might you address or overcome them?

**WHY: Motivation and Perceived Value**

- Why do you think annotating emotions is important or meaningful?
- What personal benefits do you associate with the annotation of positive or negative emotions?
- What challenges or barriers do you foresee in the emotion annotation process?

**HOW: Preferred Methods and Tools for Annotation**

- How would you go about annotating your emotions?
- What tools or methods would you prefer to use (e.g., paper journals, apps with Likert scales or emojis, voice recordings)?
- How much time would you be willing to dedicate to emotion annotation per day or week?
- What features or types of support would make the annotation process easier or more engaging?
- Are there specific functions or aids (e.g., reminders, visualizations, AI feedback) that would help you annotate more effectively?
- How could the emotion annotation process be simplified or made more intuitive?

**Additional Questions**

- What are your overall expectations from the annotation process?
  - What outcomes do you hope to achieve?
  - How would you evaluate or measure the success of the process?

## B Focus Group Discussion

**Introduction**

- Brief overview of the study and purpose with presentation.
- Warm-up conversation and informed consent.

**Current Practices for Assessing Emotional States**

- How do you currently assess the emotional states of your patients?
- What tools or techniques do you use to collect data on your patients' emotions?

**Attitude towards Data and AI**

- Can you elaborate on what AI tools you would like to use in your practice?
- What are your initial thoughts on the use of AI to understand and monitor emotions?
- How do you think AI can enhance emotional well-being and mental health care?
- What are the potential benefits and drawbacks of using AI for emotional recognition in clinical settings?

**Emotion Data Collection**

- What are opportunities for the present ways of emotion annotations (as presented in the introduction), and why, according to you?
- Which emotions do you believe are important to track daily to maintain good mental well-being?
- In your experience, how easy is it for individuals to understand their emotions?
- Do you think the process would be more challenging for people who are emotionally susceptible or are suffering from some minor disorders?
- What do you see as the main challenges in collecting emotion data in everyday settings?
- What should we call the "emotion ground truth" or "emotion label," and why?
- At what resolution (e.g., frequency, granularity) should we track emotions to make effective interventions?

## C Survey Questionnaire

**Section 1: Participant Background**

- Consent to participate
- Age, Gender, Education, Occupation

**Section 2: Understanding Emotional Awareness**

- **Q1.** How often do you reflect on your emotions?
- **Q2.** How easily can you identify emotions during strong experiences?
- **Q3.** How often do you feel mixed emotions?
- **Q4.** Do you use any tools (e.g., journaling, mood tracking apps)?
- **Q5.** Think about a recent time when you felt a strong emotion. What emotion did you feel? (Please write a brief description of the situation and the emotion you identified)
- **Q6.** Looking back at the situation you described in question above and how accurate do you think your emotion label was?
- **Q7.** Name up to 5 positive emotions you feel daily and their impact in your daily life.
- **Q8.** Name up to 5 negative emotions you feel daily and their impact and in your daily life.
- **Q9.** Which emotions are easiest to identify, and why?
- **Q10.** Which emotions are hardest to identify, and why?
- **Q11.** Can you differentiate between similar emotions (e.g., sadness vs. disappointment or anger and frustration)? Explain how?
- **Q12.** What does the intensity of an emotion mean to you? Explain?
- **Q13.** What best describes an "emotion" (select all that apply)?

**Section 3: Attitudes Toward Daily Emotion Annotation**

- **Q14.** How confident are you in labeling your emotions accurately?
- **Q15.** How well can you label mixed emotions?
- **Q16.** How would you prefer to annotate emotions (e.g., text, emojis, scale)?
- **Q17.** Explain why you chose a particular option in Q16?
- **Q18.** What factors are most important when labeling emotions? (e.g., context, physical response)
- **Q19.** Why did you choose your preferred annotation method?
- **Q20.** Do cultural or societal factors influence your emotion labeling? Please explain.
- **Q21.** Would you like to annotate emotions daily?
- **Q22.** If Yes, Why would you like to annotate your emotions daily?
- **Q23.** If No, Why not would you like to annotate your emotions daily?
- **Q24.** How easy is daily emotion annotation for you?
- **Q25.** How frequently can you annotate emotions?
- **Q26.** Would you annotate negative emotions (e.g., anger, stress)? Why or why not?
- **Q27.** Would you annotate positive emotions (e.g., calm, joy)? Why or why not?

| Question No. | Completed Response | Response Rate (%) | Required Question | Word Count Summary |
|---|---|---|---|---|
| Q4 | 41 | 54.67 | Yes | Range = 1–68, Mean = 6.74, SD = 15.23 |
| Q5 | 70 | 93.33 | Yes | Range = 1–105, Mean = 22.59, SD = 24.44 |
| Q7 | 69 | 92.00 | Yes | Range = 1–121, Mean = 14.93, SD = 26.28 |
| Q8 | 68 | 90.67 | Yes | Range = 1–78, Mean = 13.00, SD = 17.49 |
| Q9 | 72 | 96.00 | Yes | Range = 1–47, Mean = 10.88, SD = 10.91 |
| Q10 | 61 | 81.33 | Yes | Range = 1–43, Mean = 10.13, SD = 10.58 |
| Q11 | 73 | 97.33 | Yes | Range = 1–110, Mean = 18.39, SD = 21.32 |
| Q12 | 66 | 88.00 | Yes | Range = 2–100, Mean = 19.91, SD = 17.37 |
| Q17 | 62 | 82.67 | No | Range = 2–119, Mean = 21.37, SD = 20.20 |
| Q20 | 26 of 44 | 59.09 | No | Range = 5–196, Mean = 36.15, SD = 40.51 |
| Q22 | 25 of 28 | 89.29 | No | Range = 3–39, Mean = 14.20, SD = 10.19 |
| Q23 | 45 of 47 | 95.74 | No | Range = 1–48, Mean = 14.53, SD = 11.66 |
| Q26 | 66 | 88.00 | No | Range = 1–57, Mean = 14.23, SD = 11.96 |
| Q27 | 62 | 82.67 | No | Range = 1–61, Mean = 12.79, SD = 10.97 |

Table 14. Quantitative Summary of Open-Ended Responses in our Survey