Exploring Data-Driven Approaches to Stress Management: A Systematic Review of Stress Tracking, Intervention, and System Evaluation Methods

Youngji Koh
School of Computing
KAIST
Daejeon, Republic of Korea
youngji@kaist.ac.kr

Jeonghyun Kim
School of Computing
KAIST
Daejeon, Republic of Korea
jeonghyun.kim@kaist.ac.kr

Kwangyoung Lee
Department of Industrial Design
KAIST
Daejeon, Republic of Korea
kwangyoung@kaist.ac.kr

Yugyeong Jung
School of Computing
KAIST
Daejeon, Republic of Korea
yugyeong.jung@kaist.ac.kr

Hwajung Hong
Department of Industrial Design
KAIST
Deajeon, Republic of Korea
hwajung@kaist.ac.kr

Uichin Lee*
School of Computing
KAIST
Daejeon, Republic of Korea
uclee@kaist.ac.kr

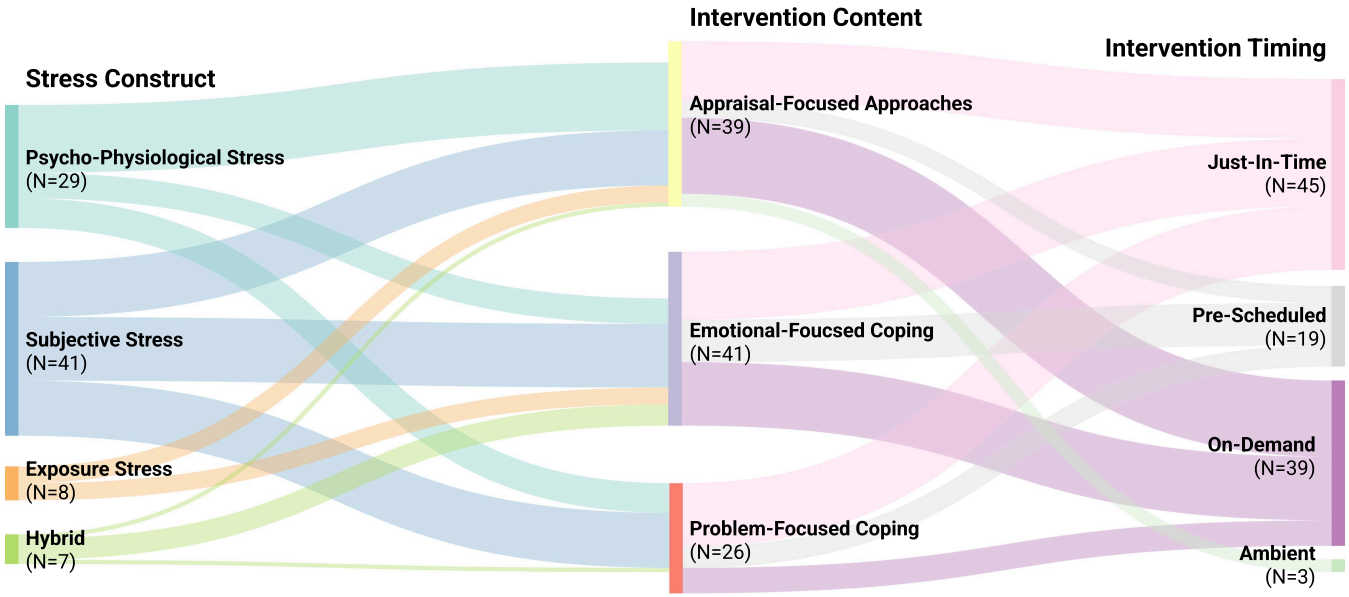

Figure 1: A Sankey diagram illustrating the landscape of data-driven stress management systems. This figure visualizes the flow and distribution from how stress is defined (Stress Construct), to the content of the intervention (Intervention Content), and finally to its delivery timing (Intervention Timing) across the reviewed literature. The N value in each node denotes the total number of instances classified under that category. As a single study may fall into multiple categories (e.g., using both 'Emotion-Focused' and 'Problem-Focused' strategies), the total flow may expand across stages.

*Corresponding author

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791194

Abstract

Advances in ubiquitous and wearable sensing and HCI research have made stress monitoring increasingly accessible, enabling the development of personalized stress management technologies. Yet, stress is a subjective and contextual experience, making effective intervention design challenging. Prior studies often isolate stress detection or intervention, without providing an integrated view of how these components connect and are evaluated in real-world

use. To address this gap, we conducted a systematic review of 2,152 papers and selected 52 empirical studies where stress tracking informed interventions. Using a framework based on three stress constructs (subjective stress, psycho-physiological stress, and exposure stress), we analyzed how definitions of stress shape detection indicators, intervention design and timing, and evaluation methods. We show that stress conceptualization strongly influences system design, and we propose a conceptual framework linking detection, intervention, and evaluation to guide future user-centered stress management technologies.

CCS Concepts

• **Human-centered computing → User studies**; **Ubiquitous and mobile computing design and evaluation methods**.

Keywords

stress management, stress detection, self-tracking, intervention systems, evaluation methods, systematic review

ACM Reference Format:

Youngji Koh, Jeonghyun Kim, Kwangyoung Lee, Yugyeong Jung, Hwajung Hong, and Uichin Lee. 2026. Exploring Data-Driven Approaches to Stress Management: A Systematic Review of Stress Tracking, Intervention, and System Evaluation Methods. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3772318.3791194

1 Introduction

Recent advancements in ubiquitous and wearable sensing technologies have enabled the monitoring of stress in situ, capturing behavioral signals [75], facial expressions [113], and physiological responses such as heart rate and electrodermal activity (EDA) [35, 97]. In parallel, subjective measures such as ecological momentary assessments (EMA) capture lived experiences that complement objective signals [46]. These technological advances create new opportunities for timely and accessible stress management interventions, especially in everyday life when support is most needed [88].

Commercial devices have already incorporated stress tracking and intervention features. For instance, Fitbit Sense monitors EDA and heart rate variability (HRV) and offers guided breathing or mindfulness exercises. Similarly, Samsung Galaxy Watch and Apple Watch estimate stress through HRV and support interventions such as breathing reminders and reflection sessions. These systems highlight the mainstream adoption of stress management technologies. Building on this trend, research in HCI has explored more advanced directions, including context-aware and adaptive interventions that integrate stress detection with tailored support [33, 68].

While the rise of self-tracking technologies offers a promising avenue for support, designing effective interventions is uniquely challenging due to the inherent complexity of stress. We acknowledge that stress is not unequivocally problematic; it can be beneficial (eustress) or intentionally induced for training purposes [74, 104]. However, this review primarily focuses on contexts where stress is perceived as a burden requiring management or mitigation. Stress is not a simple, uniform signal but a deeply subjective and contextual experience that varies across individuals and situations [52]. This complexity makes it crucial to systematically examine not only how systems detect stress (ranging from manual subjective reports to automatic detection with sensors), but also how these detection outputs are linked to interventions in ways that matter to users. When this connection is weak or unclear, interventions risk feeling irrelevant or burdensome. A user-centered perspective is therefore crucial for making stress management systems effective in everyday life. Such a perspective also requires systematic evaluation, since only through evaluation we can assess whether detection methods and interventions genuinely reflect users' needs and lived experiences. A review that synthesizes how systems have connected detection, intervention, and evaluation can provide insights for guiding the design of future stress management technologies.

However, prior literature reviews have offered only a fragmented perspective. Most have focused either on stress detection [1, 10, 27, 64] or on intervention strategies [51, 76, 86], leaving open the question of how these two components can be connected. Only limited attempts have considered detection and intervention together [20, 95], yet these reviews paid limited attention to the diversity of intervention strategies or to evaluation practices. This absence of an integrative perspective motivates the need for a review that systematically examines stress detection, intervention, and evaluation as a unified whole.

To address this gap, our review analyzes existing research on stress management systems in HCI. We adopt a conceptual framework (visualized in Figure 2) that bridges sensing technologies with theoretical constructs. Drawing inspiration from Lazarus and Folkman's transactional model [52], we leverage this framework to organize prior work into three broad categories based on how stress is defined and operationalized as an inference target: (1) subjective stress, defined by an individual's cognitive appraisal of whether a situation is threatening or challenging; (2) psycho-physiological and expressive stress, encompassing bodily changes and behavioral manifestations that occur as reactions to stress, and (3) stress defined in terms of exposure to external stressors or tasks, focusing on the stimulus itself.

Using this categorization, we examine not only how systems specify stress as an inference target and what indicators are used to infer it, but also how these inferences guide the design and timing of intervention strategies and how their effectiveness has been evaluated. This integrative perspective allows us to map the connections between stress constructs, indicators, intervention designs, and evaluation practices across prior work.

Based on this approach, we ask the following research questions:

- **RQ1**: How have systems defined stress as an inference target, and what indicators have been used to infer it?
- **RQ2**: How have these inferences been connected to the design and delivery of interventions?
- **RQ3**: How have stress management systems been evaluated in practice?

To answer our research questions, we identified 2,152 papers and, through a rigorous selection process, narrowed the scope to 52 empirical studies that leverage stress tracking to inform interventions. Our analysis of these studies examines three core aspects: (1) the sensing modalities and models used to define a stress construct,

(2) the content and timing of the resulting interventions, and (3) the methods used for user evaluation. As illustrated in Figure 1, we provide an overview of how the design process diverges depending on the guiding stress construct. The design process clearly diverges based on the guiding stress construct. Systems defining stress psycho-physiologically via biosignals (e.g., HRV, EDA) typically lead to appraisal-focused interventions like biofeedback and visualization, often delivered just-in-time. In contrast, systems targeting subjective stress connect to a broader range of active coping strategies, such as cognitive behavior therapy (CBT) and mindfulness, with more flexible delivery. In terms of user evaluation, most systems were validated through feasibility studies focused on usability and short-term effects, rather than rigorous efficacy trials with larger and more diverse populations. These studies typically relied on a mix of psychological scales and standardized usability surveys. Building on these findings, our discussion addresses the challenges and future directions for each of these three domains: self-tracking, intervention design, and user evaluation.

The main contributions of this review are as follows:

- We provide an integrative review that jointly examines stress detection, intervention strategies, and evaluation practices in stress management systems.
- We analyze the critical link between how stress is conceptualized and measured and the resulting design choices for interventions.
- Based on our findings, we identify key challenges and opportunities that can guide future research on the design and evaluation of user-centered stress management technologies.

2 Background and Related Work

This section provides background on how stress has been conceptualized and measured in HCI, reviews prior work on stress management systems, and discusses how our review positions itself relative to existing surveys.

2.1 Defining and Measuring Stress in HCI

The term '*stress*' was famously defined by Hans Selye as 'the nonspecific response of the body to a demand,' with the '*stressor*' being the stimulus [84]. This foundational view has been significantly expanded by the transactional model of stress and coping [52]. This model posits that a stressor does not automatically lead to a stress response. Instead, the process is mediated by an individual's cognitive appraisal, which refers to the subjective evaluation of whether the event represents a threat, harm, or challenge. For instance, an appraisal of 'challenge' or intentional exposure can lead to positive engagement (eustress) or serve training purposes (e.g., stress inoculation) [74, 104], whereas 'threat' often leads to distress. This perspective accounts for substantial individual differences in stress reactivity and emphasizes that stress is fundamentally an interaction between the person, who engages in appraisal and subsequent responses, and the environment, which provides the stressors.

While the transactional model provides a rich theoretical foundation, building computational systems for stress management requires translating this concept into a concrete, measurable target for modeling. In the HCI and ubiquitous computing literature, researchers have primarily operationalized stress in three distinct ways, which constitute the conceptual framework adopted in this review (Figure 2). The first and most common approach conceptualizes stress as a set of internal physiological changes and external expressive behaviors that can be objectively measured. We refer to this as *Psycho-Physiological Stress*, emphasizing the bodily responses (e.g., heart rates, facial expressions, or voice tone) that manifest when individuals react to a stressful stimulus (i.e., either in the lab settings or real-life scenarios). A second approach, grounded in the notion of cognitive appraisal, defines stress as a perceived experience. This construct, which we call *Subjective Stress*, reflects how a situation is interpreted as threatening or challenging. It is most commonly measured through self-appraisal methods via self-reporting their stress levels (known as perceived stress), but can also be assessed through external evaluators, for example, when clinicians or observers evaluate an individual's stress state based on their observable responses. Finally, a third strategy operationalizes stress by an individual's exposure to a known stress-inducing context or task. This construct, termed *Causative Stress* or *Exposure Stress*, is not concerned with the user's reaction but rather with the presence of the stressor itself (e.g., staying in noisy environments or conducting stressful tasks, such as job interviewing).

To model these different constructs, researchers draw on a wide range of signals captured from sensors and devices. *Subjective indicators* capture an individual's perceived stress directly through self-reports, including validated questionnaires such as the Perceived Stress Scale (PSS) [16], in-the-moment ecological momentary assessments (EMA), or simple Likert-scale ratings prompted by a system. *Physiological indicators* reflect bodily changes related to the "fight-or-flight" response, in which the sympathetic nervous system (SNS) is activated and the parasympathetic nervous system (PNS) is suppressed [11]. These processes result in measurable changes such as increased heart rate, elevated blood pressure, and alterations in electrodermal activity, which are frequently monitored using wearable sensors [29, 81, 85]. Similarly, *Expression indicators* capture the outward bodily manifestation of stress, such as facial expressions that reveal an individual's feelings [22], or vocal cues including higher pitch or faster speaking rate [4, 80]. *Behavioral indicators* extend beyond direct expression and reflect the individual's adaptive or maladaptive changes in behaviors as part of stress coping. Behavioral indicators include variations in digital interactions (e.g., typing speed and error rates, and mouse movements), but also fluctuations in physical activities and mobility patterns captured by motion sensors [32, 73]. Finally, *Exposure indicators* differ from the previous indicators as they do not capture the individual's response, but rather the external environment or situation *demands* that may *trigger* stress. These indicators can include data from an individual's digital footprint, such as the number of unread emails or upcoming deadlines in a calendar, as well as physical surroundings measured by sensors, like ambient noise levels from a microphone.

2.2 Stress Management Systems in HCI

Building on the operationalization of stress into psycho-physiological, subjective, and exposure constructs, prior HCI work has explored various stress management systems. These systems function by detecting stress indicators and delivering targeted interventions. To achieve this, they leverage the diverse indicators detailed in the

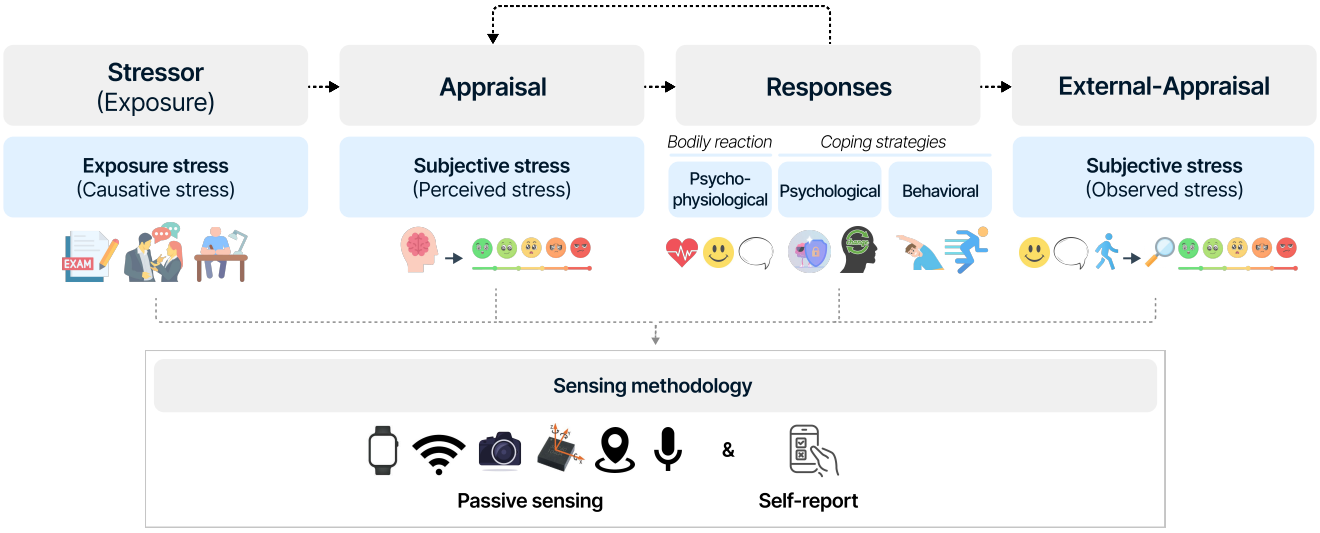

Figure 2: The conceptual framework used in this review. Building on the transactional model from Lazarus and Folkman [52], this framework operationalizes stress into three inference targets: (1) Stressor (Exposure), (2) Cognitive Appraisal (Subjective), and (3) Response (Psycho-physiological). It illustrates how stress is shaped through interpretation and coping, and how multimodal sensing methodologies, including passive sensing and self-report, can be used to detect stress across stages.

previous section: physiological, expressive, and behavioral indicators are captured to model a user's psycho-physiological responses; subjective indicators like self-reports are used to understand their cognitive appraisal; and contextual indicators help identify the presence of stressors. Once stress is detected through these channels, the system delivers interventions to mitigate its effects (Figure 3).

While the detection component is guided by these stress constructs, the design of the subsequent intervention represents a distinct challenge. Designing effective interventions requires careful consideration of both their timing and content. The design of intervention content has often been guided by the transactional model of stress, which describes stress processing in two stages: appraisal, the interpretation of the situation, and coping, the actions taken in response.

Within this framework, stress interventions can be designed to engage with either stage. Some systems target the appraisal stage, using biofeedback or data visualizations to help users reflect on and gain awareness of their current state [59, 110]. Others intervene directly in the coping stage by providing concrete coping strategies. These strategies are often categorized as problem-focused, which aims to resolve the source of the stress, and emotion-focused, which aims to manage the resulting emotions. The timing of these interventions can also vary, from just-in-time delivery upon stress detection to pre-scheduled or on-demand support. Some research in workplace contexts has even explored shared, ambient systems that support multiple users simultaneously [107, 108].

2.3 Positioning This Review

Although research on systems for stress detection and intervention has been active in HCI, prior surveys have largely treated stress detection and intervention as separate domains. This fragmentation limits the field's ability to conceptualize stress technologies as integrated systems, while also obscuring empirical insights into how sensing-inference-intervention configurations impact stress reduction in practice.

Reviews on stress detection have mainly concentrated on technological aspects of sensing, typically surveying methods that employ passive monitoring using wearable or contactless technologies [10, 27, 64]. In contrast, reviews on stress interventions have emphasized the efficacy of specific strategies within particular contexts, such as managing workplace stress [51], supporting university students [76], or promoting mindfulness in healthy individuals [86].

Only a few reviews have attempted to bridge this gap between detection and intervention, yet their scope has been limited. For example, Dobson et al. examined sensor-based systems delivering in-the-moment interventions for anxiety, with limited discussion of diverse stress intervention strategies [20]. Another review focused on commercial products without systematically analyzing user studies or evaluation practices [95]. As a result, the field still lacks a perspective that jointly considers stress sensing methods, intervention design, and evaluation practices in real-world contexts.

To address this gap, our review provides a holistic analysis of integrated stress management systems, examining the entire pipeline from detection to intervention. We introduce a framework that organizes the literature according to how stress is defined and operationalized, namely as psycho-physiological, perceived, or causative stress. Using this framework, we systematically analyze the connections between detection indicators, the design and timing of interventions, and the evaluation methods employed in user studies.

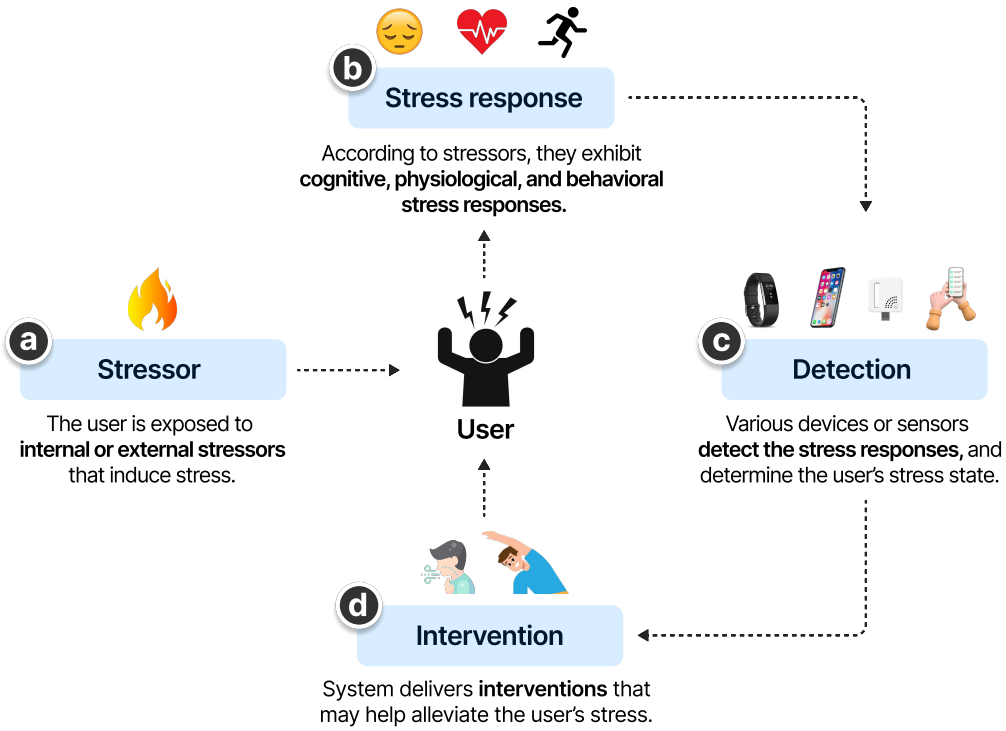

Figure 3: Overview of stress intervention system. (a) Users are exposed to internal or external stressors that induce stress. (b) These stressors evoke cognitive, physiological, and behavioral stress responses. (c) Various devices and sensors detect these responses to determine the user's stress state. (d) Based on detection results, the system delivers appropriate interventions that may help alleviate the user's stress.

In doing so, this review offers a comprehensive map of the landscape, identifies key trends and challenges, and outlines directions for future research in proactive stress management technologies.

3 Methods

In this review, we focused on research articles that *implemented interventions informed by stress tracking results*. To identify relevant literature, we followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Figure 4). We conducted a systematic search in three major academic databases, including the ACM Digital Library, IEEE Xplore, and Web of Science. The review included studies published up to July 2025, with no lower bound applied.

To ensure comprehensive coverage, search terms were selected based on an initial scoping survey and refined through discussions among all authors. The keywords were organized into three groups (Stress, Detection, and Intervention). These groups were connected with `AND`, while keywords within each group were connected with `OR`. We applied these search queries to full-text fields searchable in each database.

- **Stress**: stress, distress, eustress
- **Detection**: stress detection, stress sensing, stress classification, stress monitoring, self-tracking, personal tracking, personal informatics, lifelogging, quantified self
- **Intervention**: intervention, coping, stress management, stress reduction, mindfulness, emotion regulation, emotional support

A goal of this review is to understand the *holistic mechanism* of how sensing technologies translate into effective stress management in real-world contexts. In HCI research, evaluating health behavior change technologies requires looking "beyond efficacy" to investigate usage patterns, user perceptions, and how different system components interact [47]. Therefore, we treated evaluation not merely as a validation of clinical outcomes, but as an essential component to verify the utility of the sensing-intervention loop.

We did not include design-only studies within the scope of this review because, without empirical user feedback, it is difficult to assess whether the sensing mechanism effectively supported the intervention and whether the timing of the intervention was receptive to users. We acknowledge that some technology-oriented works publish system designs and evaluations separately. However, this review focuses on *integrated systems* presented within a single

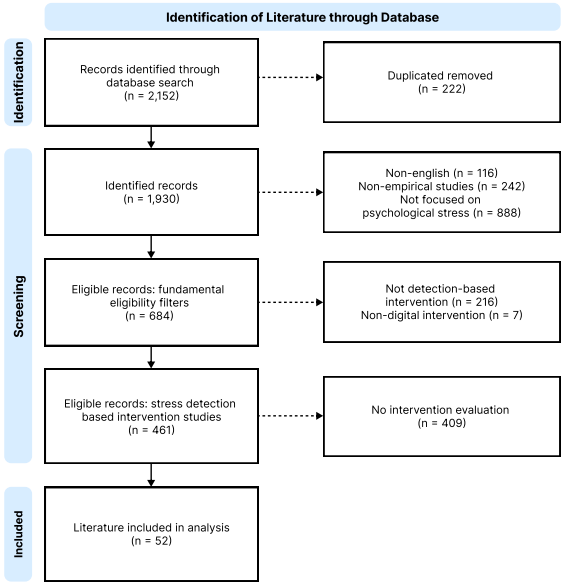

Figure 4: PRISMA Flow diagram of the literature selection process for the scoping review.

study. This scoping decision ensures that the sensing and intervention components were designed and evaluated as a cohesive unit, allowing us to analyze the immediate interplay between detection capabilities and user engagement. Based on this scope, we explicitly applied the following criteria:

Inclusion criteria:

- Studies implementing a digital intervention related to psychological stress
- Interventions informed by sensing or detection results
- Papers reporting a user study or formal evaluation of the intervention (qualitative, quantitative, or mixed-methods)

Exclusion criteria:

- Non-English or non-empirical publications
- Studies not focused on psychological stress
- Implementations lacking a detection-based intervention
- Non-digital interventions
- Papers without a user study or evaluation

The initial database search yielded 2,152 records. After removing 222 duplicates, 1,930 records remained for screening. Title and abstract screening excluded 1,246 articles (non-English: $n = 116$; non-empirical: $n = 242$; not focused on psychological stress: $n = 888$). Full-text assessment of the remaining 684 articles excluded an additional 223 articles (no detection-based intervention: $n = 216$; non-digital interventions: $n = 7$). Finally, from the 461 eligible articles, 409 lacked a user study or formal evaluation and were excluded. A total of 52 articles met all criteria and were included in the final analysis.

4 Results

In this section, we first provide an overview of the collected studies, summarizing their publication trends and application contexts. We then examine how systems have defined stress as an inference target and the indicators used (RQ1), investigate how these inferences connect to intervention design and delivery (RQ2), and explore how the systems have been evaluated in practice (RQ3).

4.1 Overview of the Reviewed Papers

Publication Trends. Figure 5 illustrates the publication years of the 52 selected papers. Interest in data-driven stress management systems has steadily increased, particularly from 2020 to July 2025. The number of publications peaked in 2024 with ten papers, and at least one paper has appeared every year since 2009.

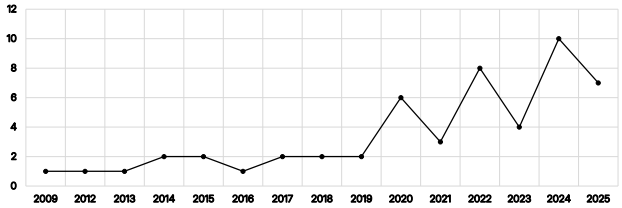

Figure 5: A time-series chart shows the number of studies in our corpus from 2009 to July 2025.

Stress Contexts. Across the 52 studies, we identified six primary contexts: everyday life, workplace, laboratory-induced stress, parenting and caregiving, driving and commuting, and academic settings (Table 1). Everyday life stress was most common, often centered on students' routines and wellbeing[38, 45], with some work extending to military personnel [105] and close relationships [36]. Workplace stress was the second largest category, spanning office work [107, 108], remote work [33], and high-pressure professions such as healthcare and emergency response [5, 34], highlighting tensions between productivity and wellbeing. Lab-induced stress was elicited through controlled tasks, ensuring reliable ground truth labeling. Typical tasks included socially stressful interactions [82], virtual job interviews [2], Stroop or mental arithmetic tests [7, 101], and game-based stressors [68, 69]. Parenting and caregiving contexts, though less examined, highlighted relational and distributed stress across families and care teams. Studies explored stay-at-home mothers managing childcare [37, 89], parent-child learning interactions [102], school-based autism support [41], and caregiving for individuals with intellectual disabilities or dementia [30, 48]. Driving and commuting stress was investigated in both simulations and real-world contexts, including controlled driving tasks [58, 59], long-distance driving [98], and everyday commuting scenarios [26]. Finally, academic stress was examined in a single study, which focused on classroom and learning settings [9].

4.2 Stress Tracking Methods

To answer RQ1, we analyzed how each study defined stress as an inference target, measured stress indicators using different sensing modalities, and modeled these indicators to detect stress.

Table 1: An overview of contexts for stress tracking-based intervention systems.

Context	Number of Studies	Percentage (%)
Everyday life stress	18	35
Workplace stress	13	25
Lab-induced stress	8	15
Parenting/Caregiving stress	7	14
Driving/Commuting stress	4	8
Academic stress	1	2

4.2.1 Stress Constructs as Inference Targets. Our analysis reveals that prior work has predominantly operationalized stress into two core constructs: Psycho-Physiological Stress and Subjective Stress. Of the 52 studies reviewed, 22 targeted Psycho-Physiological Stress and 20 focused on Subjective Stress. Exposure Stress was a less common target, appearing in 7 studies. There were also hybrid approaches that integrated multiple constructs, found in 3 studies.

Psycho-Physiological Stress defines stress as the body's direct, measurable reaction. Within this construct, two main inference targets appeared: *Physiological Stress*, where stress was treated as changes in biosignals (e.g., EDA, HRV), and *Affective Stress*, where stress was inferred from facial expressions. In most systems, physiological stress was defined directly from sensor data using threshold-based rules. A notable exception is the study by Verma et al. [98]. In their work, physiological data were used as the ground truth to define high-stress states in a driving scenario. Their system inferred stress by analyzing behavioral indicators (e.g., speeding, hard braking) and exposure indicators of the driving environment (e.g., vehicle speed, congestion levels, road types).

Only one study targeted affective stress, operationalizing it through facial expression analysis. In this case, stress was defined by observable affective cues (e.g., eyebrow distance, emotion categories such as 'anger' or 'neutral'). [60]

Subjective Stress defines stress as an individual's perception and appraisal of their own state. In most cases, the inference target is the user's self-report, though a few studies also relied on external appraisals provided by family caregivers or clinical staff [30, 48]. Our analysis shows that this construct was operationalized in three main ways. First, some studies either mapped stress levels directly from reported values or derived them by applying simple thresholds (e.g., stressed vs. not stressed) [38, 65]. Second, in other cases, subjective reports served as ground truth labels for modeling. The target was the self-reported stress state, while the indicators were drawn from diverse data sources such as behavioral and exposure signals (e.g., activity, workload, social setting) [15, 24] or physiological measures [30]. Third, nine systems elicited subjective reports in response to specific probes. In these studies, physiological or exposure indicators functioned as triggers for collecting subjective stress. For example, stress EMAs were administered at the time of calendar event registration to record anticipated stress [53].

Exposure Stress was the least common construct, appearing in seven studies. In this category, stress was defined directly through exposure to external stimuli (e.g., simulated job interviews) [2]. Most studies utilized physiological indicators to infer exposure stress states (e.g., increased heart rate while involving in interviewing) [2, 70], typically under controlled lab settings. An exception was the work of Jo et al. [37], which adopted a personal informatics perspective. In their study, exposure stress was self-tracked by participants across contextual dimensions such as activity, social relationships, and time of day. Rather than inferring stress, the system supported users in reflecting on these contextual indicators to better understand their own stress and its situational correlates [37].

Hybrid approaches were found in three studies. One system combined psycho-physiological stress and exposure stress, computing a stress score by averaging five normalized signals (i.e., email volume, calendar density, time of day, negative-positive facial expression activity, and heart rate) [33]. Two additional studies combined subjective and physiological constructs. Ren et al. [78] directly mapped both indicators for visualization purposes. Wakschlag et al. [100] applied a rule-based scheme. A day was marked as high-risk for stress if either subjective stress or physiological stress was elevated. This approach primarily addressed missing data.

4.2.2 Stress Indicators. While stress constructs define what counts as stress (i.e., the inference target), stress indicators specify how it was measured. Stress tracking systems capture different facets of stress indicators, which can be grouped into four primary dimensions: physiological (e.g., PPG/EDA), expression (e.g., facial/vocal expression), subjective (e.g., self-reported stress), and behavioral/exposure (e.g., work hours, noise levels). These indicators can be directly used to define stress for intervention delivery or to indirectly infer stress (e.g., using behavioral indicators to infer subjective stress, or using physiological indicators to infer exposure stress). Furthermore, it is possible to track multiple stress indicators; a total of 25 studies used two or more indicators (e.g., physiological with perceived) to provide a more robust operationalization of stress constructs. Table 2 summarizes the distribution of studies across these categories.

To illustrate how these categories have been operationalized, we describe each in turn with representative data sources, raw signals, and derived features. Table 3 provides a structured overview.

Subjective indicators relied on participants' subjective reports of stress, most commonly captured through ecological momentary assessment (EMA) or short mobile questionnaires. These reports reflected stress intensity and emotions, and were often used as standalone detection targets or as ground truth for validating sensor-based approaches [38, 45, 90].

Physiological indicators are the most prevalent and primarily focused on autonomic arousal. Wearable devices such as wristbands and smartwatches (e.g., Garmin, Empatica, Fossil) were commonly used to capture physiological signals, including ECG, PPG, or EDA, from which features such as heart rate, heart rate variability, and skin conductance were derived [36, 58, 82]. Chest-worn devices collected ECG and respiration [82, 107], while fNIRS and EEG headbands were used in laboratory studies to measure hemodynamic or brain activity [7, 101]. Other form factors included sensors clipped to the ear [26] or worn on the hand and fingers [93, 102]. Across these modalities, derived features largely served arousal estimation (e.g., HRV indices, skin conductance levels, respiration rate).

Expression indicators, though less common, represent expressive cues of stress. Webcams, facial landmark analysis, and video-based

Table 2: Distribution of papers by stress indicators.

Stress Indicators					
Subjective	**Physiological**	**Expression**	**Behavioral / Exposure**	**Papers**	**Total**
✓				[34, 55, 65, 90, 112]	5
	✓			[5, 9, 26, 31, 36, 41, 49, 58, 68, 69, 77, 82, 93, 99, 101, 102, 107–110]	20
		✓		[60]	1
			✓	[37]	1
✓	✓			[30, 54, 62, 63, 78, 89, 100, 105]	8
✓		✓		[48]	1
✓			✓	[15, 24, 38, 39, 45, 53]	6
✓	✓		✓	[17]	1
	✓		✓	[2, 6, 7, 21, 59, 70, 87, 98]	8
	✓	✓	✓	[33]	1

emotion recognition were used to capture stress-related affective states [33, 60], while one study employed acoustic features of speech (e.g., tone, speed, tenor) [48]. These approaches derived features such as facial expression dynamics or prosodic variations.

Behavioral and exposure indicators capture stress-related changes in daily routines and technology use or situational demands that trigger stress. Smartphones provided data such as GPS traces, accelerometer readings, app usage, and call logs [45], while wearable sensors such as shoe-mounted IMUs tracked movement patterns [24]. In workplace and driving contexts, stress was inferred from computer logging (e.g., email, calendar, keyboard/mouse activity) [33] or vehicle-mounted sensors (e.g., speed, acceleration, braking) [98]. Features derived from these sources emphasized routines, workload, mobility, and activity patterns. In some cases, contextual information such as activity type or location was collected via self-reports rather than sensors [37–39]. In addition, behavioral data were sometimes used not as direct indicators of stress, but to provide situational context (e.g., step counts or physical activity levels [17, 87]) or to filter out motion artifacts. For example, accelerometer data were used to identify periods of intense movement that could distort physiological signals [49, 63, 105].

Taken together, these approaches demonstrate how diverse modalities, including physiological, behavioral, perceived, and affective features, have been leveraged to operationalize stress detection.

4.2.3 Sensing Devices. As shown in the table 4, most studies relied on commercial wearable devices such as Empatica, Garmin, Samsung, and Shimmer, with only a limited number utilizing custom-built sensors. Specialized companies, such as Empatica, Shimmer, Plux, BioStamp, and Thought Technology, fulfilled niche demands for specific physiological signals, particularly in research contexts. Notably, devices designed for research purposes generally retailed for over $1,000, reflecting the high costs associated with advanced sensing capabilities.

Mainstream consumer brands, such as Garmin, Samsung, and Fossil, were also selected, predominantly for smartwatch-based sensing, with prices typically under $500. This pricing and positioning difference highlights a clear divide between research-oriented manufacturers and consumer electronics companies. Wrist-based devices were the dominant choice; however, other locations such as the chest (ECG bands/patches) and ear (PPG clips) remained essential for capturing specific physiological signals. In particular, chest-worn ECG bands were notable for their relatively low cost, making them a popular option for heart signal acquisition.

In summary, the landscape of wearable device selection in research demonstrates distinct segmentation based on device purpose and cost, with high-priced research models serving specialized needs, and cost-effective consumer wearables supporting broader usage. Signal type and sensor placement further define device choice for stress sensing in this field.

4.2.4 Stress State Inference. Stress indicators can be directly used as measures for stress (e.g., subjective or exposure indicators) or indirectly used to infer stress states via machine learning. Out of the 52 papers, 41 developed modeling approaches using rule-based or machine learning methods, typically relying on ground truth labels from controlled stimuli, self-reports, or physiological thresholds. The 11 studies emphasized descriptive reporting or visualization without explicit classification [5, 26, 34, 37, 39, 53, 55, 70, 78, 93, 108]. Below, we describe how ground truth labels were obtained and summarize how stress states were modeled based on these labels.

Ground Truth. Rule-based or machine learning methods require ground-truth data for configuring rules or training machine learning models. For classifying stress states from collected features, prior studies obtained ground truth labels using three common strategies: stimulus-driven, subjective, and physiological response labeling. Stimulus-driven labeling marks task segments intended to elicit stress, subjective labeling typically operationalizes perceived or observed stress, and physiological labeling captures responses to a stress stimulus.

Stimulus-driven labeling is based on the concept of '*exposure stress*' (or causative stress). Ground truth labels are given when an individual is exposed to a stress stimulus. Traditionally, stimulus-driven labeling follows strict protocol-defined task segments (e.g., baseline: no stress vs. induction: stress). Widely used stress-induction protocols include the Trier Social Stress Test (TSST), a computer version of the Paced Auditory Serial Addition Task (PASAT-C), the Stroop color-word test, mental-arithmetic tasks, and the Cold Pressor Test (CPT). In these designs, baseline and induction segments

Table 3: Overview of stress indicators with representative data sources, raw signals, and derived features. Derived features are grouped by related stress constructs: subjective experience (perceived stress level, emotions, anticipation), physiological activation (e.g., HRV, EDA, respiration, thermal and neural signals), and behavioral or exposure context (e.g., activity, workload, driving behavior, pose). Abbreviations are listed and explained in Appendix A Table 8.

Stress Indicators	Data Source	Raw Signals	Derived Features	Paper
Subjective Indicator	Mobile	Self-report	Perceived stress level	[24, 34, 38, 39, 45, 54, 55, 63, 65, 100, 112]
			Subjective emotions	[62, 89, 90]
			Anticipated stress level	[53]
	Wearable		Perceived stress level	[87]
	PC		Perceived stress level	[17]
	External report	Labeling from caregivers	Perceived stress level	[30]
			Subjective emotions	[48]
Physiological Indicator	Wearable	ECG	HR/HRV	[6, 26, 30, 41, 68, 69, 82, 100, 107]
		EDA	SCL/SCR	[2, 5, 6, 9, 17, 21, 30, 49, 58, 68–70, 78, 87, 93]
		PPG/BVP	HR/HRV	[2, 17, 21, 26, 36, 41, 54, 62, 63, 78, 87, 89, 99, 102, 105, 108, 110]
		RESP	RESP	[6, 68, 69, 93]
		SkinTemp	Thermal features	[21, 87]
		BP	BP features	[6, 93]
		EEG	EEG	[101]
		fNIRS (HBO, HHB)	HR/HRV	[7]
	Camera	Video signal (remote PPG)	HR/HRV	[33]
		SkinTemp	Thermal features	[109]
	Vehicle-embedded sensor	EDA	SCL/SCR	[59]
	Hand-held biosensor			[31]
	Chair-embedded sensor	BCG	HR/HRV	[77]
Expression Indicator	Camera	Video signal	Facial expression	[33]
	Microphone	Acoustic signal	Voice tone, Tenor, Speed	[48]
Behavioral & Exposure Indicator	Mobile	Self-report	Activity, Location, Social setting	[37–39, 45]
			Time	[37–39]
			Sleep, Phone usage	[45]
			Calendar event	[53]
	Wearable	Motion (ACC, Gyro)	Movement	[24, 87]
	PC	Computer/OS logging data	Workload	[33]
	Hospital information system	Patients per category, Supervision status		[15]
	Vehicle-embedded sensor	GPS, IMU	Speed, Location, Acceleration/deacceleration	[98]
		Pressure	Pose movement	[59]

are directly labeled as "not stressed" and "stressed" [2, 82, 101]. Similarly, Dongre et al. [21] used the standardized WESAD dataset [71], which provides protocol-defined segments (e.g., public-speaking and mental-arithmetic blocks).

Beyond these standard lab protocols, Madrid et al. used driving-simulator conditions as labels [59]. Baseline was obstacle-free or low-difficulty car-following; stress was complex urban driving with adverse weather or limited visibility, pedestrian/vehicle hazards, phone interruption, and concurrent mental arithmetic. Exposure stress can be extended beyond lab settings; for example, in the workplace setting, Howe et al. [33] defined cognitive workloads (e.g., emails, meetings) or Park et al. [66] used emotional workloads (e.g., emotionally demanding customer calls).

Subjective labeling is based on the concept of '*subjective stress*,' which can be self-reported by individuals (i.e., reporting perceived stress levels) or reported by external observers (i.e., reporting observed stress levels). In lab settings, perceived stress was recorded immediately after each protocol-defined segment or concurrently during tasks. For example, Yun et al. [109] collected minute-by-minute in-game ratings and a post-game survey. Elvitigala et al. [24] used a dataset from a prior study. The dataset [23] induced stress using Stroop, Minesweeper, and mental arithmetic tasks, and induced relaxation using introductory or nature videos. After each segment, participants reported stress on a 7-point Likert scale, and these labels were used to train a classifier.

In everyday life, self-reports were collected using the Experience Sampling Method (ESM), also known as Ecological Momentary Assessment (EMA). ESM schedules varied across studies. For example, schedules included prompts every four hours [45], approximately every 1 to 1.5 hours [38, 39], and every 30 minutes during work

Table 4: Summary of wearable devices used for stress sensing, categorized by device type (Commercial, Custom, Etc.), body location, device model, price range, and number of papers citing their usage. Price ranges were categorized into three ranges: below $500, $500–$1,000, and above $1,000.

Type	Body Location	Device	Price	Paper	Total
Commercial	Wrist	Empatica E4	> $1,000	[2, 5, 9, 17, 78]	5
		Empatica EmbracePlus	> $1,000	[21]	1
		Garmin (VivoActive, VivoSmart, Venu)	< $500	[36, 89, 99, 105]	4
		Fossil Sport	< $500	[63]	1
		Samsung Galaxy Watch 6	< $500	[62]	1
		Not Specified (Android Wear 2.0)	N/A	[54]	1
		Philips DTI-2	N/A	[49]	1
		imec Chill+	N/A	[87]	1
		Mio Fuse	< $500	[41]	1
	Hand	Shimmer EDA sensor	$500 - $1,000	[68]	1
		Thought Technology FlexComp Infiniti	> $1,000	[69]	1
	Finger	Thought Technology Biograph infiniti	> $1,000	[93]	1
	Ear	Shimmer 3 PPG optical pulse ear-clip	$500 - $1,000	[26]	1
	Head	Macrotellect BrainLink	< $500	[101]	1
	Chest	Polar H7	< $500	[41]	1
		Zephyr Bioharness BT	< $500	[68, 69]	2
		BioStamp	> $1,000	[100]	1
		Aidlab	< $500	[107]	1
	Torso	Shimmer 3 ECG	$500 – $1,000	[26]	1
		Movesense ECG patch	< $500	[30]	1
		Plux wireless biosignal toolkit	> $1,000	[82]	1
		Thought Technology Biograph infiniti	> $1,000	[93]	1
	Shoe	MBIENTLAB IMU	< $500	[24]	1
	Foot	Metafas SentiSock	N/A	[30]	1
Custom	Finger	Custom-built sensor	N/A	[6, 70, 102, 110]	4
	Arm		N/A	[58]	1
	Head		N/A	[7]	1
Etc	-	Not specified	N/A	[6]	1

hours (about ten per day) [110]. Response formats were Likert-type, most commonly 5-point [34, 37–39, 53]. Three-point [45] and seven-point variants [23] were also used. One study used a standardized questionnaire, the Perceived Stress Scale (PSS) [112]. When external observers are involved in rating, they can do in-situ labeling or post-hoc labeling (or retrospective affect judgment) [13]. For example, caregivers provided proxy labels by annotating stress and relaxation events [30].

Physiological response labeling refers to using physiological signals as the ground truth for stress, under the assumption that they reliably reflect stress responses to external stimuli. In most studies, these signals themselves were treated as the target. In a smaller number of cases, physiological responses served as ground truth against which models inferred stress from other modalities such as behavioral or expressive indicators [3, 98].

This approach is implemented through *physiological thresholds*, either as fixed cutoffs or as personalized baselines via per-participant calibration. For fixed cutoffs, Xue et al. [107] labeled stress when ECG-derived HRV fell below a predefined threshold, specifically RMSSD <50 ms. Personalized thresholds were operationalized in two main ways. First, baseline-referenced methods established per-participant baselines and defined stress as deviations from these baselines; for example, relative changes in EDA, HRV, and breathing rate after paced breathing [68, 69], absolute distance from a three-minute EDA baseline [58], and child-specific heart rate zones calibrated from resting heart rate [41]. Second, distribution-based methods converted physiological signals into categorical stress levels by segmenting each participant's data distribution. For example, skin-conductance values were divided into ordinal arousal levels [49], and algorithmic discretization methods were applied to map raw GSR signals into 0–5 stress levels [9].

Modeling. Most studies classified stress as a binary state, distinguishing between stressed and not stressed [24, 58, 107, 110]. A minority adopted multi-level targets. For example, Kim et al. [45] used three ordinal categories (Low, Moderately High, High), and other work used four-level ordinal labels (no stress, low stress, medium stress, high stress) [99].

For detection, we observed two categories of approaches. First, data-driven models ranged from simple statistical methods to classic machine learning and deep learning models. Some studies used statistical or regression models, such as multiple regression with a logit link [15]. Classic classifiers such as LDA [24], kNN [6, 7], SVM [6], decision trees [59], random forests [2, 38], and gradient-boosted trees (XGBoost) [45] were trained on multimodal features from wearables, phones, and context logs. Deep learning models, including one-dimensional CNNs [21] and multi-task learning networks [98], were also employed. Across these studies, labels were sourced from self-reports, stimulus protocols, or physiological thresholds, depending on the research design.

When these models were evaluated, accuracy and F1 were common metrics [24, 45], and AUC was also reported in some work [26, 98]. Reported performance for binary classification typically clustered around the mid 80% range. For example, Kim et al. reported

84.3% accuracy and an 84.1% F1-score for an LDA classifier [45], Dongre et al. achieved 85.1% accuracy and an 89.0% F1-score with a 1D CNN on the WESAD dataset [21], and other settings reported AUCs spanning from approximately 0.73 to 0.91 [98].

Second, rule-based detectors inferred stress by applying explicit cutoffs to physiological features. These deterministic rules were either fixed or personalized. Fixed, population-level rules used preset limits without per-participant calibration. For instance, classifying stress based on HRV SDNN falling below 50 ms [107], using the LF/HF ratio from three-minute HRV windows [110], or mapping Garmin's proprietary stress score to predefined bands (e.g., 0–25 for no stress, 26–50 for low stress) [99].

Personalized rules, in contrast, are adapted to individual baselines. For instance, one study flagged stress when a participant's EDA dipped below their personal three-minute baseline for more than two seconds [58]. Another computed a user-specific stress score by averaging five normalized indicators (email volume, calendar density, time of day, facial expression, heart rate) referenced to a first-week baseline [33]. Other personalized approaches involved calibrating an individual's HRV range during dedicated relaxation and activity periods [77] or flagging stress when RMSSD dropped below 80% of a user's baseline [102].

Evaluation of these deterministic detectors was often limited. Some papers relied on correlation with self-reports. Howe et al. [33] validated their score using Pearson correlation ($r = .20$, $p < .01$) rather than classification metrics. Zhang et al. [110] also analyzed the correlation between stress and sleep stage (lower stress during deep sleep, higher during REM sleep).

Finally, in addition to detectors that aimed to produce direct stress classifications, many systems leveraged trigger-based or hybrid workflows. Automated detection served not as the final output but as a mechanism to prompt or validate self-reports, effectively coupling objective physiological events with in-the-moment subjective annotations. For example, Song et al. used a threshold (top 25th percentile of the previous day's HRV-derived stress) to detect high-stress moments and trigger a request for a self-report annotation [89]. Other systems used physiological triggers to prompt users to confirm and label the context of a potential stress event [54, 87]. Several systems also consumed predictions from pre-trained or proprietary black-box models as triggers for user labels [62, 63].

Human-in-the-loop Model Calibration. While the majority of studies relied on static ground truth labels, feedback was utilized in some works to refine models through calibration. For instance, Kim et al. [45] implemented a human-in-the-loop approach that allowed users to adjust the system's stress predictions. These user corrections served as dynamic labels to retrain the XGBoost model, thereby personalizing the inference engine over time.

4.3 Intervention Strategies

To address RQ2, we examined intervention strategies across three dimensions: the delivery platforms, the intervention contents targeting stress constructs, and the timing of intervention delivery.

4.3.1 Intervention Delivery Platform. Intervention delivery can happen via diverse platforms, such as mobile, wearable, desktop, public display, IoT, and AR/VR devices, and user interactions range from graphical user interfaces to haptic and conversational interactions. Across the reviewed systems, mobile applications emerged as the dominant intervention platform, often integrated with wearables such as smartwatches or physiological sensors [34, 54, 55]. Mobile apps enabled just-in-time delivery of feedback, visualization, and personalized coping recommendations in daily life settings [45, 53, 110]. Wearables, including custom wristbands and commercial smartwatches, were frequently used to provide continuous sensing and biofeedback through haptic or visual cues [7, 31, 58].

Beyond mobile and wearable platforms, several studies explored alternative modalities tailored to specific contexts. Desktop and web systems were commonly used in workplace and lab-induced settings [15, 33, 49, 93]. In workplace contexts, shared public displays were also employed to externalize stress data collectively and support team reflection [5, 107, 108]. Emerging modalities included LLM-based chatbots that provide conversational coaching [21, 62], as well as embodied and IoT-based environments (e.g., social robots, ambient lighting systems) [60, 102]. VR applications were also used to create immersive environments for reflection [99].

4.3.2 Intervention Content. To understand how intervention content has been designed in prior work, we organized the strategies reported in the literature into several broad categories. To ensure a rigorous classification, we employed an inductive analysis approach using affinity diagramming. Three researchers independently extracted intervention descriptions from the selected papers and iteratively clustered them based on functional similarity. Any discrepancies in categorization were resolved through discussion between the researchers until full consensus was reached. We identified eight categories across the literature: Visualization and Externalization, Biofeedback for Physiological Awareness, Cognitive Therapy and Reframing, Mindfulness and Relaxation Techniques, Social Support (Emotional), Activity Suggestions and Proactive Planning, Social and Organizational Support (Instrumental), and Environmental and Task Adaptation. This classification serves as an analytical tool rather than a rigid taxonomy; the boundaries are not mutually exclusive, and many interventions combine elements of both stages.

Before examining each category in detail, an overview of how intervention contents have been distributed across stress constructs is presented in Figure 6. For psycho-physiological stress, intervention systems relied predominantly on visualization, with biofeedback being the next most common strategy, while coping strategies like Cognitive Behavioral Therapy (CBT) were used only to a limited extent. While visualization also remained the most common intervention for subjective stress, these systems more frequently employed CBT, mindfulness, and activity planning. Although hybrid pipelines were relatively rare, they all included more than one type of intervention content; for example, one study combined physiological and exposure signals to deliver CBT, mindfulness, and activity suggestions [33]. Across categories, visualization and externalization emerged as the most common intervention strategy, whereas emotional social support appeared only in niche contexts such as family or parenting [37, 89].

To systematically categorize the landscape of stress interventions, we draw on Lazarus and Folkman's transactional model of stress and coping [52]. In this model, stress arises when an external

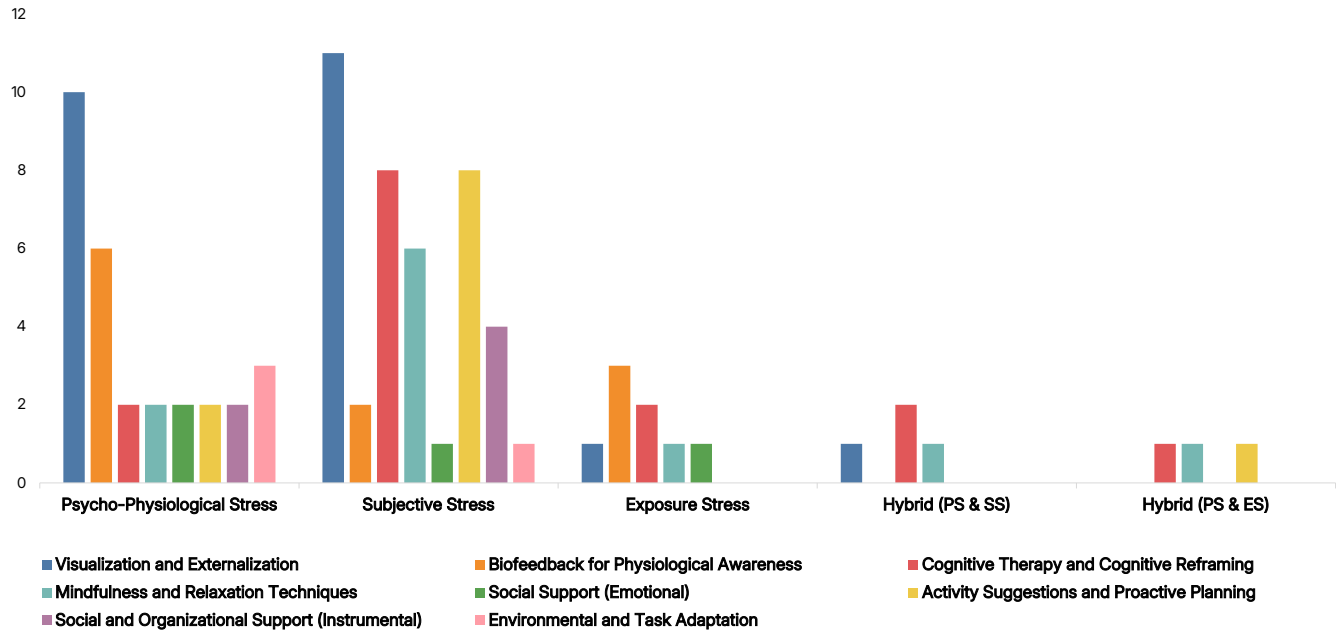

Figure 6: Distribution of Intervention Content by Stress Construct.

stressor is subjectively appraised, leading to subsequent responses that include both bodily reactions and coping strategies. Following this perspective, we categorize interventions into two broad groups: appraisal-focused approaches that primarily target the *appraisal stage*, enhancing awareness and subjective interpretation of stress, and coping-focused approaches that act at the *response stage*, guiding users toward concrete coping strategies. An overview of how the identified intervention categories are distributed across these two groups is provided in Table 5.

Appraisal-focused approaches. The appraisal-focused approach includes interventions that primarily support reflection or awareness without directly guiding users toward a specific coping action. Among the eight categories identified above, two correspond to this group: *Visualization and Externalization* and *Biofeedback for Physiological Awareness*.

Visualization and Externalization. These interventions foster awareness by displaying stress data through dashboards [24, 41, 45, 63] or sharing displays with others [36, 107, 108]. Other forms of externalization include physical interfaces that manifest a user's real-time affective state [58] and immersive VR environments that visualize stress patterns [99].

Biofeedback for Physiological Awareness. Although biofeedback has traditionally been associated with coping, the HCI systems in our review [31, 55] often utilize biofeedback to foster interoceptive awareness [83, 91], which is the ability to sense, interpret, and integrate internal body signals, such as heart rate and emotion. Strengthening interoceptive awareness improves the mind-body connection, emotional regulation, mental and physical health, and promotes mindful decision-making [42]. These interventions provide real-time feedback on physiological signals, such as through haptic vibrations [7, 31, 59] or visual cues like graphical displays [55, 59, 70] or ambient light [77]. For instance, Zhang et al. introduced a respiration-based biofeedback system that dynamically adjusted breathing frequency based on its stress-alleviation effect [110]. The core technical contribution lies in making invisible physiological signals perceptible, thereby enhancing the user's self-awareness and understanding of their body's response to stress.

Coping-focused approaches. In contrast, a coping-focused approach comprises interventions that move beyond reflection and actively guide users toward stress-coping behaviors. Following Lazarus and Folkman's framework [52], we distinguish between *emotion-focused coping*, which regulates the emotional response to stress, and *problem-focused coping*, which targets the source of stress or its situational demands. Within each type, we identified several categories of interventions.

Emotion-Focused Coping. These interventions primarily help users regulate emotions and reframe stressful experiences, fostering emotional regulation skills and encouraging more constructive interpretations of stressful situations.

Cognitive Therapy and Reframing. This category includes interventions grounded in principles from Cognitive Behavioral Therapy (CBT), positive psychology, and meta-cognitive strategies, which guide users in actively modifying negative thought patterns [65]. One prominent type involves guided cognitive exercises, which are often delivered as prompts or structured interfaces. These techniques range from positive self-talk reminders [24], cognitive reframing, and positive reinterpretation [33, 53], to guided practices such as journaling [62, 105]. For instance, recent work by Zhao et al. [112] introduced a guided gratitude journaling system that

Table 5: Overview of distribution of intervention contents across categories.

Stage	Type	Sub-Category	Example Interventions
Appraisal (Stress perception and interpretation)	Awareness / Appraisal-Focused	Visualization & Externalization	Dashboards, shared displays, physical interfaces, immersive VR visualization
		Biofeedback for Physiological Awareness	Haptic feedback, visual feedback, socially guided biofeedback
Response (Coping and stress management)	Emotion-Focused Coping	Cognitive Therapy & Reframing	CBT-based exercises, positive self-talk, cognitive reframing, journaling, LLM-based counseling, VR-based exposure training
		Mindfulness & Relaxation	Guided breathing, PMR, mindfulness meditation, sensory relaxation tools (e.g., warm stone)
		Social Support (Emotional)	Peer/family/partner connections, emotional disclosure, group workshops
	Problem-Focused Coping	Activity Suggestions & Proactive Planning	Micro-interventions (music, walking), habit formation, counterfactual/context planning
		Social/Organizational Support (Instrumental)	Supervisor/teacher workload adjustment, team-based resource management, task allocation optimization
		Environmental & Task Adaptation	Notification control, adaptive lighting/music, task or game difficulty adjustment

scaffolded writing through structured prompts and employed generative AI to automatically create illustrative diary cards from user entries, providing a multi-modal and shareable form of positive reflection.

A second emerging type specifically leverages Large Language Models (LLMs) to power conversational agents and chatbots that deliver CBT-based dialogue and psychotherapy [21, 34, 62]. Finally, leveraging the behavioral component of CBT, some systems focus on skill training and exposure, using technologies like VR to create simulated environments where users can practice and rehearse coping skills, such as for job interviews [2].

Mindfulness and Relaxation Techniques. This category consists of interventions that guide users through specific physical or mental exercises to actively regulate their stress response. It includes structured practices like guided deep breathing, Progressive Muscle Relaxation (PMR), and mindfulness-based interventions such as body scan meditation [48, 105]. For example, one system employed an interactive social agent that provided instructions, managed the training flow, and delivered verbal feedback to support techniques such as conscious breathing [82]. This category also encompasses sensory relaxation tools, such as a warm stone [6].

Social Support (Emotional). Interventions in this category mediate emotional social support by prompting peers, family members, or caregivers to connect with the user in stressful moments [30, 89, 102], or by fostering group reflection and workshops [5, 37]. In addition, one system delivered empathic responses directly through a social robot [60].

Problem-Focused Coping. These interventions directly address the causes or situational demands of stress by helping users plan, adapt, or restructure their environment.

Activity Suggestions and Proactive Planning. This category provides actionable guidance that can be divided into two types: immediate suggestions for in-the-moment relief and proactive tools for future planning. The first type, immediate activity suggestions, refers to micro-interventions recommending short breaks or diversions. This includes offering mood boosters like music or short videos [6, 24, 33], or encouraging physical activities like a walk based on behavioral data [24, 45]. Some systems deliver these suggestions in a personalized manner, offering location-based tips or using companion robots to provide timely advice [60, 90]. The second type, proactive planning, helps users manage future stress by preparing for events or adjusting contexts [53]. For instance, some systems help users form proactive habits by setting implementation intentions ("if-then" plans) [55], while others provide data-driven, counterfactual suggestions for altering future contexts based on historical data [38]. Note that these activity suggestions can serve both as emotion-focused coping (i.e., to manage the emotional response to stress) or, if the activity is part of a strategy to address the stressor directly or restructure the situation, it can be considered problem-focused coping. Thus, the categorization of these interventions can depend on how they are used in context, either to alleviate emotional distress or tackle the root causes of stress.

Social and Organizational Support (Instrumental). This involves organizational or structural interventions, which modify workplace or institutional processes. This includes systems that empower supervisors or teachers to adjust workloads or implement coping strategies based on stress data [34, 41, 93], and decision support systems that optimize task allocation, such as patient-to-physician assignments, to mitigate stress [15].

Environmental and Task Adaptation. These interventions adapt the user's environment or tasks in response to detected stress, such as toggling notifications [24], dynamically adjusting lighting [77] and music [101], or modifying task difficulty to maintain engagement [68, 109].

4.3.3 *Intervention Timing.*

The delivery timing of interventions in the reviewed studies can be categorized into four types: just-in-time, on-demand, ambient, and pre-scheduled. Many systems integrated multiple timing strategies to provide comprehensive support. As illustrated in Figure 7, the choice of timing was often associated with the target stress construct. Just-in-time interventions were

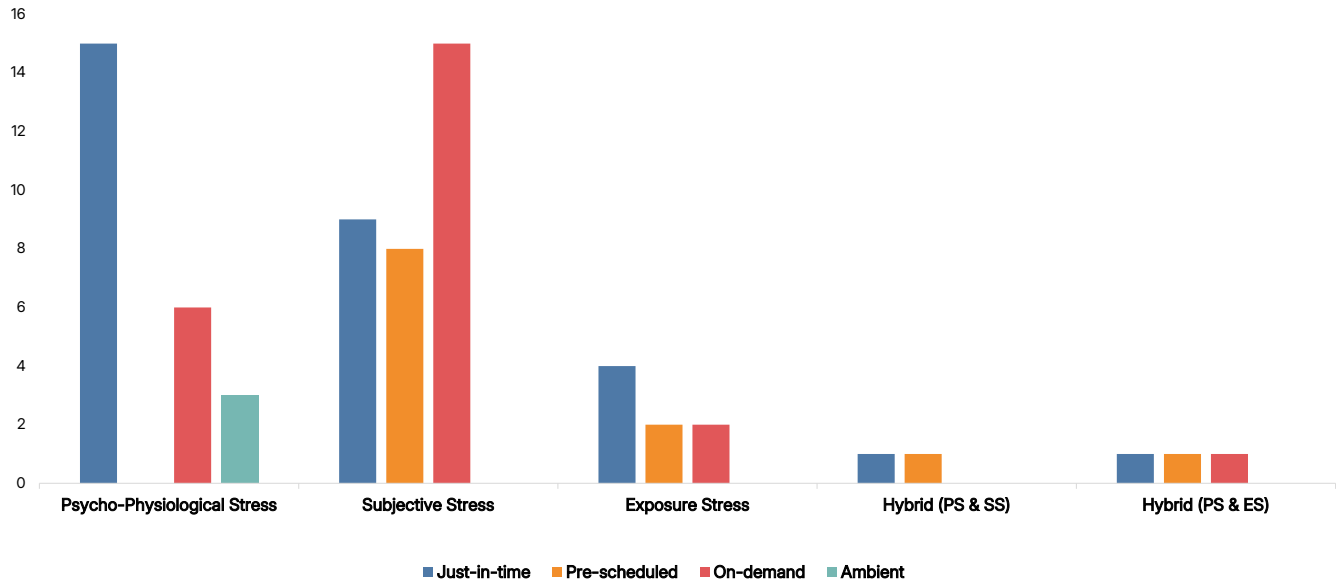

Figure 7: Distribution of Intervention Timing by Stress Construct.

predominant for psycho-physiological stress, while on-demand approaches were most common for managing subjective stress.

Just-in-time interventions, defined as those delivered at the onset of the target behavior to be mitigated [43], were the most prevalent strategy, appearing in 30 studies. These interventions were typically triggered by real-time physiological thresholds [7, 31, 58] or specific behavioral patterns, such as prolonged sedentary behavior [24, 77]. The frequency of these interventions was inherently event-driven rather than fixed. In most cases, the frequency was not explicitly limited, as the intervention was triggered whenever the stress-related criteria were met [45, 58]. However, to avoid being intrusive, some systems imposed a cap, such as a maximum of four interventions per day [33]. In other designs, the frequency was made user-configurable, allowing individuals to adjust the sensitivity by setting their own thresholds for stress levels or sedentary time [24].

On-demand interventions, which can be accessed by the user whenever they choose, were also highly common, found in 24 studies. This modality was primarily used for two distinct purposes: retrospective reflection, where users reviewed past data on dashboards or in VR environments [5, 99]; and proactive access to coping tools, such as initiating a chatbot conversation when needed [33].

Pre-scheduled interventions, which are planned in advance through fixed schedules, alarms, or structured sessions, were reported in 12 studies. The frequency of these interventions varied across studies. For example, some systems delivered regular stress reports at fixed times throughout the day (e.g., 11 AM, 3 PM, 7 PM, 11 PM) [45], provided reflective prompts or educational content on a daily or weekly basis [21, 55, 63], or incorporated structured workshops held periodically over multi-week programs [37]. Other systems enabled users to self-schedule interventions by registering future events and setting reminder times [53].

Finally, *Ambient interventions*, which provide information occasionally while remaining within the user's periphery [44], were the least common strategy, appearing in 3 studies. This approach was most often implemented through always-visible shared displays in workplaces [107, 108] or in-game indicators, such as a continuously displayed breathing rate (BR) and arrow icons showing increases or decreases throughout the game [69].

We also reveal the relationship between intervention content and delivery timing (Table 6). It is important to note that many of the 52 reviewed studies employed multiple intervention content types, and a single content type could be delivered via multiple timing strategies. Therefore, the counts in our analysis represent the frequency of each specific content-timing pairing, not the number of unique studies, and a single paper can contribute to multiple counts. The most prominent pairing observed was On-demand timing with Visualization & Externalization, which occurred 16 times. This suggests a primary use case for on-demand systems as tools for user-driven, retrospective reflection. In contrast, Just-in-time (JIT) timing was strongly associated with interventions that prompt immediate action. It was frequently paired with Biofeedback for Physiological Awareness (10 instances), Mindfulness & Relaxation Techniques (7 instances), followed closely by Cognitive Therapy & Reframing (6 instances) and Activity Suggestions & Proactive Planning (8 instances).

This table also shows that Cognitive Therapy & Reframing appeared with relatively balanced frequency across the three main timing strategies: JIT (6 instances), Pre-scheduled (6 instances), and On-demand (9 instances). In contrast to this distribution, other strategies were observed in more specific pairings. For instance, all 3 instances of Environmental & Task Adaptation were triggered via

Table 6: Distribution of intervention timing by intervention content.

Intervention Content	Intervention Timing			
	Just-in-time	Pre-scheduled	On-demand	Ambient
Visualization and Externalization.	4	4	16	2
Biofeedback for Physiological Awareness	10	-	2	1
Cognitive Therapy and Reframing	6	6	9	-
Mindfulness and Relaxation Techniques	7	3	4	-
Social Support (Emotional)	3	1	2	-
Activity Suggestions and Proactive Planning	8	3	5	-
Social and Organizational Support (Instrumental)	4	2	1	-
Environmental and Task Adaptation	3	-	-	-

JIT, and Ambient timing appeared with Visualization & Externalization (2 instances) as well as once with Biofeedback for Physiological Awareness (1 instance).

4.3.4 Adaptive Intervention Strategies. Beyond static content delivery, a subset of studies employed feedback loops to dynamically adjust intervention strategies, effectively establishing a closed-loop system. This adaptation appeared in two forms: explicit feedback-based learning and implicit control loops.

Explicit Feedback-based Learning. Some systems utilized machine learning (ML) or reinforcement learning (RL) to optimize content based on user input. PopTherapy [65] utilized a model that learned from stress deltas (scores before and after intervention) to recommend the most effective content. Similarly, Ko et al. [48] employed an RL approach where the user's binary evaluation of whether a tip was "helpful" served as a reward signal, enabling the system to learn and prioritize preferred strategies.

Implicit Closed-loop Control. In contrast, adaptation could also occur without active user intervention. Akmandor et al. [6] adopted an implicit control loop that monitored Heart Rate Variability (HRV) in real-time, automatically switching to a different technique if no physiological improvement was detected.

4.4 Evaluation Methods

To investigate RQ3, focusing on how these systems were validated, we analyze the evaluation methods and metrics reported in prior work. Detailed information on all 52 studies is provided in Appendix B Table 9.

4.4.1 Evaluation Context. We categorized the 52 studies based on four methodological aspects: study environment, target user group, experimental design, number of participants, and study duration.

Study Environment. The results revealed that a majority of the studies ($N = 28$) were conducted in field settings. This approach allowed for the evaluation of systems in naturalistic contexts such as workplaces [24, 33], universities [45, 53], and during daily life [63]. Conversely, 18 studies were conducted in controlled laboratory settings, which enabled precise measurement of psycho-physiological responses to standardized tasks or stimuli [2, 31, 58]. A small number of studies ($N = 5$) employed a hybrid approach, combining in-situ data collection with lab-based evaluations [38, 39, 99]. Finally, one study was conducted in a showroom environment [77].

Target User Group. The reviewed literature addressed a diverse range of participant groups, organized into several distinct clusters of research focus. The most frequently studied population was students ($N = 15$) and workers ($N = 12$). Student cohorts appeared in both lab and field contexts [45, 82]. Worker-focused studies primarily involved office and information workers ($N = 8$), with additional research on healthcare professionals[15, 34], military personnel [105], and teachers [49]. A small number of studies ($N = 2$) recruited cohorts of students and professionals to capture stress experiences across both academic and workplace settings [63, 69].

Beyond these groups, some studies drew on the general population ($N = 9$) for broad applicability [65, 110], while others examined more specific contexts such as driving scenarios ($N = 4$) [58, 98]. Finally, a substantial cluster targeted clinical and family/dyadic populations ($N = 10$), including individuals with Autism Spectrum Disorder (ASD) [2, 90], those with intellectual disabilities [30], and pregnant women [100]. Related studies also examined caregiver-patient dyads [41, 48], parents of infants [89], parent-child dyads[102], and couples or close friends [36].

Experimental design. The studies can be classified into three main categories. Non-comparative designs were the most prevalent approach, used in 23 studies. These studies typically focused on the feasibility and user experience of a system within one group [24, 63]. The remaining studies employed comparative designs to assess efficacy. Within-subjects designs were used in 14 studies, where all participants experienced multiple conditions [45, 58]. Between-subjects designs were used in 14 studies, comparing outcomes between distinct experimental and control groups [33, 53]. One study used a comparative analysis based on simulation [15].

Number of Participants. Across the 52 reviewed studies, the number of participants varied substantially, reflecting differences in research scope, feasibility, and target populations. The participant counts ranged from as few as 2 individuals to as many as 214. The average number of participants was 35.9, while *the median was 23*, indicating that half of the studies recruited fewer than two dozen participants. The distribution was highly skewed, as several large-scale studies (e.g., $N = 100, N = 126, N = 214$) [90, 100, 112] increased the overall mean, while the majority of studies involved smaller cohorts ($Q1 = 12, Q3 = 36$). The standard deviation was 39.07, suggesting large variability in study sizes. This variation highlights a trade-off in methodological practice: small-sized studies ($N < 20$)

often enabled fine-grained, controlled evaluation of system feasibility or mechanisms of effect [5, 58], whereas large-sized studies ($N > 70$) were primarily conducted in field or hybrid settings to assess broader applicability and ecological validity [33, 90, 100]. Mid-sized studies (20–40 participants) constituted the most common range, balancing practical recruitment and budget constraints with the need for statistical robustness.

Study duration. The durations of the reviewed studies varied widely, reflecting differences in research objectives. Mid-term studies (1 week to 2 months) were the most common ($N = 24$), particularly for field deployments evaluating user adaptation and real-world effects [24, 33]. Short-term studies (< 24 hours), often single-session, were characteristic of lab-based experiments ($N = 18$) measuring immediate responses [31, 58]. Finally, six studies were long-term (> 2 months), providing insights into sustained engagement and longitudinal outcomes [30, 48, 98]. In addition, four lab-based studies did not report study duration, limiting comparability across designs.

4.4.2 Evaluation Metrics. To assess the effectiveness, usability, and impact of the stress management systems, the reviewed studies employed a combination of quantitative and qualitative metrics. Quantitative metrics were primarily centered around four key areas: changes in user stress, system usability, user engagement with the system, and user perception of system performance.

Changes in Stress. The most common method for assessing stress reduction was the use of validated psychological scales, typically administered pre- and post-study. Commonly applied instruments included the Perceived Stress Scale (PSS) [45, 55, 82], the State-Trait Anxiety Inventory (STAI) [31, 107], and the Depression, Anxiety and Stress Scale (DASS) [33, 105]. Several studies also examined momentary stress changes using self-reported ratings collected immediately before and after interventions [33, 60]. A subset of studies measured stress objectively using physiological signals, such as Heart Rate (HR), Heart Rate Variability (HRV), and Electroencephalography (EEG) data [6, 26, 101].

System Usability and User Experience. To evaluate how easy and satisfying the systems were to use, a majority of studies utilized standardized usability questionnaires. The System Usability Scale (SUS) was the most frequently employed metric [2, 24, 38, 41]. Other standardized scales included the User Experience Questionnaire (UEQ) [45, 82]. Additionally, many studies created their own questionnaires to assess specific aspects like comfort, satisfaction, and perceived usefulness on Likert scales [58, 60].

System Usage and Engagement. User adherence, acceptability, and engagement were primarily measured through the analysis of system logs. The logs captured metrics such as the frequency and duration of use, the number of features accessed, and the number of interventions completed [24, 55, 63]. Questionnaires were also used to measure concepts like willingness to use the system [31], user engagement with the User Engagement Scale (UES) [99], and standardized measures of acceptability (AIM) and feasibility (FIM) [41].

User Perception of System Performance. Several studies quantitatively evaluated how users perceived the system's performance. This included measuring users' agreement with system-detected stress levels [24] and their perception of the algorithm's accuracy and trustworthiness [45]. Qualitative metrics were universally employed to gain a deeper understanding of the user experience and the nuanced impact of the systems. The most common method was semi-structured interviews conducted at the end of the study period [24, 45, 55, 107]. Other widely used methods included diary studies to capture in-situ experiences [38, 39], open-ended survey questions for feedback [33, 65], and think-aloud protocols to observe user interactions in real time [5]. These qualitative approaches were primarily used to explore themes such as the system's perceived effectiveness and usability [24, 41], its impact on user self-reflection and trust [45, 99], and suggestions for future improvement [33, 53].

5 Discussion

Based on our review, we discuss the design implications for future stress management systems, organized into three key dimensions: sensing, intervention, and methodology. Table 7 provides a comprehensive summary of the current challenges and corresponding opportunities discussed in this section.

5.1 Stress Management System Design Implications

Synthesizing our findings, we argue that stress management systems should be viewed not as isolated components but as socio-technical pipelines where sensing, modeling, and intervention choices create cascading effects on user experience. We structure implications around two dimensions: establishing a process-centric sensing framework and designing adaptive, closed-loop interventions.

5.1.1 Toward Process-Centric and Multi-Construct Framework. Our review in Section 4.2.1 reveals that the vast majority of current studies (49 out of 52) focus on inferring a single stress construct, heavily skewed towards appraisal or response. While valid for detection, this fragmented approach oversimplifies the complex stress process defined by Lazarus and Folkman [52], which unfolds from a cause (stressor) through appraisal to a response, reducing it to a single outcome metric. This reductionism leads to two critical design failures: context loss and perceptual mismatch.

Context Loss and the Black Box. The absence of stressor data creates a context vacuum. Systems present users with "what" happened (e.g., a physiological change) without explaining "why" it occurred. This lack of causal information renders the sensing model a "black box," hindering users' data literacy and their ability to interpret or act on their own data [18].

Perceptual Mismatch and Prescriptive Pressure. A focus on a single construct often leads to a discrepancy between the user's subjective feeling and the system's external feedback. For instance, a user might feel excited (positive appraisal), but the system flags high arousal as stress. This cognitive dissonance can lead to distrust [18] or, conversely, over-reliance, where users suppress their own feelings to match the algorithmic judgment [19]. Crucially, this risks shifting the role of emotion-aware technology from being a descriptive tool, which objectively reflects the user's state, to a prescriptive tool that dictates how one should feel. Such systems risk enforcing social norms (e.g., valuing constant calmness) and diminishing user agency by invalidating justified emotional responses [92]. This concern aligns with the design challenges highlighted by Terzimehić et al. [94]. Addressing the tension where

Table 7: Summary of Design Implications and Methodological Recommendations

Category	Current Challenges (Limitations)	Future Directions (Opportunities)
Sensing & Modeling *(Toward Process-Centric)*	• **Context Loss:** Single-construct focus (Response-only) creates a "Black-box" • **Perceptual Mismatch:** Discrepancy between user feeling and sensor data • **High User Burden:** Intrusive collection of multi-modal data	• **Multi-Construct Framework:** Capturing Stressor (Context), Appraisal, and Response • **Advanced Sensemaking:** Causal discovery, counterfactuals, and LLM interpreters • **Human-in-the-loop Calibration:** Users label states to correct ground truth
Intervention *(Toward Adaptive Agents)*	• **Generic Content:** Symptom-oriented (e.g., breathing) w/o addressing root causes • **Static Timing (Open-loop):** Mis-timed JIT causes interruption and fatigue • **High Attrition:** Users disengage due to low relevance	• **Closed-Loop Adaptation:** Learning preferences via MAB/RL algorithms • **Proactive Support Agents:** Linking context to solutions (e.g., study planning) • **Receptivity-Aware JITAI:** User-controlled thresholds and dynamic timing
Methodology *(Toward Rigor & Validity)*	• **Study Design:** Predominance of non-comparative (45%) and short-term (< 8 weeks) studies • **Evaluation Metric:** Focus on efficacy/efficiency, neglecting user experience • **Deployment Barriers:** Data inaccuracy and hardware volatility	• **Rigorous Validation:** Prioritizing longitudinal RCTs and comparative designs • **Experiential Evaluation:** Mixed-methods to understand "how" users experience systems • **Transparent Reporting:** Sharing code/data and documenting real-world challenges

external interfaces compete with internal attention, they argue that systems should not replace human internal processing, but rather act as mirrors that facilitate internal body awareness.

Design Implication: Multi-Construct Sensemaking and Calibration. To address these issues, future systems must transition to a multi-construct data collection strategy that captures the full triad of the stress process: the context (stressor), the user's perception (appraisal), and the bodily reaction (response). This aligns with the personal informatics process model by Li et al. [56], which emphasizes a cyclical process of data collection, reflection, and behavior modification. However, implementing such a framework presents significant challenges, particularly regarding the burden of data collection and the complexity of data analysis for sensemaking. Capturing comprehensive multimodal data can be intrusive and prone to quality issues. To mitigate this, systems should adopt automated data management tools [40] to ensure data quality while minimizing user effort, alongside context-aware automatic journaling [61] to reduce manual reporting burdens. Furthermore, to decipher the complex interplay between these constructs and provide actionable insights beyond simple visualization [14], we need advanced sensemaking tools, such as those supporting causal discovery [39] and counterfactual 'what-if' exploration [38]. Building on these analytic foundations, we suggest leveraging Large Language Model (LLM) agents as intelligent assistants for personal informatics [57, 79]. LLMs can enhance sensemaking by translating opaque sensor outputs into contextualized feedback (e.g., "Your stress index is high this morning because your sleep quality was poor").

However, bridging the sensemaking gap is not enough; systems must also provide mechanisms to correct ground truth errors when discrepancies arise. While our review indicates that only a small subset of studies actively utilized user feedback to refine stress models, embracing *human-in-the-loop calibration*, where users explicitly label or correct their state to retrain the system (e.g., [45]), is essential. This approach corrects ground truth errors and transforms the system from a prescriptive authority into a supportive tool that respects user agency.

5.1.2 Human-in-the-loop and Personalized JITAI for Sustained Engagement. While receptivity is a critical success factor for interventions, it is severely constrained by the sensing limitations discussed above. The resulting lack of context leads to generic content and rigid timing, which frequently discourages user engagement.

Generic Content and Static Just-in-Time Interventions. First, regarding content, most studies neglect personalization, defaulting to generic, symptom-oriented interventions (e.g., breathing exercises). This limits receptivity, as prior work shows that content relevance strongly shapes engagement [72]. Without addressing the root cause (stressor), users may become dependent on temporary symptom relief rather than building sustainable, problem-focused coping capacity. Second, regarding timing, 30 studies employed Just-in-Time (JIT) interventions triggered immediately upon stress detection. However, Howe et al. [33] found that JIT interventions delivered at unwanted moments led to lower satisfaction despite high participation. This misalignment often results in intervention fatigue and attrition over time, consistent with Eysenbach's law of attrition [25], as users disengage when systems fail to align with their individual traits and goals [67].

Design Implication 1: Feedback-Driven Closed-Loop Adaptation. While the majority of prior studies focused on analyzing

pre-post effects, we found that only a small subset (N=4) established a closed-loop system. Crucially, to sustain long-term engagement, such loops must utilize user feedback for a dual purpose: refining the stress inference model (sensing) and optimizing the intervention strategy (acting). Systems should evolve into a holistic loop where feedback calibrates the ground truth of the model as discussed above, while simultaneously adapting intervention content and timing preferences via adaptive algorithms (e.g., reinforcement learning or multi-armed bandits [48, 65, 103, 106]). By integrating these calibrated insights with rich contextual data, technologies such as LLM-based personalized messaging [106] can bridge the gap between sensing and intervening. This integration enables systems to evolve into active agents [79] that support comprehensive sensemaking, linking detected stress responses to personal contexts (e.g., identifying an upcoming exam) and proposing actionable solutions (e.g., a study plan). This shifts support from symptom alleviation (emotion-focused) to root-cause problem solving (problem-focused).

Design Implication 2: Receptivity-Aware JITAIs. Finally, to address timing issues, future work should focus on receptivity modeling. Instead of the static "detect-and-deliver" logic, systems should incorporate a human-in-the-loop approach where users control engagement parameters. For instance, allowing users to adjust sensitivity thresholds (e.g., minimum stress duration), as demonstrated by Elvitigala et al. [24], can minimize intrusion. By respecting user willingness to engage, systems can transform from interruptions into supportive companions. Moreover, recent LLM-based JITAI work suggests that receptivity-aware adaptation need not be limited to threshold tuning [28]: e.g., LLMs can flexibly interpret multimodal contextual cues and user persona to dynamically decide whether and how to intervene, enabling richer and more nuanced receptivity modeling than rule-based systems allow.

5.2 Methodological Challenges and Design Recommendations for User Studies

In this section, we discuss meta-level strategies for designing user studies and conducting field deployments, bridging the gap between ideal methodological rigor and the practical realities of sensing-based research.

5.2.1 Need for Rigorous User Study Designs. Our review highlights a pressing need to elevate the rigor of study designs to ensure both internal and external validity.

Comparative Design and Study Duration. Currently, approximately 45% of the reviewed studies employed non-comparative designs. To validate actual system efficacy beyond simple usability, future research should prioritize comparative designs, such as Randomized Controlled Trials (RCTs) or within-subject counterbalanced strategies. Furthermore, we found that only 11% of studies exceeded eight weeks in duration. Given that health-related habit formation typically requires weeks to months [50], longitudinal studies are essential to observe long-term behavioral trends rather than short-term novelty effects.

Experiential Evaluation Beyond Efficacy. Regarding evaluation metrics, most studies relied on pre-post analyses of psychological scales (e.g., PSS, STAI) or physiological changes (e.g., HRV), supplemented by standard usability scores (e.g., SUS). However, evaluation must go beyond proving efficiency, as Klasnja et al. [47] argue; i.e., HCI research should also focus on the experiential and contextual aspects of system use. We advocate for mixed-method approaches that investigate the underlying mechanisms of success or failure, understanding not just "if" a system works, but "how" users experience it in their daily lives.

5.2.2 Barriers to Large-Scale RCTs. Conducting large-scale RCTs in the wild presents distinct challenges regarding deployment stability and resource constraints. Unlike controlled lab settings, wild deployments expose participants to diverse, uncontrolled environments, inevitably leading to data inaccuracy and missing values that degrade analysis reliability [40]. In addition to these environmental factors, researchers face financial and logistical hurdles. The cost of research-grade equipment (often exceeding $1,000 per unit) creates a barrier to recruitment, while hardware volatility poses risks to longitudinal research; for instance, the discontinuation of devices like the Empatica E4 or the "black-box" nature of frequently changing consumer models complicates validation and reproducibility.

5.2.3 Recommendations for Rigorous & Reproducible Research. To overcome these barriers, we propose the following recommendations for the community:

Shared Data Pipelines & Modular Architecture. While reproducible pipelines like Rapids [96] and BDP [8] exist, greater community effort is needed to address device heterogeneity. We call for the development of modular architectures that are resilient to hardware changes, allowing researchers to swap sensing layers without rebuilding the entire system.

Transparent Reporting & Reproducibility. Even when using consumer devices, reporting must be systematic. Machine learning models for stress prediction should be documented transparently. Following Zhang et al. [111], researchers should share collected datasets and code when possible to enable reproducible research.

Detailed Reporting of Real-world Challenges. Publications should report not only successful outcomes but also the technical realities of deployment. Researchers should detail data quality issues, such as missing data rates [40]. Additionally, since wearable adherence varies significantly by context [12], reporting engagement and adherence metrics is essential to help future researchers anticipate and mitigate attrition.

6 Conclusion

In this integrative review, we analyzed the critical links between detection, intervention, and evaluation in HCI stress management systems. Our findings reveal a field that, while technologically innovative, often disconnects objective physiological sensing from users' subjective appraisals and contexts. This fragmentation leads to generic, decontextualized interventions and evaluations that prioritize technical accuracy over real-world efficacy. To bridge this gap, we advocate for a paradigm shift toward a process-centric framework. This approach moves beyond simple symptom detection to capture the full stress process, integrating the stressor (context), appraisal (perception), and response (physiology/behavior). Crucially, the utility of this framework extends beyond stress management. By modeling the causal mechanism between environmental demands, internal interpretation, and bodily reactions, this

framework offers a generalized template applicable to other mental health domains, such as anxiety regulation and mood disorders, where context and appraisal are equally pivotal. By embracing this integrated, user-centered approach, the HCI community can build the next generation of systems, evolving from prescriptive monitors into supportive, active agents that foster genuine user agency and long-term well-being.

Acknowledgments

The corresponding author of this work is Uichin Lee. This research was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (RS-2025-02305801).

References

[1] Alaa Abd-Alrazaq, Mohannad Alajlani, Reham Ahmad, Rawan AlSaad, Sarah Aziz, Arfan Ahmed, Mohammed Alsahli, Rafat Damseh, and Javaid Sheikh. 2024. The performance of wearable AI in detecting stress among students: systematic review and meta-analysis. *Journal of Medical Internet Research* 26 (2024), e52622.

[2] Deeksha Adiani, Aaron Itzkovitz, Dayi Bian, Harrison Katz, Michael Breen, Spencer Hunt, Amy Swanson, Timothy J Vogus, Joshua Wade, and Nilanjan Sarkar. 2022. Career interview readiness in virtual reality (CIRVR): a platform for simulated interview training for autistic individuals and their employers. *ACM Transactions on Accessible Computing (TACCESS)* 15, 1 (2022), 1–28.

[3] Ana C Aguiar, Mariana Kaiseler, Hugo Meinedo, Traian E Abrudan, and Pedro R Almeida. 2013. Speech stress assessment using physiological and psychological measures. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. 921–930.

[4] Mehmet Berkehan Akçay and Kaya Oğuz. 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication* 116 (2020), 56–76.

[5] Surely Akiri, Vasundhara Misal, Sanaz Taherzadeh, J Lee Jenkins, Gary Williams, Helena Mentis, and Andrea Kleinsmith. 2024. Enhancing stress understanding through team reflection: technology-driven insights in high-stress training scenarios. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work*. 1–18.

[6] Ayten Ozge Akmandor and Niraj K Jha. 2017. Keep the stress away with SoDA: Stress detection and alleviation system. *IEEE Transactions on Multi-Scale Computing Systems* 3, 4 (2017), 269–282.

[7] Anita Beigzadeh, Vahid Yazdnian, and Seyed Kamaledin Setarehdan. 2025. Mental stress detection and performance enhancement using fNIRS and wrist vibrator biofeedback. *Biomedical Signal Processing and Control* 107 (2025), 107877.

[8] Brinnae Bent, Ke Wang, Emilia Grzesiak, Chentian Jiang, Yuankai Qi, Yihang Jiang, Peter Cho, Kyle Zingler, Felix Ikponmwosa Ogbeide, Arthur Zhao, et al. 2021. The digital biomarker discovery pipeline: An open-source software platform for the development of digital biomarkers using mHealth and wearables data. *Journal of clinical and translational science* 5, 1 (2021), e19.

[9] Pablo Calcina-Ccori, Eduardo S Rodriguez-Canales, and Edgar Sarmiento-Calisaya. 2022. A persuasive system for stress detection and management in an educational environment. In *Global IoT Summit*. Springer, 239–249.

[10] Yekta Said Can, Bert Arnrich, and Cem Ersoy. 2019. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of biomedical informatics* 92 (2019), 103139.

[11] Walter Bradford Cannon. 1939. The wisdom of the body. (1939).

[12] Alexandre Chan, Daniella Chan, Hui Lee, Chiu Chin Ng, and Angie Hui Ling Yeo. 2022. Reporting adherence, validity and physical activity measures of wearable activity trackers in medical research: a systematic review. *International Journal of Medical Informatics* 160 (2022), 104696.

[13] Patrick Chipman, Sidney K D'Mello, Barry Gholson, Art Graesser, Bethany McDaniel, and Amy Witherspoon. 2006. Detection of emotions during learning with AutoTutor. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 28.

[14] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. 2017. Understanding self-reflection: how people reflect on personal data through visual data exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 173–182.

[15] Marta Cildoz, Amaia Ibarra, and Fermin Mallor. 2020. Coping with stress in emergency department physicians through improved patient-flow management. *Socio-Economic Planning Sciences* 71 (2020), 100828.

[16] Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. 1983. A global measure of perceived stress. *Journal of health and social behavior* (1983), 385–396.

[17] Glen Debard, Nele De Witte, Romy Sels, Marc Mertens, Tom Van Daele, and Bert Bonroy. 2020. Making wearable technology available for mental healthcare through an online platform with stress detection algorithms: the Carewear project. *Journal of Sensors* 2020, 1 (2020), 8846077.

[18] Xianghua Ding, Shuhan Wei, Xinning Gui, Ning Gu, and Peng Zhang. 2021. Data engagement reconsidered: a study of automatic stress tracking technology in use. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

[19] Vipula Dissanayake, Vanessa Tang, Don Samitha Elvitigala, Elliott Wen, Michelle Wu, and Suranga Nanayakkara. 2022. Troi: Towards understanding users perspectives to mobile automatic emotion recognition system in their natural setting. *Proceedings of the ACM on Human-Computer Interaction* 6, MHCI (2022), 1–22.

[20] Rosie Dobson, Linwei Lily Li, Katie Garner, Taria Tane, Judith McCool, and Robyn Whittaker. 2023. The use of sensors to detect anxiety for in-the-moment intervention: scoping review. *JMIR mental health* 10, 1 (2023), e42611.

[21] Poorvesh Dongre. 2024. Physiology-driven empathic large language models (EmLLMs) for mental health support. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–5.

[22] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.

[23] Don Samitha Elvitigala, Denys JC Matthies, and Suranga Nanayakkara. 2020. StressFoot: Uncovering the potential of the foot for acute stress sensing in sitting posture. *Sensors* 20, 10 (2020), 2882.

[24] Don Samitha Elvitigala, Philipp M Scholl, Hussel Suriyaarachchi, Vipula Dissanayake, and Suranga Nanayakkara. 2021. StressShoe: a DIY toolkit for just-in-time personalised stress interventions for office workers performing sedentary tasks. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*. 1–14.

[25] Gunther Eysenbach. 2005. The law of attrition. *Journal of medical Internet research* 7, 1 (2005), e402.

[26] Stephen H Fairclough and Chelsea Dobbins. 2020. Personal informatics and negative emotions during commuter driving: Effects of data visualization on cardiovascular reactivity & mood. *International Journal of Human-Computer Studies* 144 (2020), 102499.

[27] Shruti Gedam and Sanchita Paul. 2021. A review on mental stress detection using wearable sensors and machine learning techniques. *IEEE Access* 9 (2021), 84045–84066.

[28] David Haag, Devender Kumar, Sebastian Gruber, Dominik P Hofer, Mahdi Sareban, Gunnar Treff, Josef Niebauer, Christopher N Bull, Albrecht Schmidt, and Jan David Smeddinck. 2025. The Last JITAI? Exploring Large Language Models for Issuing Just-in-Time Adaptive Interventions: Fostering Physical Activity in a Prospective Cardiac Rehabilitation Setting. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[29] 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.

[30] Sep Hesselmans, Franka JM Meiland, Esmee Adam, Erwin van de Cruijs, Arthur Vonk, Fransje van Oost, Dwayne Dillen, Stefan de Vries, Eric Riegen, Reon Smits, et al. 2024. Effect of stress-based interventions on the quality of life of people with an intellectual disability and their caregivers.

[31] Victoria Hollis, Alon Pekurovsky, Eunika Wu, and Steve Whittaker. 2018. On being told how we feel: how algorithmic sensor feedback influences emotion perception. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–31.

[32] Karen Hovsepian, Mustafa Al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 493–504.

[33] Esther Howe, Jina Suh, Mehrab Bin Morshed, Daniel McDuff, Kael Rowan, Javier Hernandez, Marah Ihab Abdin, Gonzalo Ramos, Tracy Tran, and Mary P Czerwinski. 2022. Design of digital workplace stress-reduction intervention systems: Effects of intervention type and timing. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–16.

[34] Mahmudul Islam, Sami Rashid, Lishan Rafid, Tasnuba Badrul, Ashraful Islam, and Beenish Moalla Chaudhry. 2024. Design and evaluation of a smartwatch-based physiological signal-driven workplace stress management mhealth tool for bangladeshi healthcare professionals. In *2024 Advances in Science and Engineering Technology International Conferences (ASET)*. IEEE, 1–10.

[35] Shiyi Jiang, Farshad Firouzi, Krishnendu Chakrabarty, and Eric B Elbogen. 2021. A resilient and hierarchical IoT-based solution for stress monitoring in everyday settings. *IEEE Internet of Things Journal* 9, 12 (2021), 10224–10243.

[36] Yanqi Jiang, Xianghua Ding, Xiaojuan Ma, Zhida Sun, and Ning Gu. 2023. IntimaSea: exploring shared stress display in close relationships. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[37] Eunkyung Jo, Austin L Toombs, Colin M Gray, and Hwajung Hong. 2020. Understanding parenting stress through co-designed self-trackers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[38] Gyuwon Jung and Uichin Lee. 2025. CounterStress: Enhancing Stress Coping Planning through Counterfactual Explanations in Personal Informatics. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.

[39] Gyuwon Jung, Sangjun Park, and Uichin Lee. 2024. Deepstress: supporting stressful context sensemaking in personal informatics systems using a quasi-experimental approach. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[40] Yugyeong Jung, Hei Yiu Law, Hadong Lee, Junmo Lee, Bongshin Lee, and Uichin Lee. 2025. DataSentry: Building Missing Data Management System for In-the-Wild Mobile Sensor Data Collection through Multi-Year Iterative Design Approach. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[41] Isha Kaur, Rima Kamel, Evan Sultanik, Jessica Tan, Carla A Mazefsky, Lauren Brookman-Frazee, James C McPartland, Matthew S Goodwin, Jeffrey Pennington, Rinad S Beidas, et al. 2025. Supporting emotion regulation in children on the autism spectrum: co-developing a digital mental health application for school-based settings with community partners. *Journal of pediatric psychology* 50, 1 (2025), 129–140.

[42] Sahib S Khalsa, Ralph Adolphs, Oliver G Cameron, Hugo D Critchley, Paul W Davenport, Justin S Feinstein, Jamie D Feusner, Sarah N Garfinkel, Richard D Lane, Wolf E Mehling, et al. 2018. Interoception and mental health: a roadmap. *Biological psychiatry: cognitive neuroscience and neuroimaging* 3, 6 (2018), 501–513.

[43] Jaejeung Kim, Joonyoung Park, Hyunsoo Lee, Minsam Ko, and Uichin Lee. 2019. LocknType: Lockout task intervention for discouraging smartphone app use. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[44] Tanyoung Kim, Hwajung Hong, and Brian Magerko. 2010. Design requirements for ambient display that supports sustainable lifestyle. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems*. 103–112.

[45] Taewan Kim, Haesoo Kim, Ha Yeon Lee, Hwarang Goh, Shakhboz Abdigapporov, Mingon Jeong, Hyunsung Cho, Kyungsik Han, Youngtae Noh, Sung-Ju Lee, et al. 2022. Prediction for retrospection: Integrating algorithmic stress prediction into personal informatics systems for college students' mental health. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–20.

[46] Zachary D King, Judith Moskowitz, Begum Egilmez, Shibo Zhang, Lida Zhang, Michael Bass, John Rogers, Roozbeh Ghaffari, Laurie Wakschlag, and Nabil Alshurafa. 2019. Micro-stress EMA: A passive sensing framework for detecting in-the-wild stress in pregnant mothers. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 3, 3 (2019), 1–22.

[47] Predrag Klasnja, Sunny Consolvo, and Wanda Pratt. 2011. How to evaluate technologies for health behavior change in HCI research. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 3063–3072.

[48] Eunjung Ko, Karen M Rose, Kristina C Gordon, Sooyoung Kim, Ye Gao, Peng Wang, Lahiru Wijayasingha, Hongning Wang, John A Stankovic, and Kathy D Wright. 2025. Feasibility and Acceptability of the Smarthealth Intervention for Dementia Caregivers. A Qualitative Analysis of a Single-Group Pilot Study. *Journal of Advanced Nursing* (2025).

[49] Rafal Kocielnik and Natalia Sidorova. 2015. Personalized stress management: enabling stress monitoring with lifelogexplorer. *KI-K"unstliche Intelligenz* 29, 2 (2015), 115–122.

[50] Phillippa Lally, Cornelia HM Van Jaarsveld, Henry WW Potts, and Jane Wardle. 2010. How are habits formed: Modelling habit formation in the real world. *European journal of social psychology* 40, 6 (2010), 998–1009.

[51] Anthony D Lamontagne, Tessa Keegel, Amber M Louie, Aleck Ostry, and Paul A Landsbergis. 2007. A systematic review of the job-stress intervention evaluation literature, 1990–2005. *International journal of occupational and environmental health* 13, 3 (2007), 268–280.

[52] Richard S Lazarus and Susan Folkman. 1984. *Stress, appraisal, and coping*. Springer publishing company.

[53] Kwangyoung Lee, Hyewon Cho, Kobiljon Toshnazarov, Nematjon Narziev, So Young Rhim, Kyungsik Han, YoungTae Noh, and Hwajung Hong. 2020. Toward future-centric personal informatics: Expecting stressful events and preparing personalized interventions in stress management. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[54] Aniek Lentferink, Matthijs L Noordzij, Anouk Burgler, Randy Klaassen, Youri Derks, Hilbrand Oldenhuis, Hugo Velthuijsen, and Lisette van Gemert-Pijnen. 2022. On the receptivity of employees to just-in-time self-tracking and eCoaching for stress management: a mixed-methods approach. *Behaviour & information technology* 41, 7 (2022), 1398–1424.

[55] Aniek Lentferink, Hilbrand Oldenhuis, Hugo Velthuijsen, and Lisette van Gemert-Pijnen. 2023. How Reflective Automated e-Coaching Can Help Employees Improve Their Capacity for Resilience: Mixed Methods Study. *JMIR human factors* 10 (2023), e34331.

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

[57] Jiachen Li, Xiwen Li, Justin Steinberg, Akshat Choube, Bingsheng Yao, Xuhai Xu, Dakuo Wang, Elizabeth Mynatt, and Varun Mishra. 2025. Vital Insight: Assisting Experts' Context-Driven Sensemaking of Multi-modal Personal Tracking Data Using Visualization and Human-in-the-Loop LLM. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 3 (2025), 1–37.

[58] Diana MacLean, Asta Roseway, and Mary Czerwinski. 2013. MoodWings: a wearable biofeedback device for real-time stress intervention. In *Proceedings of the 6th international conference on PErvasive Technologies Related to Assistive Environments*. 1–8.

[59] Juan Manuel Madrid, Carlos A Arce-Lopera, and Fabian Lasso. 2018. Biometric interface for driver's stress detection and awareness. In *Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 132–136.

[60] Nihara B Mayurawasala, Udaka A Manawadu, Dilmi P Kulugammana, and P Ravindra S De Silva. 2024. A Companion Robot for Reducing Stress and Increasing Workability. In *2024 International Conference on Image Processing and Robotics (ICIPRoB)*. IEEE, 1–6.

[61] 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.

[62] Sameer Neupane, Poorvesh Dongre, Denis Gracanin, and Santosh Kumar. 2025. Wearable Meets LLM for Stress Management: A Duoethnographic Study Integrating Wearable-Triggered Stressors and LLM Chatbots for Personalized Interventions. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[63] Sameer Neupane, Mithun Saha, Nasir Ali, Timothy Hnat, Shahin Alan Samiei, Anandatirtha Nandugudi, David M Almeida, and Santosh Kumar. 2024. Momentary stressor logging and reflective visualizations: Implications for stress management with wearables. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.

[64] Muhammad Nouman, Sui Yang Khoo, MA Parvez Mahmud, and Abbas Z Kouzani. 2021. Recent advances in contactless sensing technologies for mental health monitoring. *IEEE Internet of Things Journal* 9, 1 (2021), 274–297.

[65] Pablo Paredes, Ran Gilad-Bachrach, Mary Czerwinski, Asta Roseway, Kael Rowan, and Javier Hernandez. 2014. PopTherapy: Coping with stress through pop-culture. In *Proceedings of the 8th international conference on pervasive computing technologies for healthcare*. 109–117.

[66] Eunji Park, Duri Lee, Yunjo Han, James Diefendorff, and Uichin Lee. 2024. Hide-and-seek: Detecting Workers' Emotional Workload in Emotional Labor Contexts Using Multimodal Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–28.

[67] Joonyoung Park and Uichin Lee. 2023. Understanding disengagement in just-in-time mobile health interventions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 2 (2023), 1–27.

[68] Avinash Parnandi and Ricardo Gutierrez-Osuna. 2015. Physiological modalities for relaxation skill transfer in biofeedback games. *IEEE journal of biomedical and health informatics* 21, 2 (2015), 361–371.

[69] Avinash Parnandi and Ricardo Gutierrez-Osuna. 2017. Visual biofeedback and game adaptation in relaxation skill transfer. *IEEE Transactions on Affective Computing* 10, 2 (2017), 276–289.

[70] JAPH Perera, LM Rathnarajah, and HB Ekanayake. 2016. Biofeedback based computational approach for working stress reduction through meditation technique. In *2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 132–140.

[71] A Philip Schmidt, R Duerichen Reiss, and Introducing WESAD Kristof Van Laerhoven. 2018. a multimodal dataset for wearable Stress and Affect Detection. In *Proceedings of the International Conference on Multimodal Interaction*.

[72] Martin Pielot, Bruno Cardoso, Kleomenis Katevas, Joan Serrà, Aleksandar Matic, and Nuria Oliver. 2017. Beyond interruptibility: Predicting opportune moments to engage mobile phone users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–25.

[73] Ming-Zher Poh, Tobias Loddenkemper, Nicholas C Swenson, Shubhi Goyal, Joseph R Madsen, and Rosalind W Picard. 2010. Continuous monitoring of electrodermal activity during epileptic seizures using a wearable sensor. In *2010 annual international conference of the IEEE engineering in medicine and biology*. IEEE, 4415–4418.

[74] Mores Prachyabrued, Disathon Wattanadhirach, Richard B Dudrow, Nat Krairojananan, and Pusit Fuengfoo. 2019. Toward virtual stress inoculation training of prehospital healthcare personnel: A stress-inducing environment design and investigation of an emotional connection factor. In *2019 IEEE conference on virtual reality and 3d user interfaces (vr)*. IEEE, 671–679.

[75] Nafiul Rashid, Trier Mortlock, and Mohammad Abdullah Al Faruque. 2023. Stress Detection using Context-Aware Sensor Fusion from Wearable Devices. *IEEE Internet of Things Journal* (2023).

[76] Cheryl Regehr, Dylan Glancy, and Annabel Pitts. 2013. Interventions to reduce stress in university students: A review and meta-analysis. *Journal of affective disorders* 148, 1 (2013), 1–11.

[77] Xipei Ren, Bin Yu, Yuan Lu, Biyong Zhang, Jun Hu, and Aarnout Brombacher. 2019. LightSit: An unobtrusive health-promoting system for relaxation and fitness microbreaks at work. *Sensors* 19, 9 (2019), 2162.

[78] Xipei Ren, Xiaoyu Zhang, Renyao Zou, Ran Yan, and Bin Yu. 2025. EmoVis: exploring data-enabled analogue journaling to promote self-reflection for mental wellness among college students. *Behaviour & Information Technology* 44, 4 (2025), 859–881.

[79] Zhiwei Ren, Junbo Li, Minjia Zhang, Di Wang, Xiaoran Fan, and Longfei Shangguan. 2025. Toward Sensor-In-the-Loop LLM Agent: Benchmarks and Implications. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*. 254–267.

[80] Klaus R Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech communication* 40, 1-2 (2003), 227–256.

[81] 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*. 400–408.

[82] Tanja Schneeberger, Naomi Sauerwein, Manuel S Anglet, and Patrick Gebhard. 2021. Stress management training using biofeedback guided by social agents. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 564–574.

[83] André Schulz and Claus Vögele. 2015. Interoception and stress. *Frontiers in psychology* 6 (2015), 993.

[84] Hans Selye. 1978. *The stress of life, Rev.* McGraw Hill.

[85] Cornelia Setz, Bert Arnrich, Johannes Schumm, Roberto La Marca, Gerhard Tröster, and Ulrike Ehlert. 2009. Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on information technology in biomedicine* 14, 2 (2009), 410–417.

[86] Manoj Sharma and Sarah E Rush. 2014. Mindfulness-based stress reduction as a stress management intervention for healthy individuals: a systematic review. *Journal of evidence-based complementary & alternative medicine* 19, 4 (2014), 271–286.

[87] Myriam Sillevis Smitt, Mehdi Montakhabi, Jessica Morton, Cora van Leeuwen, Klaas Bombeke, and An Jacobs. 2022. Users' perceptions of a digital stress self-monitoring application: Research insights to design a practical innovation. In *International Conference on Human-Computer Interaction*. Springer, 325–341.

[88] Joshua M Smyth and Kristin E Heron. 2016. Is providing mobile interventions" just-in-time" helpful? An experimental proof of concept study of just-in-time intervention for stress management. In *2016 IEEE Wireless Health (WH)*. IEEE, 1–7.

[89] Seokwoo Song, Naomi Yamashita, and John Kim. 2020. Bodeum: Encouraging working parents to provide emotional support for stay-at-home parents in Korea. In *Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 38–49.

[90] Kirsten L Spaargaren, Yvette Roke, Sander M Begeer, Annemieke van Straten, Heleen Riper, Kirstin Greaves-Lord, and Anke M Scheeren. 2025. A randomized controlled trial into the effectiveness of a mobile health application (SAM) to reduce stress and improve well-being in autistic adults. *Autism* (2025), 13623613251346885.

[91] Phoebe A Staab, A Jess Williams, MacKenzie DA Robertson, and Petr Slovak. 2024. "Can you be with that feeling?": Extending Design Strategies for Interoceptive Awareness for the Context of Mental Health. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21.

[92] 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*. 782–793.

[93] Chiew Seng Sean Tan, Johannes Sch¨oning, Kris Luyten, and Karin Coninx. 2014. Investigating the effects of using biofeedback as visual stress indicator during video-mediated collaboration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 71–80.

[94] Nađa Terzimehić, Renate Häuslschmid, Heinrich Hussmann, and MC Schraefel. 2019. A review & analysis of mindfulness research in HCI: Framing current lines of research and future opportunities. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[95] Himanshu Thapliyal, Vladislav Khalus, and Carson Labrado. 2017. Stress detection and management: A survey of wearable smart health devices. *IEEE Consumer Electronics Magazine* 6, 4 (2017), 64–69.

[96] Julio Vega, Meng Li, Kwesi Aguillera, Nikunj Goel, Echhit Joshi, Kirtiraj Khandekar, Krina C Durica, Abhineeth R Kunta, and Carissa A Low. 2021. Reproducible analysis pipeline for data streams: open-source software to process data collected with mobile devices. *Frontiers in digital health* 3 (2021), 769823.

[97] Luis M Vela, Henry Crandall, Taehwan Lim, Fu Zhang, Austin Gibbs, Andrew RJ Mitchell, Adrian Condon, Lisa M Diamond, Huanan Zhang, and Benjamin Sanchez. 2023. IoMT-enabled stress monitoring in a virtual reality environment and at home. *IEEE Internet of Things Journal* (2023).

[98] Rohit Verma, Bivas Mitra, and Sandip Chakraborty. 2024. On-the-Go Automated Break Recommendation for Stress Avoidance during Highway Driving. *Journal on Autonomous Transportation Systems* 2, 1 (2024), 1–25.

[99] Nadine Wagener, Marit Bentvelzen, Bastian D¨anekas, Paweł W Woźniak, and Jasmin Niess. 2023. VeatherReflect: Employing weather as qualitative representation of stress data in virtual reality. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 446–458.

[100] Lauren S Wakschlag, Darius Tandon, Sheila Krogh-Jespersen, Amelie Petitclerc, Ashley Nielsen, Rhoozbeh Ghaffari, Leena Mithal, Michael Bass, Erin Ward, Jonathan Berken, et al. 2021. Moving the dial on prenatal stress mechanisms of neurodevelopmental vulnerability to mental health problems: A personalized prevention proof of concept. *Developmental psychobiology* 63, 4 (2021), 622–640.

[101] Chao Wang, Yi Wu, Nabil Sabor, Qing Zhang, Min Wang, Yongfu Li, and Guoxing Wang. 2023. Configurable Multimodal Therapeutic System for Promoting Emotional Relaxation. In *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 1–4.

[102] Pinhao Wang, Lening Huang, Guang Dai, Jing Li, Jun Hu, Emilia Barakova, Cheng Yao, and Fangtian Ying. 2024. Evaluating the Role of Interactive Encouragement Prompts for Parents in Parent–Child Stress Management. *Applied Sciences* 15, 1 (2024), 256.

[103] Xingbo Wang, Janessa Griffith, Daniel A Adler, Joey Castillo, Tanzeem Choudhury, and Fei Wang. 2025. Exploring Personalized Health Support through Data-Driven, Theory-Guided LLMs: A Case Study in Sleep Health. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–15.

[104] Sebastian Weiß and Wilko Heuten. 2023. Don't Panic!-Influence of Virtual Stressor Representations from the ICU Context on Perceived Stress Levels. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[105] Brent D Winslow, Rebecca Kwasinski, Jeffrey Hullfish, Mitchell Ruble, Adam Lynch, Timothy Rogers, Debra Nofziger, William Brim, and Craig Woodworth. 2022. Automated stress detection using mobile application and wearable sensors improves symptoms of mental health disorders in military personnel. *Frontiers in Digital Health* 4 (2022), 919626.

[106] Ruolan Wu, Chun Yu, Xiaole Pan, Yujia Liu, Ningning Zhang, Yue Fu, Yuhan Wang, Zhi Zheng, Li Chen, Qiaolei Jiang, et al. 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*. 1–24.

[107] Mengru Xue, Rong-Hao Liang, Jun Hu, Bin Yu, and Loe Feijs. 2022. Understanding how group workers reflect on organizational stress with a shared, anonymous heart rate variability data visualization. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[108] Mengru Xue, Rong-Hao Liang, Bin Yu, Mathias Funk, Jun Hu, and Loe Feijs. 2019. AffectiveWall: designing collective stress-related physiological data visualization for reflection. *IEEE Access* 7 (2019), 131289–131303.

[109] Chang Yun, Dvijesh Shastri, Ioannis Pavlidis, and Zhigang Deng. 2009. O'game, can you feel my frustration? Improving user's gaming experience via stresscam. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2195–2204.

[110] Jin Zhang, Hao Tang, Dawei Chen, and Qian Zhang. 2012. deStress: Mobile and remote stress monitoring, alleviation, and management platform. In *2012 IEEE global communications conference (GLOBECOM)*. IEEE, 2036–2041.

[111] Panyu Zhang, Gyuwon Jung, Jumabek Alikhanov, Uzair Ahmed, and Uichin Lee. 2024. A reproducible stress prediction pipeline with mobile sensor data. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 8, 3 (2024), 1–35.

[112] Liang Zhao, Weiwei Zheng, Zihan Zeng, Bokai Chen, Qi Li, Xin Wang, Lei Cao, and Feng Yu. 2024. GratitudeGuider: A Smart Positive Psychological Intervention System for Mental Wellbeing. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 191–195.

[113] Nan Zhao, Elena C Kodama, and Joseph A Paradiso. 2022. Mediated atmosphere table (MAT): Adaptive multimodal media system for stress restoration. *IEEE Internet of Things Journal* 9, 23 (2022), 23614–23625.

Appendix

A Terminology

Table 8: Glossary of physiological and motion-related terms.

Term (Abbreviation)	Description
Electrocardiogram (ECG)	A test that measures the electrical activity of the heart.
Electrodermal Activity (EDA)	A measure of the skin's electrical conductance, related to sweat gland activity.
Photoplethysmography (PPG)	A non-invasive method for monitoring heart rate by detecting blood volume changes.
Blood Volume Pulse (BVP)	The pulse wave detected by PPG, indicating changes in blood volume with each heartbeat.
Respiration Rate (RESP)	The number of breaths a person takes per minute.
Skin Temperature (SkinTemp)	The temperature of the skin surface, which can indicate changes in the body's thermal regulation.
Blood Pressure (BP)	The force of blood against the walls of the arteries, commonly measured in systolic/diastolic values.
Electroencephalography (EEG)	A test that measures electrical activity in the brain.
Functional Near-Infrared Spectroscopy (fNIRS)	A technique for measuring brain activity by monitoring blood oxygenation.
Oxyhemoglobin (HBO)	Hemoglobin bound to oxygen, used to transport oxygen in the blood.
Deoxyhemoglobin (HHB)	Hemoglobin without oxygen, reflecting oxygen usage in tissues.
Ballistocardiography (BCG)	The measurement of the body's mechanical movements caused by the heartbeat.
Accelerometer (ACC)	A sensor that measures the rate of change in velocity (acceleration) of an object.
Gyroscope (Gyro)	A sensor that measures the orientation or angular velocity of an object.
Global Positioning System (GPS)	A satellite-based navigation system that provides location and time information.
Inertial Measurement Unit (IMU)	A device that combines accelerometers, gyroscopes, and sometimes magnetometers to track motion.
Heart Rate (HR) / Heart Rate Variability (HRV)	The number of heartbeats per minute and the variation in the intervals between heartbeats.
Skin Conductance Level (SCL)	The baseline level of skin conductance, typically slow-changing, indicating arousal or stress.
Skin Conductance Response (SCR)	The rapid fluctuations in skin conductance in response to stimuli, often linked to emotional reactions.

B Summary of Study Design, Participants, and Evaluation Context

Table 9: Comprehensive summary of included studies, study designs, participant groups, sample sizes (S1 and S2 in paper [59] denote participant numbers in Study 1 and Study 2, respectively), environments, and durations. Study design is classified as *Non-comparative* for studies that deploy or evaluate a single system or intervention without an explicit comparison condition or control group, *Between-subjects* when participants are assigned to one of multiple conditions and experience only that condition, and *Within-subjects* when each participant experiences multiple conditions or phases and thus serves as their own control.

Paper	Study Design	Target User	N	Environment	Duration
[24]	Non-comparative	Children with ASD & Caregivers	10	Field	4 weeks
[45]	Non-comparative	Couples or close friends	36	Field	25 days
[107]	Non-comparative	Dementia patients and caregivers	24	Field	1 week
[58]	Non-comparative	Drivers & Commuters	11	Lab	1 hour
[110]	Non-comparative	General population	30	Lab	20 min
[33]	Non-comparative	General population	86	Field	4 weeks
[89]	Non-comparative	Healthcare professionals	18	Field	2 weeks
[31]	Non-comparative	Immigrant mothers	64	Lab	15 min
[99]	Non-comparative	Individuals with intellectual disability	20	Hybrid (Field+Lab)	2 hours data + 15 min use
[63]	Non-comparative	Infant parents	122	Field	14 weeks
[5]	Non-comparative	Office/Information workers	8	Lab	25 min
[55]	Non-comparative	Office/Information workers	28	Field	6 weeks
[38]	Non-comparative	Office/Information workers	12	Hybrid (Field+Lab)	6 weeks data + 1 week use
[62]	Non-comparative	Office/Information workers	2	Field	22 days
[37]	Non-comparative	Office/Information workers (remote)	5	Field	2 weeks
[21]	Non-comparative	Office/Information workers (SW company)	8	Field	1 day
[36]	Non-comparative	Students	19	Field	4 weeks
[53]	Non-comparative	Students	47	Field	4 weeks
[39]	Non-comparative	Students	24	Hybrid (Field+Lab)	6 weeks data + 1 week use
[112]	Non-comparative	Students	126	Field	2 weeks
[82]	Non-comparative	Students	71	Lab	140 min
[2]	Non-comparative	Students (paramedic)	17	Lab	1 hour
[65]	Non-comparative	Students & Professionals	95	Field	4 weeks
[59]	Non-comparative	Teachers (vocational school)	7 (S1), 6 (S2)	Lab	Not specified
[41]	Non-comparative	Physicians (emergency department)	73	Field	1 week
[98]	Between-subjects	Adults with ASD	7	Field	10 months
[93]	Between-subjects	Adults with ASD	20	Lab	90 min
[109]	Between-subjects	General Population	14	Lab	2 hours
[6]	Between-subjects	General population	33	Lab	90 min
[70]	Between-subjects	Military personnel	21	Lab	105 min
[34]	Between-subjects	Office/Information workers (remote)	24	Field	2 weeks
[60]	Between-subjects	Pregnant women	12	Lab	Not specified
[77]	Between-subjects	Students	50	Living-lab/Showroom	10 min
[108]	Between-subjects	Students	24	Lab	1 hour
[101]	Between-subjects	Students	22	Lab	Not specified
[68]	Between-subjects	Students	25	Lab	18 min
[69]	Between-subjects	Students	24	Lab	50 min
[54]	Between-subjects	Students	17	Field	2 weeks
[102]	Between-subjects	Students & Professionals	72	Lab	1.5 hours

Paper	Study Design	Target User	N	Environment	Duration
[7]	Within-subjects	Drivers & Commuters	20	Lab	Not specified
[49]	Within-subjects	Drivers & Commuters	21	Field	4 weeks
[26]	Within-subjects	Drivers & Commuters	8	Hybrid (Field+Lab)	4 days + Not specified
[9]	Within-subjects	General population	7	Field	55 min
[15]	Within-subjects	General population	30	Hybrid (Field+Lab)	50 days + Not specified
[78]	Within-subjects	General population	32	Field	7 days
[87]	Within-subjects	General population	12	Field	2 weeks
[105]	Within-subjects	General population	30	Field	12 weeks
[17]	Within-subjects	Instructors and workers	5	Field	1 week
[100]	Within-subjects	Parent-Child dyads	100	Field	14 weeks
[90]	Within-subjects	Students	214	Field	1 month
[30]	Within-subjects	Students	41	Field	6 months
[48]	Within-subjects	Students	20	Field	4 months