

Critiquing Self-report Practices for Human Mental and Wellbeing Computing at Ubicomp

NAN GAO*, Department of Computer Science and Technology, Tsinghua University, China and University of New South Wales (UNSW), Australia

SOUNDARIYA ANANTHAN*, University of New South Wales (UNSW), Australia

CHUN YU, Department of Computer Science and Technology, Tsinghua University, China

YUNTAO WANG, Department of Computer Science and Technology, Tsinghua University, China

FLORA D. SALIM, University of New South Wales (UNSW), Australia

Computing human mental and wellbeing is crucial to various domains, including health, education, and entertainment. However, the reliance on self-reporting in traditional research to establish ground truth often leads to methodological inconsistencies and susceptibility to response biases, thus hindering the effectiveness of modelling. This paper presents the first systematic methodological review of self-reporting practices in Ubicomp within the context of human mental and wellbeing computing. Drawing from existing survey research, we establish guidelines for self-reporting in human wellbeing studies and identify shortcomings in current practices at Ubicomp community. Furthermore, we explore the reliability of self-report as a means of ground truth and propose directions for improving ground truth measurement in this field. Ultimately, we emphasize the urgent need for methodological advancements to enhance human mental and wellbeing computing.

CCS Concepts: • **Human-centred computing** → **Ubiquitous and mobile computing**; • **Applied computing**;

Additional Key Words and Phrases: Human-centred computing; affective computing; self-report; ground truth; experience sampling method; survey; mental wellbeing; critical review

ACM Reference Format:

Nan Gao, Soundariya Ananthan, Chun Yu, Yuntao Wang, and Flora D. Salim. 2023. Critiquing Self-report Practices for Human Mental and Wellbeing Computing at Ubicomp. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 37, 4, Article 52 (January 2023), 23 pages. <https://doi.org/10.1145/1122445.1122456>

1 INTRODUCTION

Recently, Human-Centred Computing (HCC) [71] has gained immense importance due to its potential to enhance the interaction between humans and computers. It focuses on designing effective computer systems that take into account personal, social, and cultural factors, and addresses issues such as the relationships between

*Both authors contributed equally to this research.

Authors' addresses: **Nan Gao**, Department of Computer Science and Technology, and Tsinghua University, Beijing, China and University of New South Wales (UNSW), Sydney, Australia, 1466, nangao@tsinghua.edu.cn; **Soundariya Ananthan**, University of New South Wales (UNSW), Sydney, Australia, 1466, a.soundariya@gmail.com; **Chun Yu**, Department of Computer Science and Technology, and Tsinghua University, Beijing, China, chunyu@mail.tsinghua.edu.cn; **Yuntao Wang**, Department of Computer Science and Technology, and Tsinghua University, Beijing, China, yuntaowang@tsinghua.edu.cn; **Flora D. Salim**, flora.salim@unsw.edu.au, University of New South Wales (UNSW), Sydney, Australia, 1466, flora.salim@unsw.edu.au.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Association for Computing Machinery.

2474-9567/2023/1-ART52 \$15.00

<https://doi.org/10.1145/1122445.1122456>

computing technology and art, social, and cultural issues [71]. HCC has benefitted multiple fields such as Human-Computer Interaction (HCI) [124], Computer-Supported Cooperative Work (CSCW) [114], User-Centred Design [2], Cognitive Psychology [126], Sociology [95], and Anthropology [103], etc. The utilization of HCC technologies presents significant potential for enhancing human wellbeing through the development of early detection and intervention techniques for mental health. Such techniques, including emotion recognition [39], engagement detection [55], and interventions (e.g., cognitive training programs [70]), can assist individuals in achieving and maintaining optimal emotional and mental states.

Despite several decades of research in HCC fields, those approaches have not yet been successful in transitioning from research trials to practical implementation in real-world scenarios, particularly when it comes to measuring human mental states and wellbeing such as emotion [39], depression [146], engagement [55], anxiety [67], and more. In contrast, human physical activity recognition has demonstrated remarkable levels of accuracy, ranging from 83 to 100% [8], enabling its successful application in various real-world contexts, such as fitness trackers [94] and fall detection systems [96]. Nevertheless, the assessment of mental wellbeing remains uniquely challenging due to its subjective nature, which encompasses emotions, thoughts, and subjective experiences that are inherently difficult to objectively quantify and measure. This difficulty in achieving objective measurements has resulted in relatively low accuracy, rendering it inadequate for real-world applications and effective interventions in practical settings.

The primary reason for the low performance in computing mental wellbeing arises from the common practice of using self-report as the ground truth. This approach contrasts with the assessment of physical activity, which benefits from more objective measures such as direct observation and expert annotation. Mental wellbeing modelling, however, depends heavily on individual self-reports of experiences and emotions, introducing potential inaccuracies. For instance, Gao et al. [55] employed physiological and environmental sensing to predict student engagement, using self-report data from the *In-Class Student Engagement Questionnaires* (ISEQ) [50] as ground truth. Similarly, Wang et al. [146] tracked depression dynamics in college students using mobile phone and wearable sensing, with self-reported depression scores from the PHQ-8 [84] and PHQ-4 [82] questionnaires as the ground truth. While this reliance on self-reporting simplifies data collection, it introduces methodological inconsistencies (e.g., sample unrepresentativeness, threats to the reliability and validity of survey instruments) and susceptibility to various response biases (e.g., social desirability, extreme responding, and recall bias), which can significantly affect the effectiveness of mental wellbeing models.

The use of self-reporting in human mental computing research presents two primary concerns: the standardization of self-reporting practices themselves, and the choice of this research method. Challenges such as sample representativeness can distort data, which in turn affects the effectiveness of modelling derived from this data. Factors like compensation schemes, non-response rates and withdrawal mechanisms also significantly impact the quality of self-report data [75]. Additionally, HCC studies, which differ from traditional survey research, require special considerations such as the integration of sensing data collection, the frequency of experience sampling methods (ESM), and the relation of these signals with psychological. Given these complexities, self-reporting, while seemingly straightforward, demands careful planning and meticulous execution in HCC research as the credibility of the self-report data is vital to prevent the risk of ‘garbage in, garbage out’ [77].

While a few studies have explored the limitations of relying on self-report measures as ground truth for HCC [28, 54], little progress has been made in developing effective solutions to address these issues. Gao et al. [54] investigated the reliability of self-report and found that physiologically measured learning engagement and perceived engagement are not always consistent, underscoring the potential pitfalls of relying solely on subjective annotations as the basis for establishing ground truth. However, their study did not propose specific solutions to this problem. Das et al. [28] found that the prediction performance of mental wellbeing depended on the method used to establish ground truth, with psychological-related features being more effective for self-report stress and behavioural-related signals being more effective for objective arousal signals (high arousal

duration). While this approach represents an initial attempt to use alternative measures of ground truth, there are still concerns about whether arousal signals inferred from heart rate can be considered a reliable measure of mental wellbeing.

In this work, we aim to raise awareness within the Ubicomp community, echoing standardising self-report practices and exploring reliable methods for establishing ground truth in HCC studies. To this end, we analyse 49 human mental computing studies in the *ACM International Joint Conference on Pervasive and Ubiquitous Computing* (UbiComp) and the *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (IMWUT), and conducted a systematic literature review. We demonstrate comprehensive guidelines of self-reporting practices in human mental and wellbeing computing studies, and emphasize the need for methodological evolution in this field, advocating for a shift away from traditional self-reporting towards a more reliable and diverse method. For clarity, our study focuses on HCC studies related to human mental states and wellbeing computation, excluding physical behaviours computation, which is already well-established in the field. Specifically, our contributions are as follows:

- Recognizing that self-reporting is the predominant method for measuring ground truth in HCC studies, we formulate a set of guidelines for self-reporting practices. It aims to enhance community standards and enhance the credibility of future research in the field of human wellbeing computing.
- Based on our analysis of 49 Ubicomp papers and the evaluation of self-reporting practices using the proposed guidelines, we identified substantial deficiencies and discrepancies in the self-report methodologies employed in current HCC studies.
- We discussed the reliability of self-report data as ground truth and demonstrated methodologies to enhance ground truth measurement in HCC studies. We also point out the future directions to improve human mental and wellbeing computing.

The structure of the remaining paper is as follows. Section 2 offers an overview of the background related to ground truth measures in HCC studies. Section 3 presents an examination of current self-report practices in HCC studies, highlighting common pitfalls within this domain. The reliability of self-report as a ground truth measure in human-centred computing is demonstrated in Section 4. Section 5 delves into the discussion of future directions aimed at establishing ground truth in HCC studies. Section 6 indicates future directions to advancing HCC studies. Finally, Section 7 provides the concluding remarks for the paper.

2 BACKGROUND

In this section, we introduce the commonly used self-report measures in HCC, as well as the psychological constructs that are of primary interest to researchers and data commonly used in HCC research. This section aims to provide readers with a foundational understanding of these topics and their relevance to HCC research.

2.1 Survey and Experience Sampling Method as Self-Report Measures

Self-reporting involves individuals reporting their symptoms, behaviours, beliefs, or attitudes through tests, measures, or surveys [111]. These self-reports are typically done on paper, electronically, or through interviews. Self-reporting is widely used in human-related studies, including psychological research and HCC, while in the latter, the survey and experience sampling methods are particularly popular.

Surveys are often preferred due to their cost-effectiveness and ability to gather data from a large group of participants, covering a wide range of information such as basic demographics and specific social or behavioural factors [107]. Surveys can be conducted through questionnaires or interviews, but questionnaires are favoured for their ease of administration on a large scale. Typically, they can be categorized as either cross-sectional or longitudinal, with the latter involving data collection at multiple time points [48]). However, it should be noted that longitudinal surveys may not always be optimal due to their reliance on participants' cognitive abilities. To

overcome this limitation, the *Experience Sampling Method* (ESM) and *Ecological Momentary Assessment* (EMA) [14] were proposed. These methods provide repeated snapshots of individuals' subjective information, thereby reducing reliance on memory and increasing response validity by gathering data at multiple points throughout the day or a specific period [127]. Additionally, these methods allow for the storage of contextual details such as time and location, enabling the examination of temporal changes in participants' experiences and behaviours [139]. It is worth noting that ESM and EMA are often used interchangeably in [140, 141]. In this paper, we refer to the term ESM.

2.2 Psychological Constructs in HCC Literature

Psychological constructs play a critical role in describing behaviour patterns and understanding natural phenomena. Researchers heavily rely on these constructs to explore human behaviour, emotions, and thoughts [112]. To avoid confusion, it is crucial to establish clear definitions and differentiate between similar constructs. For example, terms like affect, emotion, and mood may seem similar but have distinct meanings. It is noteworthy that certain constructs can be measured objectively. Sleep, for instance, can be measured using sleep sensors, Fitbit devices [152], or even mobile phones [142, 145, 149]. However, our research will not focus on easily measurable constructs, as they have already been extensively studied. Instead, we will primarily examine constructs that are typically measured subjectively. Particularly, we have identified several commonly used psychological constructs in HCC studies.

Anxiety, Stress and Panic. *Anxiety* and *stress* are often used interchangeably, but there is a distinction between the two. Stress is a reaction to specific events or situations that can trigger emotional responses. On the other hand, anxiety is characterized by persistent worry, tension, and uncertainty [115]. It is an ongoing state of unease that can be difficult to control which may lead to various mental and health problems if not addressed. Similarly, *anxiety* and *panic* are similar but differ in their onset and symptoms. Panic attacks occur suddenly and are intense episodes of fear, often accompanied by physical symptoms such as a racing heart, shortness of breath, and chest pain, while anxiety develops gradually and allows for anticipation and planning [130]. Some HCC studies include detecting stress [66, 105], stress resilience [3], social anxiety [67] and panic attacks [117].

Engagement. Engagement is defined as a three-part classification that includes emotional, cognitive, and behavioural components. The emotional component refers to a positive state of mind and satisfaction. The cognitive component involves intellectual commitment, while the behavioural component encompasses effort and participation [32]. Engagement has been extensively studied in the field of HCC, such as student engagement [37, 53, 55], emotional engagement [88], game engagement [69], and social engagement [65].

Depression. Depression is a serious psychological condition that significantly impacts a person's wellbeing and overall functioning. It is a medical illness that can cause persistent feelings of sadness, hopelessness, and a loss of interest or pleasure in activities that were once enjoyed, leading to difficulties in daily life and a decreased quality of life [33]. Depression is a commonly studied psychological construct in HCC literature, with research focusing on topics such as depression during Covid-19 [136, 137], trajectories of depression [17], etc.

Personality. Personality is a popular psychological construct that refers to the consistent patterns of thoughts, feelings, and behaviours that make a person unique. It is often considered as a predictor of performance in studies and work [64]. The most widely accepted theory of personality structure is the Big-5 personality traits, also known as the *Five-Factor Model* (FFM), suggesting five broad dimensions capture the major features of personality: extraversion, agreeableness, conscientiousness, neuroticism, and openness [72]. Researchers in HCC have extensively studied Big-5 personality traits using various data sources, such as predicting personality using smartphones [56, 148], social media [61], online videos [13].

Affect, Emotion and Mood. These constructs, though share similarities, have distinct characteristics and serve different roles in understanding human behaviour and experiences [41]. Affect is the immediate display of emotion, observable through physical expressions such as facial expressions, postures, and vocal tones. Emotions, on the other hand, is a complex internal experience involving subjective feelings, physiological changes (e.g., increased heart rate), and behavioural responses (e.g., smiling and frowning). It is typically triggered by specific events or stimuli. Mood, in contrast, is a longer-lasting emotional state that is not tied to a specific incident. It can persist for hours, days, or even longer and have a significant impact on an individual's wellbeing. Recent HCC studies for studying these constructs include: personalized mood [91], mood instability [119], mood changes [89], compound emotion [158], affect [121, 160], etc.

Cognitive Load and Mental Workload. Cognitive load refers to the extent of working memory resources used when a person engages in a task completion [52]. It embodies the mental effort necessary to learn new information or perform specific activities. Mental workload, on the other hand, is a more comprehensive term that encompasses the overall burden on the cognitive system, including working memory, attentional resources, and other cognitive functions. Related HCC studies include cognition load [151] modelling, interruption management [62], mental workload prediction [81], etc.

Loneliness. Loneliness is a complex emotion characterized by feeling isolated and alone, regardless of the number of social interactions. It is about the quality and meaning of social relationships rather than just the quantity. Loneliness can affect the mental state and cognitive ability by disturbing the processing of brain and increasing the risk of cardiovascular attacks [11]. It is also a commonly studied psychological constructs in HCC studies [144].

Flourishing. Flourishing is a multidimensional construct used in positive psychology to describe the optimal state of individuals characterized by good mental health, extending beyond mere happiness or life satisfaction. It encapsulates a prosperous condition when people experience a sense of purpose, personal growth and the realization of their potential, leading to a profound sense of fulfilment and contentment [68]. The measurement of flourishing scores has been extensively utilized in HCC studies such as [119, 144]

Fatigue. Fatigue stands as a prominent concern impacting both physical health and mental wellbeing, particularly in conditions like Multiple sclerosis (MS). MS is a neurological disorder that primarily affects young adults and has no known cure. Managing MS involves symptom control through support and therapeutic interventions. However, managing its symptoms is typically done through support and treatment [138]. Among the various symptoms experienced by MS patients, detecting fatigue is particularly notable and have been extensively studied in HCC communities [63, 138].

2.3 Reliability and Validity of Self-Report

The integrity of research findings in the field of HCC heavily relies on the chosen method for data collection. Among the various methods available, self-reporting is widely used for collecting data from human participants. However, self-report data face numerous challenges that can significantly impact the validity and reliability of the results. Validity refers to the degree to which a study accurately reflects or assesses the specific concept that the researcher intends to measure. It includes face validity (direct participant feedback), content validity (expert evaluation), and criterion validity (correlation with real-life constructs) [21]. On the other hand, reliability refers to the consistency of results obtained from an experiment, test, or any measuring procedure upon repeated trials [30]. It includes test-retest reliability (measuring stability over time), internal consistency (measuring agreement among questionnaire items), and inter-rater consistency (measuring agreement between observers). To mitigate the threats to validity and reliability, the common practice in HCC studies is to rely on established instruments like PHQ-8 [84], PHQ-4 [82], and GAD [129]) as the ground truth. However, common threats such as sampling bias, response bias and nonresponse bias should still be considered.

Especially, response bias refers to tendencies for participants to respond inaccurately or falsely to questions [51]. Common response biases include: (1) *Recall bias*: Participants may inaccurately remember past events or recollect them wrongly. (2) *Social desirability bias*: Participants may answer in a way that is favorable to others, leading to over-reporting or under-reporting. (3) *Agreement bias*: Participants tend to select statements with positive implications or agree to statements. (4) *Order effect bias*: Participant responses may vary based on the order of the questions. (5) *Mood bias*: Participants' mental state can impact their answers, leading to changes based on their mood. (6) *Central tendency bias*: Some participants consistently choose responses in the middle of the scale, avoiding extreme agreement or disagreement. (7) *Demand characteristic bias*: Participants may alter their responses based on their understanding of the survey's purpose. (8) *Random response bias*: Participants may guess or choose random answers when they are unsure or do not understand the question.

3 SELF-REPORT PRACTICES FOR HCC STUDIES AT UBICOMP

3.1 Source selection

We performed a comprehensive selection of papers from the *ACM Digital Library*¹. Our focus was on contributions from the Ubicomp conferences over the past decade (from Ubicomp '23 to Ubicomp '13), recognized as leading venues for research in ubiquitous computing and human mental wellbeing sensing. The sources for our collection were the *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* and the *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing*. We selected the papers with the term *human sensing, physiological signals, sensors, mobile sensing, wearable, stress, emotion, depression, mood, affect, prediction, engagement, cognitive load, anxiety, health, wellbeing, sensing, behaviour, and mental health*. Our search query, designed to capture this range of topics, is demonstrated below:

"query": Title, Keyword, Abstract: ("human sensing" or "physiological signals" or "sensors" or "mobile sensing" or "wearable" or "stress" or "emotion" or "depression" or "mood" or "affect" or "prediction" or "engagement" or "cognitive" or "anxiety" or "health" or "wellbeing" or "sensing" or "behaviour" or "mental health") AND AllField: (questionnaire or survey or "self-report") "filter": Conference Collections: UbiComp: Ubiquitous Computing E-Publication Date: (01/01/2013 TO 10/30/2023), Published in: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies

Selection criteria beyond keyword relevance included:

- (1) Emphasis on human-centered computing within the fields health, wellbeing, and affect prediction.
- (2) Exclusion of workshop and poster papers.
- (3) Utilization of wearable devices or smartphones for data acquisition.
- (4) Publications dated from 2013 to 2023.

3.2 Screening Criteria

An initial keyword search returned a total of 1,257 papers. After a preliminary review of titles and abstracts, 108 papers were identified that satisfied our selection criteria. Subsequent full-text review led to the exclusion of papers not providing adequate ground truth data collection methodologies. To analyze the information consistently, we tabulated critical data from each paper, such as questionnaires used, methodologies employed, types of devices utilized, and other pertinent details relevant to our study.

This process resulted in a refined collection of 49 papers. Table 1 presents the reviewed Ubicomp papers and their respective types of ground truth. The predominant method for collecting ground truth is self-reporting,

¹See the ACM Digital Library at <https://dl.acm.org/>

Table 1. List of Papers with Different Types of Ground Truth

Types of Ground Truth	Paper Counts	Reference
Established survey (unmodified)	30	[3, 6, 26, 43, 62, 63, 67, 76, 89, 91, 102, 104, 117, 119–121, 134, 136–138, 143–145, 148, 151–154, 160]
Established survey (adapted)	8	[17, 55, 59, 66, 69, 88, 147, 149]
Custom-designed survey	8	[7, 57, 79, 87, 92, 109, 142, 155, 159]
Direct observation	3	[38, 65, 105]

which was utilized in 93.87% (46 out of 49) of the papers. This preference highlights the prevalence of self-report as the primary approach in human mental wellbeing computing within the Ubicomp field. Among the self-reporting instruments, 38 out of 46 papers relied on well-established surveys for data collection, such as the *Perceived Stress Scale* (PSS), *Big Five Inventory* (BFI), and *Patient Health Questionnaire* (PHQ), among others. However, 8 out of 38 studies made modifications to these established surveys to better align them with the specific context of their research. Furthermore, a notable finding is that 8 papers developed custom-designed surveys tailored to their specific study goals. More details regarding these custom-designed surveys will be discussed in Section 3. In contrast, the utilization of live observation as a method for collecting ground truth data was relatively limited, indicating a heavier reliance on self-report data in assessing mental and wellbeing states within the Ubicomp research landscape.

3.3 Coding Procedure

Our analysis is based on the survey design principles in Arlene Fink’s *How to Conduct Surveys: A Step by Step Guide* (Sixth Edition) [46]. This comprehensive manual has become a vital resource for researchers as it provides guidance on various aspects of survey design, such as the selection of survey types, respondent inclusion criteria, survey frequency, and the intricacies of data analysis and result interpretation. The widespread acceptance and influence of this manual can be seen from its over 6000 citations to date. In addition to Fink’s manual, we also referenced other leading publications in survey research, including *Good Practice in the Conduct and Reporting of Survey Research* by Kelley et al. [75] (more than 3200 citations to date), *The Survey Handbook* by Fink [45] (more than 2900 citations to date), De et al. [31] (more than 1100 citations to date), enriching our procedural framework with recognized standards.

As a result, we coded each paper in terms of the information reported about recruitment and participants. Besides, through examining the patterns and characteristics unique to human-centred computing studies, we assess the instruments, environments, sensing measures, data collection and post-processing methods for each paper. Table 2 describes the information we captured for each study.

3.4 Result

In this subsection, we present the findings from our review, organized according to the sections of our code-book. We identify and discuss problematic issues while highlighting examples of best practices observed in the reviewed works.

3.4.1 Recruitment. In survey research, the selection of a representative sample from a well-defined sampling frame is critical for external validity, as it enables researchers to extrapolate findings to the broader population. Selecting who will be included in the sample requires careful consideration of various factors to achieve a comprehensive population profile [75]. However, our review reveals that none of the examined studies provided a

Table 2. Codebook used for the analysis of our sample

Item	Description
Recruitment	
1. Sample method	How were potential subjects identified? (e.g., random, systematic, convenience)
2. Sample Size	How was the sample size decided?
3. Recruitment Method	How, where, how many times, and by whom potential subjects were approached?
4. Participants approached	How many participants were approached?
5. Participants agreed	How many approached participants agree to participate?
6. Non-response information	How did those who agreed differ from those who did not agree with participating the study?
Participants	
7. Age	Are the range, mean and STD of participants' age reported?
8. Gender/Sex	Is the gender/sex of participants reported?
9. Ethnicity	Is the ethnicity of participants reported?
10. Occupation/job	Is the occupation/job of participants reported?
11. Illness or health care	Is the illness or health care information of participants reported?
12. Consent form	Did participants sign consent forms?
13. Compensation	Does the compensation mentioned?
14. Mechanism to leave study	Were participants informed with the mechanism to leave study?
Instruments and environment	
15. Psychological constructs	The psychological constructs studied in the research
16. Existing/new instrument	For new instruments, should provide a section outlining the steps taken to develop or test the tool, including the results of psychological testing.
17. Questionnaire/ESM	Whether the instrument is a survey-based or ESM method
18. Natural/lab settings	Does the data collected in natural settings or lab settings?
19. Sites	Does specific detail provided related to scenarios?
Sensing Measures	
20. Device type	Types and details of the sensing device
21. Commercial/custom	Whether the device is commercial or customized?
22. Sensing signals	Types of signals
23. Relation to constructs	Is the target psychological constructs measurable by the signals?
Data collection	
24. Survey administration	How was the survey administrated (e.g., telephone, interview)
25. Data collection duration	Duration of the total data collection
26. Self-report frequency	Frequency of self-report
27. Self-report guidance	Did the research guide participants to ensure effective self-report?
28. Response rate	What was the response rate?
Post-processing	
29. Ground truth establishment*	Method for establishing ground truth
30. Data quality ensurement*	Provide specific details related to measures taken to improve data quality
31. Bias discussion*	Did researchers discuss about potential bias in the data collection?

Note: Asterisks (*) indicate the items that are not derived from existing literature

full account of their participant recruitment process. While they disclosed sample sizes, all of them failed to perform essential sample size calculations, such as power analysis, which are fundamental for ensuring statistical robustness and the attainment of the study's objectives [73].

A high rate of non-response can lead to misleading conclusions that may only reflect the views of the respondents, as indicated by Kelley [75]. French [49] found that non-respondents in patient satisfaction surveys are less likely to be satisfied than people who reply. It is critical to report the response rate and address the potential differences between respondents and non-respondents, with the implications of these differences. However, based on our review, none of Ubicomp papers report the non-response rate and information of differences, which raises concerns about data imbalance and the potential for skewed predictive models. For instance, Gao et al. [55] developed a model to predict student engagement using physiological signals, yet the voluntary nature of student participation could mean those who opt-in are inherently more engaged, skewing results. Similarly, Wang et al. [148] employed mobile sensing data to predict the Big-5 personality traits; however, the likelihood that introverted individuals may opt out of participation could lead to an unbalanced model biased against introverted traits.

3.4.2 Participants. Previous studies have highlighted the importance of reporting detailed participant information such as age, gender, ethnicity, occupation, and health status in survey research [31, 45, 75]. These demographic and personal characteristics can significantly influence self-report data, affecting its reliability and validity. For instance, age differences can impact cognitive responses, which in turn affect self-report outcomes [5]. Similarly, gender differences in emotional processing and expression are well-documented [45], suggesting that gender distribution in study samples should be carefully reported and considered.

None of the papers in our sample fully described their participant information. While all papers reported the number of participants, detailed age data (range, mean, and standard deviation) was often missing. Surprisingly, 17 papers did not report any age-related information, 26 omitted age ranges, and failed to provide mean or standard deviation values. Except for 12 papers, most reported participants' gender distribution. However, only 12 papers addressed ethnicity. The inclusion of ethnicity is crucial, as cultural factors can influence perceived mental health and wellbeing perceptions in HCC studies. Occupation, another factor influencing mental health due to varying stress levels and work environment, was not mentioned in 8 papers.

The ethical aspects of research, particularly informed consent and the mechanism to leave the study, are vital for ensuring participant autonomy and ethical research conduct [31]. Our review found that 19 papers did not report information related to consent forms, and only 9 described the mechanisms for participants to withdraw from the study.

Compensation has significant impacts on the quality of self-report data. For example, the Netherlands Official Statistics used booklets with ten stamps as gifts, and the nonresponse rate fell significantly [31]. Conversely, Stone et al. [131] observed that offering a \$250 incentive led to poor data quality, attributing this to a participant pool driven primarily by financial gain rather than genuine interest. Based on our review, 24 out of 49 papers introduce compensation or incentives, which took various forms ranging from monetary payments, technology gadgets, vouchers, and non-monetary gifts.

3.4.3 Instruments and environment. In human mental and wellbeing computing studies, the identification of psychological constructs under investigation is crucial. Based on our review, the most frequently studied constructs are depression (cited in 10 papers), stress (7 papers), mood (7 papers), and engagement (5 papers). 38 studies leveraged established survey instruments. Among these, 8 papers adapted the surveys to suit their specific contexts. For instance, Gao et al. [55] made slight modifications to the *In-class Student Engagement Questionnaires (ISEQ)* [50], originally designed for university lectures (e.g., 'I feel discouraged when I worked on the activities in class'), to the high school class context (e.g., 'I feel discouraged when we worked on something'). While such adaptations can enhance the contextual relevance of the surveys, they raise concerns about reliability and validity [21]. Modifications, even minor ones, could alter the construct being measured or affect the instrument's psychometric properties. According to De et al. [31], regardless of the form or the degree of change, it is wise

Table 3. Overview of the self-report instruments in human-centred computing research

Category	Questionnaire	Variants	Papers
Anxiety	State-Trait Anxiety Inventory (STAI) [128]	40 items	[26, 59, 120, 121]
	Generalized Anxiety Disorder (GAD) [129]	GAD-7	[136]
	Social Interaction Anxiety Scale (SIAS) [100]	20 items	[67]
Stress	Perceived Stress Score (PSS) [24]	14 items	[66, 119, 144]
	Trier Social Stress Test (TSST) [80]	Lab	[121]
Engagement	In-class Student Engagement Questionnaires (ISEQ) [50]	6 items	[55]
	University Student Engagement Inventory (USEI) [99]	15 items	[88]
	Game Engagement Questionnaire (GEQ) [1]	19 items	[69]
Depression	Patient Health Questionnaire (PHQ) [83]	PHQ-4 PHQ-8 PHQ-9	PHQ-4: [147, 154] PHQ-8: [17, 147] PHQ-9: [3, 136, 137, 143, 144]
	Center for Epidemiological Studies Depression Scale (CES-D) [90]	20 items	[63]
	Beck Depression Inventory-II (BDI-II) [12]	21 items	[152–154]
	Depression Anxiety and Stress Scale (DASS) [60]	21 items	[119]
Personality	Big Five Personality (BFI) [72]	44 items	BFI-44: [76, 91, 102, 144, 148]
		60 items	BFI-60: [120]
Affect	Positive and Negative Affect (PANAS-X) [27]	10 items	[89, 120, 144, 154, 160]
Mood	Multidimensional Mood Questionnaire (MDMQ)	24 items	[43]
Cognitive load	NASA TLX [35]	6 items	[62, 134, 151]
	Shipley Scales [18]	40 items	[120]
Fatigue	Fatigue Severity Scale (FSS) [85]	9 items	[63, 138]
Loneliness	UCLA Loneliness Scale [118]	20 items	[144]
Flourishing	Flourishing Scale [36]	8 items	[119, 144]
Panic	Panic Disorder Severity Scale (PDSS) [123]	7 items	[117]
	Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) [25]	N/A	[117]
Functioning	Quality of Life in Neurological Disorders (Neuro-QoL) [20]	8 items	[6]

to consider adapted questions as new questions and to test them accordingly. However, none of the papers that utilized adopted established surveys examined the effects of these adaptations.

8 papers utilized custom-designed self-report tools, often to ask quick questions using the ESM method. According to Kelley et al. [75], ‘if a new survey tool is used, an entire section should be used to describe the steps undertaken to develop and test the tool, including psychometric assessment results’. Unfortunately, in the 8 papers analyzed, custom-designed self-report tools were used without any test of reliability and validity. Recognizing the significance of employing well-established survey instruments in HCC studies, we have compiled a list of commonly utilized self-report instruments in Table 3, for reference and use in future research.

3.4.4 Sensing Measures. Unlike traditional survey research, HCC studies predominantly rely on self-report data as the ground truth, complemented by sensing data as predictive indicators. The initial step involves confirming the psychological constructs to be analyzed. Subsequently, it's crucial to determine the types of signals to be collected, methodologies for data acquisition, and strategies to assure signal integrity. A critical aspect to be addressed is the feasibility of measuring targeted psychological constructs using these signals. The selection of appropriate signals is often contingent on the nature of the psychological constructs and the signal types. Broadly categorizing, these signals include physiological data from wearable devices, mobile sensing data from smartphones or desktops, environmental sensors, or a combination. Table 4 shows an overview of the devices used in HCC studies at Ubicomp.

While all papers introduced the device and collected signals, 31 papers failed to explicitly explain all of the used signals to the psychological constructs under investigation, especially for papers utilized smartphone sensing. Among the physiological signals, Electrodermal Activity (EDA) demonstrates promise in reflecting the activation of *Sympathetic Nervous System* (SNS), mediating involuntary responses to emotion arousal, thereby serving as a potential measure for affective and cognitive states [15].

3.4.5 Data Collection. The quality of research data is closely tied to the method of survey administration chosen. Different modes of survey administration, such as phone, online, or in-person, can introduce mode effects, where the method itself influences the responses obtained. It is crucial for researchers to report the mode of survey administration to ensure transparency and accurately interpret the findings. Surprisingly, 14 Ubicomp papers failed to report the methods of survey administration. In addition to the mode of administration, the frequency of surveys can also impact the quality of responses, particularly in natural settings. Most Ubicomp papers have wisely adopted a daily data collection approach. However, one Ubicomp paper asked participants to report 15 times a day, and another paper required participants to report five times a day.

To mitigate bias and enhance response rates, it is crucial to provide clear guidance to participants regarding the research objectives and expectations. Only 10 Ubicomp papers reported providing such guidance to participants. Furthermore, only 3 papers reported the response rates, which is an important metric for assessing the representativeness and reliability of the collected data.

3.4.6 Post-processing. Different from traditional survey research, HCC studies require special considerations of ground truth. 6 Ubicomp papers have implemented specific methods to validate the collected ground truth. For example, one paper normalized reported mood using z-score, as '*some people may be more positive or negative about their mood*'. Another paper used linear interpolation computed for missing values calculation, which served as the ground truth. To ensure data quality, 11 Ubicomp papers adopted various measures. One paper '*allowed participants to choose the data to collect and use*', another one mentioned their ability to reach participants as needed. Additionally, one paper reported that '*participants were unaware of the true study purpose to prevent bias*'. Another paper set a trap question with a known answer (e.g., location) and assessed reliability based on answer completion time and the trap question response. Instead, one paper accommodated participants who preferred a pen-and-paper survey over using an app, emphasizing flexibility in data collection methods. However, most Ubicomp papers didn't transparently disclose whether specific measures were taken to ensure the quality of collected data. Regarding bias discussions, 22 Ubicomp papers acknowledged potential biases, such as '*when participants were continuously asked to take survey they may be distracted*', '*be not sure whether the participants reported a motion truthfully or not*', '*ground truth collection may be labour intensive, expensive and somewhat inaccurate*', etc.

Table 4. Overview of the devices used in HCC studies at Ubicomp

<i>Device</i>	<i>Category</i>	<i>Parameters</i>	<i>Num.</i>	<i>Ref.</i>
<i>Smartphone</i>	Smartphone	Physical activity, sleep, app usage	23	[17, 43, 67, 76, 89, 92, 102, 104, 119, 136, 137, 142–145, 147–149, 152–155, 159]
<i>Empatica E4</i>	Wristband	BVP, EDA, HR, ST	9	[38, 55, 57, 59, 69, 87, 88, 121, 151]
<i>Garmin Vivosmart</i>	Smartwatch	Sleep, oxygen level, HR, breathing, physical activity	3	[104, 109, 120]
<i>Polar H7</i>	Chest strap	Activity speed, distance, HR	4	[26, 79, 92, 105]
<i>Q sensor</i>	Wrist sensor	EDA, ST, actigraphy	2	[62, 65]
<i>Fitbit</i>	Smartwatch	Sleep, physical activity	2	[153, 154]
<i>Fitbit Flex 2</i>	Smartwatch	Activity, calories burned, distance, sleep, steps	1	[152]
<i>Fitbit Charge 2</i>	Smartwatch	Activity, calories burned, distance, HR, sleep, steps	1	[3]
<i>GENEActiv</i>	Smartwatch	Physical activity, sleep, everyday behaviour	1	[63]
<i>Withings Aura</i>	Bedside unit	Sleep, HR	1	[138]
<i>Withings Activite Steel</i>	Smartwatch	Activity, calories burned, distance, sleep, steps	1	[138]
<i>Withings Body Cardio</i>	Smartscale	Body composition, HR	1	[138]
<i>Sociometric badge</i>	Badge	Face-to-face interaction, conversation, physical proximity, physical activity	1	[160]
<i>Samsung Galaxy Tab S2</i>	Tablet	Physical activity, sleep, app usage	1	[69]
<i>Microsoft Kinect</i>	Add-on device	Postures and body movement	1	[69]
<i>Microsoft Band 2</i>	Smartwatch	HR, EDA, ST, physical activity	1	[147]
<i>Moto 360</i>	Smartwatch	HR, step count	1	[43]
<i>ekgMove</i>	Chest belt	ECG, HR, HRV, steps, activity, and energy expenditure	1	[43]
<i>Zephyr BioPatchTM</i>	Body wear	HR, respiratory rate, physical activity	1	[117]
<i>AutoSense</i>	Chest belt	HR, EDA, ECG, lung volume, breathing rate, ST	1	[66]
<i>HTC Vive Pro Eye</i>	Headset	Eye tracking	1	[151]
<i>BodyMedia Sensewear Pro3</i>	Armband	EDA, ECG, ST, sweat, heat flux	1	[134]
<i>Zephyr Bioharness BT</i>	Chest strap	ECG, HR, breathing rate, oxygen level	1	[134]
<i>Lightstone Fingertip Sensor</i>	Finger sensor	EDA, HR	1	[134]
<i>Neulog GSR module</i>	Finger sensor	GSR	1	[79]
<i>Chillband</i>	Wristband	EDA, ST, ACC	1	[155]
<i>Healthpatch</i>	Chest strap	ECG, ACC	1	[155]
<i>OMsignal</i>	Body wear	ECG	1	[155]
<i>PRO-Diary</i>	Smartwatch	Actigraphy	1	[6]

4 RELIABILITY OF SELF-REPORT AS GROUND TRUTH FOR HUMAN-CENTRED COMPUTING

Recent HCC studies have highlighted significant concerns regarding the reliability of self-reported data, primarily due to improper self-reporting practices and inherent limitations in survey research methods. While it is difficult to obtain the absolute truth about human mental wellbeing, researchers have employed various methods to address these issues. These methods include investigating the misalignment between self-report and well-established technologies [54], comparing self-report with physiological measurements [54], comparing the performance of prediction models [29], or directly comparing self-report with real ground truths, such as monitoring security behaviours through recorded videos citewash2017can.

Gao et al. [54] conducted a study to examine the reliability of self-report measures in establishing ground truth for predicting student engagement. They discovered that the physiological measurement of engagement and the perceived engagement reported by individuals were not always consistent. This finding suggests that relying solely on subjective annotations may introduce potential unreliability when establishing ground truth. In a related study, Kaur [74] explored the relationship between *Automated Emotion Recognition* (AER) technologies and self-reported affect in the context of information work at technology companies. They revealed a misalignment between the continuous observed emotion from the AER tool and the discrete reported affect by individuals.

Das et al. [28] found that when what people feel differs from what people say they feel. They identified a semantic gap that hinders accurate predictions in emotion recognition. Furthermore, they highlighted that predicting mental wellbeing using passive data, such as offline sensor data or online social media, is influenced by how the ground truth is measured, whether through objective arousal measurement or self-report. In another study, Wash et al. [150] collected behavioural data and survey responses from 122 participants. They discovered that only a small number of behaviours, particularly those related to tasks that require individuals to perform specific, regular actions, exhibited non-zero correlations. Interestingly, several important security behaviours that were directly monitored did not align with self-reported responses accurately. They concluded that self-report measures are reliable only for certain behaviours. However, it is worth noting that monitoring security behaviours leans more towards physical measurements rather than psychological ones, making them easier to quantify.

5 IMPROVING GROUND TRUTH MEASUREMENT IN HCC STUDIES

Improving the practice of obtaining ground truth data is crucial for HCC studies. To achieve this, efforts have been categorized into two directions: firstly, enhancing engagement and response rate in self-report mechanism, and secondly, exploring more flexible and innovative methods for collecting ground truth data.

5.1 Enhancing Engagement and Response Rate

Self-report surveys, known for their convenience and cost-effectiveness, can suffer from low response rates, undermining their validity. To enhance response rates, various efforts have been made to improve engagement, such as the use of *Casual Affective Triggers* (CATs) [22], conversational agents [78], gamification [140] and incentivizing participation [47, 125, 156].

Chounta et al. [22] suggest the use of CATs, such as engaging images or interactive artifacts, in the survey interface or email notifications. They found that CATs can improve emotional engagement, thereby motivating respondents to participate and complete surveys. This strategy can lead to higher response rates and more robust, representative data. The development of conversational agents helps to overcome low engagement by simulating human-like conversations and interacting with survey participants interactively, thereby improving the response quality [78]. The friendly and active nature of a conversational agent can help to create a more comfortable and personalized survey experience. Participants may feel more at ease when answering questions, leading to increased response quality and potentially reducing response bias.

Employing gamification techniques in surveys, as described by Van den Broeck et al. [140], can significantly enhance user engagement. By incorporating game-like elements such as animations and leaderboards, the survey experience becomes more enjoyable, encouraging more accurate and reliable responses. The use of incentives to increase response and participation rates is a common practice. Studies [47, 125, 156] have shown that incentives and follow-up messages can improve participation rates. The Bayesian Truth Serum (BTS) technique [9] addresses potential biases introduced by incentives, ensuring the reliability of responses.

5.2 Enabling Flexible Ways for Collecting Ground Truth

Diverse approaches exist for collecting ground truth data in the domain of mental wellbeing recognition, particularly in the areas of emotion [116, 122] and stress [4, 10]. Advances in physical and behavioural sciences have enabled the development of algorithms capable of accurately recognizing simple emotions such as happy, sad, angry, etc. These algorithms may utilize various modalities such as video [157], audio [110], and text [101], the latter often referred to as ‘*sentiment analysis*’. Initially, these algorithms relied on annotated data as ground truth. However, due to their high accuracy in emotion detection, they have increasingly been adopted as a new form of ground truth for continuous emotion annotation in this field [74, 132].

It is crucial to acknowledge that while methods developed for emotion recognition, particularly those utilizing video and audio data, have shown promise, their effectiveness in addressing more complex psychological constructs like depression and personality traits remains limited. These algorithms identifying these more nuanced psychological states are not yet sufficient to consider them as reliable ground truth sources [23, 97]. This limitation indicates the need for ongoing research and development in the field. However, it also presents unique opportunities: these methods can serve as supplementary ground truth sources, contributing to a more holistic understanding of mental states and wellbeing. To provide a clearer perspective, we present a summary of the current practices in collecting ground truth data across various methods below.

Induction Test. Induction tests are designed to elicit physiological responses to emotions like stress, serving as a widely-accepted benchmark for establishing the ground truth about these emotional states [19, 40, 106]. Key tests include the *Cold Pressor Test*, *Trier Social Stress Test*, *Montreal Imaging Stress Task*, *Maastricht Acute Stress Test*, *Paced Auditory Serial Addition Task*, and *Mannheim Multicomponent Stress Test* [10]. However, to capture the complexities of real-world emotions, Almazrouei et al. [4] conducted online stress induction tests for natural assessment. Additionally, Larradet et al. [86] introduced a mobile application linked to wearable devices for continuous monitoring of emotional states, using the *Ortony, Clore, and Collins* (OCC) model to prompt users to record emotional triggers, thereby bridging the gap between lab and real-world settings.

Vision-Based Detection. This method utilizes facial expressions, eye movements, and physical behaviours like head movement and pupil size variation to analyze human mental states such as stress and emotions [58, 74]. Some established algorithms include the *Facial Action Coding System* (FACS) for categorizing facial expressions [42], Facebook’s *DeepFace* for facial recognition [133], Google’s *Cloud Vision API* for emotional analysis in images. Due to the high accuracy of vision-based emotion detection, it is increasingly used as ground truth in HCC studies. For instance, Tag et al. [132] utilizes the Affectiva API and the AWARE framework to monitor emotions through smartphone cameras. This application considers the captured emotions as ground truth for analyzing emotion trajectories for smartphone users.

Text-Based Detection. Text-based emotion detection, commonly known as sentiment analysis, utilizes algorithms and natural language processing (NLP) techniques to identify emotions from textual content. The Google Cloud Natural Language API is a prominent tool in this domain, utilising machine learning for nuanced sentiment analysis. Advanced models such as BERT [34] have significantly improved the accuracy of emotion detection in text, making them invaluable in HCC studies as supplementary ground truth sources. For instance, Terzimehic et al. [135] utilized *Ortony, Clore, and Collins* (OCC) model to derive emotions from love or breakup

letters to smartphones, which served as the ground truth for analyzing the emotional shifts experienced by smartphones before and during the COVID-19 pandemic.

Speech-Based Detection. Speech-based emotion detection differs from text-based analysis as it involves interpreting vocal cues and prosody to discern emotional states. This method analyzes the tone, pitch, and rhythm of speech, which convey a wealth of emotional information beyond the spoken words. Popular algorithms and tools such as *Mel-Frequency Cepstral Coefficients* (MFCCs), OpenSMILE toolkit [44] and the Emo-DB database [16], have been utilized to capture the emotion during speech.

Multimodal Detection. Multimodal approaches integrate various data types to provide a more comprehensive analysis of emotions. Sharma et al. [122] exemplify this by combining video, social media, and Twitter feedback to assess student emotions in educational settings. Non-verbal behaviours in job interviews are analyzed by studying facial expressions, speech patterns, and prosody to evaluate performance [108]. Rodrigues et al. [116] developed a multimodal system for detecting stress and fatigue in drivers, integrating physiological, psychological, and georeferenced data. The ground truth derived from these multimodal sources enables the study of emotions in a continuous, dynamic context, offering a richer understanding of emotional states across various scenarios.

6 RESEARCH GAPS AND FUTURE WORK

Through our review, we have identified the challenges in the practice of self-reporting in HCC studies. To contribute to this field, future researchers can explore several directions.

Improve Self-Report Data Collection: Self-report practices in HCC studies demonstrate distinct patterns compared to traditional survey research. Ubicomp researchers could consider improving the self-report data collection process by addressing factors that influence the quality and quantity of data. This can involve engaging interactions (e.g., gamification [140], conversational chatbot [78]), designing better incentive mechanisms, determining optimal methods (e.g., time, frequency, location, user cognition load) for questionnaire delivery, and developing more suitable tools for HCC research, like *Photographic Affect Meter* [113].

Enhance Self-Report Data Quality: The common practice in HCC studies is to use self-report data directly as ground truth without assessing its credibility. Future researchers can analyse and evaluate the quality of self-report data during the post-processing periods. This could be achieved by incorporating special and trap questions during data collection [93], analysing completion time to assess response certainty [98], measuring response biases and mitigating low-quality data, thereby enhancing overall data quality used as ground truth.

Flexible Ground Truth Measurement: It is important to develop more adaptable methods for collecting ground truth data. These approaches should emphasize privacy protection, minimize the burden on users, and avoid the need for extensive installations or specialized equipment. While fully replacing self-report data may not be feasible in the short term, integrating multiple data collection modalities as discussed in Section 5.2 could address some of the limitations inherent in self-reporting.

Foster Replicable Practices: Reproducibility is a significant challenge in HCC studies due to the nature of human-based data collection. To enhance reproducibility, it is crucial to establish standardised protocols for data collection and analysis. When collecting ground truth data, researchers should report various factors, including questionnaire delivery methods, frequency, user incentives, guidance, etc. By focusing on these elements, researchers can increase the reproducibility and reliability of their research findings, fostering a more robust and trustworthy body of knowledge in the field.

7 CONCLUSION

This paper presents a comprehensive methodological review of Ubicomp papers that utilize self-report as the ground truth for human mental and wellbeing computing. Our analysis identifies deficiencies and inconsistencies

in current self-reporting practices within the Ubicomp community. We have developed a set of guidelines that aim to improve and standardize self-reporting practices, thereby enhancing the reliability and credibility of future studies in human mental and wellbeing computing.

Furthermore, we address the urgent need for methodological evolution and advocate for a shift from traditional self-reporting methods towards incorporating more reliable and diverse approaches, such as physiological data analysis and advanced data processing techniques. This shift is critical for advancing the accuracy and applicability of HCC research, particularly in real-world scenarios. By embracing these changes, the field of HCC can make more substantial and impactful contributions to the understanding and enhancement of human mental wellbeing.

ACKNOWLEDGMENTS

We sincerely thank anonymous reviewers for their effort to improve the paper.

REFERENCES

- [1] 2009. The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology* 45, 4 (2009), 624–634. <https://doi.org/10.1016/j.jesp.2009.02.016>
- [2] Chadia Abras, Diane Maloney-Krichmar, Jenny Preece, et al. 2004. User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications* 37, 4 (2004), 445–456.
- [3] Dan Adler, Vincent Tseng, Gengmo Qi, Joseph Scarpa, Srijan Sen, and Tanzeem Choudhury. 2021. Identifying Mobile Sensing Indicators of Stress-Resilience. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5 (06 2021), 1–32. <https://doi.org/10.1145/3463528>
- [4] Mohammed Almazrouei, Ruth Morgan, and Itiel Dror. 2022. A method to induce stress in human subjects in online research environments. *Behavior Research Methods* (07 2022). <https://doi.org/10.3758/s13428-022-01915-3>
- [5] Frank M Andrews and A Regula Herzog. 1986. The quality of survey data as related to age of respondent. *J. Amer. Statist. Assoc.* 81, 394 (1986), 403–410.
- [6] Anindya Das Antar, Anna Kratz, and Nikola Banovic. 2023. Behavior Modeling Approach for Forecasting Physical Functioning of People with Multiple Sclerosis. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 1, Article 7 (mar 2023), 29 pages. <https://doi.org/10.1145/3580887>
- [7] Riku Arakawa, Karan Ahuja, Kristie Mak, Gwendolyn Thompson, Sam Shaaban, Oliver Lindhiem, and Mayank Goel. 2023. LemurDx: Using Unconstrained Passive Sensing for an Objective Measurement of Hyperactivity in Children with No Parent Input. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 2, Article 46 (jun 2023), 23 pages. <https://doi.org/10.1145/3596244>
- [8] Ferhat Attal, Samer Mohammed, Mariam Dedabrishvili, Faicel Chamroukhi, Latifa Oukhellou, and Yacine Amirat. 2015. Physical human activity recognition using wearable sensors. *Sensors* 15, 12 (2015), 31314–31338.
- [9] Aurélien Baillon, Han Bleichrodt, and Georg D. Granic. 2022. Incentives in surveys. *Journal of Economic Psychology* 93 (2022), 102552. <https://doi.org/10.1016/j.joep.2022.102552>
- [10] Anjana Bali and Amteshwar Singh Jaggi. 2015. Clinical experimental stress studies: methods and assessment. *Reviews in the Neurosciences* 26 (2015), 555 – 579.
- [11] Roy Baumeister and Mark Leary. 1995. The Need to Belong: Desire for Interpersonal Attachments as a Fundamental Human Motivation. *Psychological bulletin* 117 (06 1995), 497–529. <https://doi.org/10.1037/0033-2909.117.3.497>
- [12] Aaron T. Beck, Robert A. Steer, Roberta Ball, and William F. Ranieri. 1996. Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. *Journal of personality assessment* 67 3 (1996), 588–97.
- [13] Joan-Isaac Biel, Lucía Teijeiro-Mosquera, and Daniel Gatica-Perez. 2012. Facetube: predicting personality from facial expressions of emotion in online conversational video. In *Proceedings of the 14th ACM international conference on Multimodal interaction*. 53–56.
- [14] Fionneke M Bos, Robert A Schoevers, and Marije aan het Rot. 2015. Experience sampling and ecological momentary assessment studies in psychopharmacology: a systematic review. *European Neuropsychopharmacology* 25, 11 (2015), 1853–1864.
- [15] Jason J Braithwaite, Derrick G Watson, Robert Jones, and Mickey Rowe. 2013. A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology* 49, 1 (2013), 1017–1034.
- [16] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, Benjamin Weiss, et al. 2005. A database of German emotional speech.. In *Interspeech*, Vol. 5. 1517–1520.
- [17] Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2015).

- [18] Noelle E. Carlozzi. 2011. *ShIPLEY Institute of Living Scale*. Springer New York, New York, NY, 2287–2289. https://doi.org/10.1007/978-0-387-79948-3_1070
- [19] Rossana Castaldo, Luis Montesinos, Paolo Melillo, Sebastiano Massaro, and Leandro Pecchia. 2017. To What Extent Can We Shorten HRV Analysis in Wearable Sensing? A Case Study on Mental Stress Detection.
- [20] D. Cella, J.-S. Lai, C.J. Nowinski, D. Victorson, A. Peterman, D. Miller, F. Bethoux, A. Heinemann, S. Rubin, J.E. Cavazos, A.T. Reder, R. Sufit, T. Simuni, G.L. Holmes, A. Siderowf, V. Wojna, R. Bode, N. McKinney, T. Podrabsky, K. Wortman, S. Choi, R. Gershon, N. Rothrock, and C. Moy. 2012. Neuro-QOL. *Neurology* 78, 23 (2012), 1860–1867. <https://doi.org/10.1212/WNL.0b013e318258f744> arXiv:<https://n.neurology.org/content/78/23/1860.full.pdf>
- [21] Bernard C. K. Choi and Anita W. P. Pak. 2004. A Catalog of Biases in Questionnaires. *Preventing Chronic Disease* 2 (2004).
- [22] Irene-Angelica Chounta and Alexander Nolte. 2022. The CAT Effect: Exploring the Impact of Casual Affective Triggers on Online Surveys' Response Rates (*CHI '22*). Association for Computing Machinery, New York, NY, USA, Article 583, 13 pages. <https://doi.org/10.1145/3491102.3517481>
- [23] James A Coan and John JB Allen. 2007. *Handbook of emotion elicitation and assessment*. Oxford university press.
- [24] Sheldon Cohen, Tom P Kamarck, and Robin J. Mermelstein. 1983. A global measure of perceived stress. *Journal of health and social behavior* 24 4 (1983), 385–96.
- [25] John Cooper. 2001. Diagnostic and statistical manual of mental disorders (4th edn, text revision)(DSM–IV–TR) Washington, DC: American Psychiatric Association 2000. 943 pp.£ 39.99 (hb). ISBN 0 89042 025 4. *The British Journal of Psychiatry* 179, 1 (2001), 85–85.
- [26] Jean Dos Reis Costa, Alexander Travis Adams, Malte F. Jung, François Guimbretière, and Tanzeem Choudhury. 2016. EmotionCheck: leveraging bodily signals and false feedback to regulate our emotions. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2016).
- [27] John R. Crawford and Julie D. Henry. 2004. The positive and negative affect schedule (PANAS): construct validity, measurement properties and normative data in a large non-clinical sample. *The British journal of clinical psychology* 43 Pt 3 (2004), 245–65.
- [28] Vedant Das Swain, Victor Chen, Shrija Mishra, Stephen M. Mattingly, Gregory D. Abowd, and Munmun De Choudhury. 2022. Semantic Gap in Predicting Mental Wellbeing through Passive Sensing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (*CHI '22*). Association for Computing Machinery, New York, NY, USA, Article 374, 16 pages. <https://doi.org/10.1145/3491102.3502037>
- [29] Vedant Das Swain, Victor Chen, Shrija Mishra, Stephen M Mattingly, Gregory D Abowd, and Munmun De Choudhury. 2022. Semantic Gap in Predicting Mental Wellbeing through Passive Sensing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [30] R. S. Davies. 2020. *Designing Surveys for Evaluations and Research*. https://edtechbooks.org/designing_surveys
- [31] Edith D De Leeuw, Joop Hox, and Don Dillman. 2012. *International handbook of survey methodology*. Routledge.
- [32] Triparna de Vreede, Stephanie Andel, Gert-Jan de Vreede, Paul Spector, Vivek Singh, and Balaji Padmanabhan. 2019. What is Engagement and How Do We Measure It? Toward a Domain Independent Definition and Scale. <https://doi.org/10.24251/HICSS.2019.092>
- [33] What Causes Depression. 2012. what is depression? *World Health Organization* (2012).
- [34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [35] Hannes Devos, Kathleen Gustafson, Pedram Ahmadnezhad, Ke Liao, Jonathan Mahnken, William Brooks, and Jeffrey Burns. 2020. Psychometric Properties of NASA-TLX and Index of Cognitive Activity as Measures of Cognitive Workload in Older Adults. *Brain Sciences* 10 (12 2020), 994. <https://doi.org/10.3390/brainsci10120994>
- [36] Ed Diener, Derrick Wirtz, and William Tov. 2010. New measures of well-being: Flourishing and positive and negative feelings. *Soc Indic Res* 39 (01 2010), 247–266.
- [37] Betsy DiSalvo, Dheeraj Bandaru, Qiaosi Wang, Hong Li, and Thomas Plötz. 2022. Reading the Room: Automated, Momentary Assessment of Student Engagement in the Classroom: Are We There Yet? *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–26.
- [38] Betsy DiSalvo, Dheeraj Bandaru, Qiaosi Wang, Hong Li, and Thomas Plötz. 2022. Reading the Room: Automated, Momentary Assessment of Student Engagement in the Classroom: Are We There Yet? *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 112 (sep 2022), 26 pages. <https://doi.org/10.1145/3550328>
- [39] Andrius Dzedzickis, Artūras Kaklauskas, and Vytautas Bucinskas. 2020. Human emotion recognition: Review of sensors and methods. *Sensors* 20, 3 (2020), 592.
- [40] Begum Egilmez, Emirhan Poyraz, Wenting Zhou, Gokhan Memik, Peter Dinda, and Nabil Alshurafa. 2017. UStress: Understanding college student subjective stress using wrist-based passive sensing. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. 673–678. <https://doi.org/10.1109/PERCOMW.2017.7917644>
- [41] Panteleimon Ekkekakis. 2013. *The measurement of affect, mood, and emotion: A guide for health-behavioral research*. Cambridge University Press.
- [42] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

- [43] Anja Exler, Andrea Schankin, Christoph Klebsattel, and Michael Beigl. 2016. A wearable system for mood assessment considering smartphone features and data from mobile ECGs. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (2016).
- [44] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.
- [45] Arlene Fink. 2003. *The survey handbook*. sage.
- [46] Arlene Fink. 2015. *How to conduct surveys: A step-by-step guide*. Sage Publications.
- [47] Andrew T. Fiore, Coye Cheshire, Lindsay Shaw Taylor, and G.A. Mendelsohn. 2014. Incentives to Participate in Online Research: An Experimental Examination of "Surprise" Incentives. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) (*CHI '14*). Association for Computing Machinery, New York, NY, USA, 3433–3442. <https://doi.org/10.1145/2556288.2557418>
- [48] Nick Fox, Nigel Mathers, and Amanda Hunn. 2000. *Surveys and Questionnaires*. 77–112.
- [49] Kate French. 1981. Methodological considerations in hospital patient opinion surveys. *International journal of nursing studies* 18, 1 (1981), 7–32.
- [50] Kathryn A Fuller, Nilushi S Karunaratne, Som Naidu, Betty Exintaris, Jennifer L Short, Michael D Wolcott, Scott Singleton, and Paul J White. 2018. Development of a Self-report Instrument for Measuring in-class Student Engagement Reveals that Pretending to Engage is a Significant Unrecognized Problem. *PLoS ONE* 13, 10 (2018), e0205828.
- [51] Adrian Furnham and Monika Henderson. 1982. The good, the bad and the mad: Response bias in self-report measures. *Personality and Individual Differences* 3, 3 (1982), 311–320.
- [52] Edith Galy, Magali Cariou, and Claudine Mélan. 2011. What is the relationship between mental workload factors and cognitive load types? *International journal of psychophysiology : official journal of the International Organization of Psychophysiology* 83 (10 2011), 269–75. <https://doi.org/10.1016/j.ijpsycho.2011.09.023>
- [53] Nan Gao, Mohammad Saiedur Rahaman, Wei Shao, Kaixin Ji, and Flora D Salim. 2022. Individual and group-wise classroom seating experience: Effects on student engagement in different courses. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–23.
- [54] Nan Gao, Mohammad Saiedur Rahaman, Wei Shao, and Flora D Salim. 2021. Investigating the reliability of self-report data in the wild: The quest for ground truth. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*. 237–242.
- [55] Nan Gao, Wei Shao, Mohammad Saiedur Rahaman, and Flora D Salim. 2020. n-Gage: Predicting in-class Emotional, Behavioural and Cognitive Engagement in the Wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–26.
- [56] Nan Gao, Wei Shao, and Flora D Salim. 2019. Predicting Personality Traits from Physical Activity Intensity. *Computer* 52, 7 (2019), 47–56.
- [57] Shkurta Gashi, Elena Di Lascio, and Silvia Santini. 2019. Using Unobtrusive Wearable Sensors to Measure the Physiological Synchrony Between Presenters and Audience Members. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3 (2019), 1 – 19.
- [58] Giorgos Giannakakis, Matthew Padiaditis, Dimitris Manousos, Eleni Kazantzaki, Franco Chiarugi, Panagiotis G. Simos, Kostas Marias, and Manolis Tsinakakis. 2017. Stress and anxiety detection using facial cues from videos. *Biomed. Signal Process. Control* 31 (2017), 89–101.
- [59] Martin Gjoreski, Hristijan Gjoreski, Mitja Luvstrek, and Matjazv Gams. 2016. Continuous stress detection using a wrist device: in laboratory and real life. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (2016).
- [60] Andrew T. Gloster, Howard M. Rhoades, Diane M. Novy, Jens Klotsche, Ashley C Senior, Mark E. Kunik, Nancy L. Wilson, and Melinda A. Stanley. 2008. Psychometric properties of the Depression Anxiety and Stress Scale-21 in older primary care patients. *Journal of affective disorders* 110 3 (2008), 248–59.
- [61] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE, 149–156.
- [62] Nitesh Goyal and Susan R. Fussell. 2017. Intelligent Interruption Management Using Electro Dermal Activity Based Physiological Sensor for Collaborative Sensemaking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 52 (sep 2017), 21 pages. <https://doi.org/10.1145/3130917>
- [63] Gabriel Guo, Hanbin Zhang, Liuyi Yao, Huining Li, Chenhan Xu, Zhengxiong Li, and Wenyao Xu. 2022. MSLife: Digital Behavioral Phenotyping of Multiple Sclerosis Symptoms in the Wild Using Wearables and Graph-Based Statistical Analysis. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 158 (dec 2022), 35 pages. <https://doi.org/10.1145/3494970>

- [64] Dirk Hellhammer, Arthur Stone, Juliane Hellhammer, and Joan Broderick. 2010. Measuring Stress. *Encyclopedia of behavioural neuroscience* (12 2010), 186–191. <https://doi.org/10.1016/B978-0-08-045396-5.00188-3>
- [65] Javier Hernández, Ivan Riobo, Agata Rozga, Gregory D. Abowd, and Rosalind W. Picard. 2014. Using electrodermal activity to recognize ease of engagement in children during social interactions. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2014).
- [66] Karen Hovsepian, Mustafa al’Absi, Emre Ertin, Tom P Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: towards a gold standard for continuous stress assessment in the mobile environment. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2015).
- [67] Yu Huang, Haoyi Xiong, Kevin Leach, Yuyan Zhang, Philip I. Chow, Karl C. Fua, Bethany A. Teachman, and Laura E. Barnes. 2016. Assessing social anxiety using gps trajectories and point-of-interest data. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2016).
- [68] Felicia A. Huppert and Timothy T. C. So. 2011. Flourishing Across Europe: Application of a New Conceptual Framework for Defining Well-Being. *Social Indicators Research* 110 (2011), 837 – 861.
- [69] Sinh Huynh, Seungmin Kim, Jeonggil Ko, Rajesh Krishna Balan, and Youngki Lee. 2018. EngageMon: Multi-Modal Engagement Sensing for Mobile Games. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2 (2018), 13:1–13:27.
- [70] Eider Irazoki, Leslie Maria Contreras-Somoza, José Miguel Toribio-Guzmán, Cristina Jenaro-Río, HenriëTte Van der Roest, and Manuel A Franco-Martín. 2020. Technologies for cognitive training and cognitive rehabilitation for people with mild cognitive impairment and dementia. A systematic review. *Frontiers in psychology* 11 (2020), 648.
- [71] Alejandro Jaimes, Daniel Gatica-Perez, Nicu Sebe, and Thomas S Huang. 2007. Guest Editors’ Introduction: Human-Centered Computing—Toward a Human Revolution. *Computer* 40, 5 (2007), 30–34.
- [72] Oliver P. John and Sanjay Srivastava. 1999. The Big Five Trait taxonomy: History, measurement, and theoretical perspectives.
- [73] Hyun Kang. 2021. Sample size determination and power analysis using the G* Power software. *Journal of educational evaluation for health professions* 18 (2021).
- [74] Harmanpreet Kaur, Daniel McDuff, Alex C Williams, Jaime Teevan, and Shamsi T Iqbal. 2022. “I didn’t know I looked angry”: Characterizing observed emotion and reported affect at work. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [75] Kate Kelley, Belinda Clark, Vivienne Brown, and John Sitzia. 2003. Good practice in the conduct and reporting of survey research. *International Journal for Quality in health care* 15, 3 (2003), 261–266.
- [76] Mohammed Khwaja, Sumer S. Vaid, Sara Zannone, Gabriella M. Harari, A. Aldo Faisal, and Aleksandar Matic. 2019. Modeling Personality vs. Modeling Personalidad: In-the-Wild Mobile Data Analysis in Five Countries Suggests Cultural Impact on Personality Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 88 (sep 2019), 24 pages. <https://doi.org/10.1145/3351246>
- [77] Monique F Kilkenny and Kerin M Robinson. 2018. Data quality: “Garbage in—garbage out”. , 103–105 pages.
- [78] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (CHI ’19). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3290605.3300316>
- [79] 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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 91 (sep 2019), 22 pages. <https://doi.org/10.1145/3351249>
- [80] Clemens Kirschbaum, Karl Martin Pirke, and Dirk H. Hellhammer. 1993. The ‘Trier Social Stress Test’—a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology* 28 1-2 (1993), 76–81.
- [81] Thomas Kosch, Mariam Hassib, Daniel Buschek, and Albrecht Schmidt. 2018. Look into my eyes: using pupil dilation to estimate mental workload for task complexity adaptation. In *Extended abstracts of the 2018 chi conference on human factors in computing systems*. 1–6.
- [82] Kurt Kroenke, Robert L. Spitzer, Janet B.W. Williams, and Bernd Löwe. 2009. An Ultra-brief Screening Scale for Anxiety and Depression: The PHQ-4. *Psychosomatics* 50, 6 (2009), 613–621.
- [83] Kurt Kroenke, Robert L. Spitzer, Janet B.W. Williams, and Bernd Löwe. 2010. The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. *General Hospital Psychiatry* 32, 4 (2010), 345–359. <https://doi.org/10.1016/j.genhosppsych.2010.03.006>
- [84] Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B.W. Williams, Joyce T. Berry, and Ali H. Mokdad. 2009. The PHQ-8 as a Measure of Current Depression in the General Population. *Journal of Affective Disorders* 114, 1 (2009), 163–173.
- [85] Lauren B. Krupp, Nicholas G. Larocca, Joanne Muir-Nash, and Alfred D. Steinberg. 1989. The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus. *Archives of neurology* 46 10 (1989), 1121–3.
- [86] Fanny Larradet, Radoslaw Niewiadomski, Giacinto Barresi, and Leonardo S. Mattos. 2019. Appraisal Theory-Based Mobile App for Physiological Data Collection and Labelling in the Wild. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*

- (London, United Kingdom) (*UbiComp/ISWC '19 Adjunct*). Association for Computing Machinery, New York, NY, USA, 752–756. <https://doi.org/10.1145/3341162.3345595>
- [87] Elena Di Lascio, Shkurta Gashi, Juan Sebastian Hidalgo, Beatrice Nale, Maike E. Debus, and Silvia Santini. 2020. A Multi-Sensor Approach to Automatically Recognize Breaks and Work Activities of Knowledge Workers in Academia. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4 (2020), 1 – 20.
- [88] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2018. Unobtrusive Assessment of Students' Emotional Engagement during Lectures Using Electrodermal Activity Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2 (2018), 1 – 21.
- [89] James Alexander Lee, Christos Efstratiou, and Lu Bai. 2016. OSN mood tracking: exploring the use of online social network activity as an indicator of mood changes. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (2016).
- [90] Peter M. Lewinsohn, John R. Seeley, Robert Edmund Roberts, and Nicholas B. Allen. 1997. Center for Epidemiologic Studies Depression Scale (CES-D) as a screening instrument for depression among community-residing older adults. *Psychology and aging* 12 2 (1997), 277–87.
- [91] Boning Li and Akane Sano. 2020. Extraction and Interpretation of Deep Autoencoder-based Temporal Features from Wearables for Forecasting Personalized Mood, Health, and Stress. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4 (2020), 1 – 26.
- [92] Chun-Tung Li, Jiannong Cao, and Tim M. H. Li. 2016. Eustress or distress: an empirical study of perceived stress in everyday college life. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (2016).
- [93] Mingnan Liu and Laura Wronski. 2018. Trap questions in online surveys: Results from three web survey experiments. *International Journal of Market Research* 60, 1 (2018), 32–49.
- [94] Jeffrey W Lockhart, Tony Pulickal, and Gary M Weiss. 2012. Applications of mobile activity recognition. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. 1054–1058.
- [95] Deborah Lupton. 2014. Self-tracking cultures: towards a sociology of personal informatics. In *Proceedings of the 26th Australian computer-human interaction conference on designing futures: The future of design*. 77–86.
- [96] Mitja Luvstrek and Bovstjan Kaluvza. 2009. Fall detection and activity recognition with machine learning. *Informatica* 33, 2 (2009).
- [97] David D Luxton, Russell A McCann, Nigel E Bush, Matthew C Mishkind, and Greg M Reger. 2011. mHealth for mental health: Integrating smartphone technology in behavioral healthcare. *Professional Psychology: Research and Practice* 42, 6 (2011), 505.
- [98] Neil Malhotra. 2008. Completion Time and Response Order Effects in Web Surveys. *Public Opinion Quarterly* 72, 5 (2008), 914–934.
- [99] João Marôco, Ana Lúcia Marôco, Juliana Alvares Duarte Bonini Campos, and Jennifer Fredricks. 2016. University student's engagement: development of the University Student Engagement Inventory (USEI). *Psicologia: Reflexão e Crítica* 29 (2016).
- [100] Richard P. Mattick and J.Christopher Clarke. 1998. Development and validation of measures of social phobia scrutiny fear and social interaction anxiety11Editor's note: This article was written before the development of some contemporary measures of social phobia, such as the Social Phobia and Anxiety Inventory (Turner et al., 1989). We have invited this article for publication because of the growing interest in the scales described therein. S.T. *Behaviour Research and Therapy* 36, 4 (1998), 455–470. [https://doi.org/10.1016/S0005-7967\(97\)10031-6](https://doi.org/10.1016/S0005-7967(97)10031-6)
- [101] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal* 5, 4 (2014), 1093–1113.
- [102] Lakmal Meegahapola, William Droz, Peter Kun, Amalia de Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, George Gaskell, Altangerel Chagnaa, Amarsanaa Ganbold, Tsolmon Zundui, Carlo Caprini, Daniele Miorandi, Alethia Hume, Jose Luis Zarza, Luca Cernuzzi, Ivano Bison, Marcelo Rodas Britez, Matteo Busso, Ronald Chenu-Abente, Can Günel, Fausto Giunchiglia, Laura Schelenz, and Daniel Gatica-Perez. 2023. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 176 (jan 2023), 32 pages. <https://doi.org/10.1145/3569483>
- [103] Daniel Miller and Heather A Horst. 2020. The digital and the human: A prospectus for digital anthropology. In *Digital anthropology*. Routledge, 3–35.
- [104] Shayan Mirjafari, Kizito Masaba, Ted Grover, Weichen Wang, Pino G. Audia, Andrew T. Campbell, N. Chawla, Vedant Das Swain, Munmun De Choudhury, Anind K. Dey, Sidney K. D'Mello, Ge Gao, Julie M. Gregg, Krithika Jagannath, Kaifeng Jiang, Suwen Lin, Qiang Liu, Gloria Mark, Gonzalo J. Martinez, Stephen M. Mattingly, Edward Moskal, Raghu Mulukutla, Subigya Nepal, Kari A. Nies, Manikanta D. Reddy, Pablo Robles-Granda, Koustuv Saha, Anusha Sirigiri, and Aaron D. Striegel. 2019. Differentiating Higher and Lower Job Performers in the Workplace Using Mobile Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3 (2019), 1 – 24.
- [105] Varun Mishra, Gunnar Pope, Sarah E. Lord, Stephanie Lewia, Byron M. Lowens, Kelly E. Caine, Sougata Sen, Ryan J. Halter, and David F. Kotz. 2018. The Case for a Commodity Hardware Solution for Stress Detection. *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (2018).

- [106] Óscar Martínez Mozos, Virginia Sandulescu, Sally Andrews, David Ellis, Nicola Bellotto, Radu Dobrescu, and J. M. Ferrández. 2017. Stress Detection Using Wearable Physiological and Sociometric Sensors. *International journal of neural systems* 27 2 (2017), 1650041.
- [107] Magen Mhaka Mutepfa and Roy Taper. 2019. *Traditional Survey and Questionnaire Platforms*. Springer Singapore, Singapore, 541–558. https://doi.org/10.1007/978-981-10-5251-4_89
- [108] Iftekhar Naim, M. Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. 2015. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 1. 1–6. <https://doi.org/10.1109/FG.2015.7163127>
- [109] Subigy Nepal, Shayan Mirjafari, Gonzalo J. Martínez, Pino G. Audia, Aaron D. Striegel, and Andrew T. Campbell. 2020. Detecting Job Promotion in Information Workers Using Mobile Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4 (2020), 1 – 28.
- [110] Renato Panda, Ricardo Manuel Malheiro, and Rui Pedro Paiva. 2020. Audio features for music emotion recognition: a survey. *IEEE Transactions on Affective Computing* (2020).
- [111] Delroy L Paulhus, Simine Vazire, et al. 2007. The self-report method. *Handbook of research methods in personality psychology* 1, 2007 (2007), 224–239.
- [112] Ashley L. Peterson. 2022. What is... psychological testing. <https://mentalhealthathome.org/2020/09/04/what-is-psychological-testing/>
- [113] John P Pollak, Phil Adams, and Geri Gay. 2011. PAM: a photographic affect meter for frequent, in situ measurement of affect. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 725–734.
- [114] Wanda Pratt, Madhu C Reddy, David W McDonald, Peter Tarczy-Hornoch, and John H Gennari. 2004. Incorporating ideas from computer-supported cooperative work. *Journal of biomedical informatics* 37, 2 (2004), 128–137.
- [115] Usama Rehman, Mohammad Ghazi Shah Nawaz, Neda Haseeb Khan, Korsi Dorene Kharshiing, Masrat Khursheed, Kaveri Gupta, Drishti Kashyap, and Ritika Uniyal. 2020. Depression, Anxiety and Stress Among Indians in Times of Covid-19 Lockdown. *Community Mental Health Journal* 57 (2020), 42 – 48.
- [116] João G. P. Rodrigues, Mariana Kaiseler, Ana Aguiar, João P. Silva Cunha, and João Barros. 2015. A Mobile Sensing Approach to Stress Detection and Memory Activation for Public Bus Drivers. *IEEE Transactions on Intelligent Transportation Systems* 16, 6 (2015), 3294–3303. <https://doi.org/10.1109/TITS.2015.2445314>
- [117] Jonathan Rubin, Hoda Eldardiry, Rui Abreu, Shane Ahern, Honglu Du, Ashish Pattekar, and Daniel G. Bobrow. 2015. Towards a Mobile and Wearable System for Predicting Panic Attacks. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (Osaka, Japan) (UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 529–533. <https://doi.org/10.1145/2750858.2805834>
- [118] Daniel Russell. 1996. UCLA Loneliness Scale (Version 3): Reliability, Validity, and Factor Structure. *Journal of personality assessment* 66 (03 1996), 20–40. https://doi.org/10.1207/s15327752jpa6601_2
- [119] Koustuv Saha, Larry Chan, Kaya de Barbaro, Gregory D. Abowd, and Munmun De Choudhury. 2017. Inferring Mood Instability on Social Media by Leveraging Ecological Momentary Assessments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1 (2017), 1 – 27.
- [120] Koustuv Saha, Ted Grover, Stephen Mattingly, Vedant Das Swain, Pranshu Gupta, Gonzalo Martinez, Pablo Robles-Granda, Gloria Mark, Aaron Striegel, and Munmun Choudhury. 2021. Person-Centered Predictions of Psychological Constructs with Social Media Contextualized by Multimodal Sensing. *Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies* 5 (03 2021), 32. <https://doi.org/10.1145/3448117>
- [121] Sirat Samyoun, Md Mofijul Islam, Tariq Iqbal, and John Stankovic. 2022. M3Sense: Affect-Agnostic Multitask Representation Learning Using Multimodal Wearable Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6 (07 2022), 1–32. <https://doi.org/10.1145/3534600>
- [122] Dr. Archana Sharma* and Dr. Vibhakar Mansotra. 2019. Multimodal Decision-level Group Sentiment Prediction of Students in Classrooms. . 4902–4909 pages. <https://doi.org/10.35940/ijitee.I3549.1081219>
- [123] M.Katherine Shear, Paola Rucci, Jenna Williams, Ellen Frank, Victoria Grochocinski, Joni Vander Bilt, Patricia Houck, and Tracey Wang. 2001. Reliability and validity of the Panic Disorder Severity Scale: replication and extension. *Journal of Psychiatric Research* 35, 5 (2001), 293–296. [https://doi.org/10.1016/S0022-3956\(01\)00028-0](https://doi.org/10.1016/S0022-3956(01)00028-0)
- [124] Gaurav Sinha, Rahul Shahi, and Mani Shankar. 2010. Human computer interaction. In *2010 3rd International Conference on Emerging Trends in Engineering and Technology*. IEEE, 1–4.
- [125] Michael Smith, Maryam Witte, Sarah Rocha, and Mathias Basner. 2019. Effectiveness of incentives and follow-up on increasing survey response rates and participation in field studies. *BMC Medical Research Methodology* 19 (12 2019), 230. <https://doi.org/10.1186/s12874-019-0868-8>
- [126] Robert L Solso, M Kimberly MacLin, and Otto H MacLin. 2005. *Cognitive psychology*. Pearson Education New Zealand.
- [127] Bettina Sonnenberg, Michaela Riediger, Cornelia Wrzus, and Gert G. Wagner. 2012. Measuring time use in surveys – Concordance of survey and experience sampling measures. *Social Science Research* 41, 5 (2012), 1037–1052. <https://doi.org/10.1016/j.ssresearch.2012.03.013>

- [128] Charles Donald Spielberger, Richard L. Gorsuch, and Robert E. Lushene. 1970. Manual for the State-Trait Anxiety Inventory.
- [129] Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of internal medicine* 166, 10 (2006), 1092–1097.
- [130] Th. Steimer. 2002. The biology of fear- and anxiety-related behaviors. *Dialogues in Clinical Neuroscience* 4 (2002), 231 – 249.
- [131] Arthur A Stone, Ronald C Kessler, and Jennifer A Haythomthwatte. 1991. Measuring daily events and experiences: Decisions for the researcher. *Journal of personality* 59, 3 (1991), 575–607.
- [132] Benjamin Tag, Zhanna Sarsenbayeva, Anna L Cox, Greg Wadley, Jorge Goncalves, and Vassilis Kostakos. 2022. Emotion trajectories in smartphone use: Towards recognizing emotion regulation in-the-wild. *International Journal of Human-Computer Studies* 166 (2022), 102872.
- [133] Yaniv Taigman, Ming Yang, Marc Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1701–1708.
- [134] Brandon T. Taylor, Anind K. Dey, Daniel P. Siewiorek, and Asim Smailagic. 2015. Using physiological sensors to detect levels of user frustration induced by system delays. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2015).
- [135] Nada Terzimehić, Sarah Aragon-Hahner, and Heinrich Hussmann. 2023. The Tale of a Complicated Relationship: Insights from Users' Love/Breakup Letters to Their Smartphones before and during the COVID-19 Pandemic. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 1 (2023), 1–34.
- [136] ML Tlachac, Ricardo Flores, Miranda Reisch, Katie Houskeeper, and Elke Rundensteiner. 2022. DepreST-CAT: Retrospective Smartphone Call and Text Logs Collected during the COVID-19 Pandemic to Screen for Mental Illnesses. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6 (07 2022), 1–32. <https://doi.org/10.1145/3534596>
- [137] ML Tlachac, Ricardo Flores, Miranda Reisch, Rimsha Kayastha, Nina Taurich, Veronica Melican, Connor Bruneau, Hunter Caouette, Joshua Lovering, Erial Toto, and Elke Rundensteiner. 2022. StudentSADD: Rapid Mobile Depression and Suicidal Ideation Screening of College Students during the Coronavirus Pandemic. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6 (07 2022), 1–32. <https://doi.org/10.1145/3534604>
- [138] C. Tong, Matthew J. Craner, Matthieu Vegreville, and Nicholas D. Lane. 2019. Tracking Fatigue and Health State in Multiple Sclerosis Patients Using Connected Wellness Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3 (2019), 1 – 19.
- [139] Timothy J. Trull and Ulrich W. Ebner-Priemer. 2009. Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: introduction to the special section. *Psychological assessment* 21 4 (2009), 457–62.
- [140] Niels Van Berkel, Jorge Goncalves, Simo Hosio, and Vassilis Kostakos. 2017. Gamification of mobile experience sampling improves data quality and quantity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–21.
- [141] Meredith L. Wallace, Molly H. Carter, and Satish Iyengar. 2018. The Current State of EMA and ESM Study Design in Mood Disorders Research: A Comprehensive Summary and Analysis.
- [142] Rui Wang, M. S. Hane Aung, Saeed Abdullah, Rachel M Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John M. Kane, Michael Merrill, Emily A. Scherer, Vincent Wen-Sheng Tseng, and Dror Ben-Zeev. 2016. CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2016).
- [143] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Seattle, Washington) (*UbiComp '14*). Association for Computing Machinery, New York, NY, USA, 3–14. <https://doi.org/10.1145/2632048.2632054>
- [144] Rui Wang, Gabriella M. Harari, Peilin Hao, Xia Zhou, and Andrew T. Campbell. 2015. SmartGPA: how smartphones can assess and predict academic performance of college students. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2015).
- [145] Rui Wang, Weichen Wang, M. S. Hane Aung, Dror Ben-Zeev, Rachel M Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John M. Kane, Emily A. Scherer, and Megan Walsh. 2017. Predicting Symptom Trajectories of Schizophrenia using Mobile Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1 (2017), 1 – 24.
- [146] Rui Wang, Weichen Wang, Alex DaSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. 2018. Tracking Depression Dynamics in College Students using Mobile Phone and Wearable Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–26.
- [147] Rui Wang, Weichen Wang, Alex W DaSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. 2018. Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2 (2018), 1 – 26.

- [148] Weichen Wang, Gabriella M. Harari, Rui Wang, Sandrine R. Müller, Shayan Mirjafari, Kizito Masaba, and Andrew T. Campbell. 2018. Sensing Behavioral Change over Time. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2 (2018), 1 – 21.
- [149] Weichen Wang, Subigya Nepal, Jeremy F. Huckins, Lessley Hernandez, Vlado Vojdanovski, Dante Mack, Jane Plomp, Arvind Pillai, Mikio Obuchi, Alex daSilva, Eilis Murphy, Elin Hedlund, Courtney Rogers, Meghan Meyer, and Andrew Campbell. 2022. First-Gen Lens: Assessing Mental Health of First-Generation Students across Their First Year at College Using Mobile Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 95 (jul 2022), 32 pages. <https://doi.org/10.1145/3543194>
- [150] Rick Wash, Emilee Rader, and Chris Fennell. 2017. Can People Self-Report Security Accurately? Agreement Between Self-Report and Behavioral Measures. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2228–2232.
- [151] Justin C. Wilson, Suku Nair, Sandro Scielzo, and Eric C. Larson. 2021. Objective Measures of Cognitive Load Using Deep Multi-Modal Learning: A Use-Case in Aviation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 40 (mar 2021), 35 pages. <https://doi.org/10.1145/3448111>
- [152] Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K. Villalba, Janine M. Dutcher, Michael J. Tumminia, Tim Althoff, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Jennifer Mankoff, and Anind K. Dey. 2019. Leveraging Routine Behavior and Contextually-Filtered Features for Depression Detection among College Students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3 (2019), 1 – 33.
- [153] Xuhai Xu, Prerna Chikersal, Janine M. Dutcher, Yasaman S. Sefidgar, Woosuk Seo, Michael J. Tumminia, Daniella K. Villalba, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Afsaneh Doryab, Paula S. Nurius, Eve Riskin, Anind K. Dey, and Jennifer Mankoff. 2021. Leveraging Collaborative-Filtering for Personalized Behavior Modeling: A Case Study of Depression Detection among College Students. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 41 (mar 2021), 27 pages. <https://doi.org/10.1145/3448107>
- [154] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff. 2023. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 190 (jan 2023), 34 pages. <https://doi.org/10.1145/3569485>
- [155] Han Yu and Akane Sano. 2023. Semi-Supervised Learning for Wearable-Based Momentary Stress Detection in the Wild. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 2, Article 80 (jun 2023), 23 pages. <https://doi.org/10.1145/3596246>
- [156] Shengchao Yu, Howard E Alper, Angela-Maithy Nguyen, Robert M. Brackbill, Lennon Turner, Deborah J. Walker, Carey B. Maslow, and Kimberly Caramanica Zweig. 2017. The effectiveness of a monetary incentive offer on survey response rates and response completeness in a longitudinal study. *BMC Medical Research Methodology* 17 (2017).
- [157] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. 2007. A survey of affect recognition methods: audio, visual and spontaneous expressions. In *Proceedings of the 9th international conference on Multimodal interfaces*. 126–133.
- [158] Xiao Zhang, Wenzhong Li, Xu Chen, and Sanglu Lu. 2018. Moodexplorer: Towards Compound Emotion Detection via Smartphone Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–30.
- [159] X. Zhang, Wenzhong Li, Xu Chen, and Sanglu Lu. 2018. MoodExplorer: Towards Compound Emotion Detection via Smartphone Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1 (2018), 176:1–176:30.
- [160] Yanxia Zhang, Jeffrey Olenick, Chu Hsiang Daisy Chang, Steve W. J. Kozlowski, and H. Hung. 2018. TeamSense: Assessing Personal Affect and Group Cohesion in Small Teams through Dyadic Interaction and Behavior Analysis with Wearable Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2 (2018), 150:1–150:22.