On Being Told How We Feel: How Algorithmic Sensor Feedback Influences Emotion Perception

VICTORIA HOLLIS, University of California, Santa Cruz, United States
ALON PEKUROVSKY, University of California, Santa Cruz, United States
EUNIKA WU, University of California, Santa Cruz, United States
STEVE WHITTAKER, University of California, Santa Cruz, United States

Algorithms and sensors are increasingly deployed for highly personal aspects of our everyday lives. Recent work suggests people have imperfect understanding of system outputs, often assuming sophisticated capabilities and deferring to feedback. We explore how people construe algorithmic interpretations of emotional data in personal informatics systems. A survey (n=188) showed strong interest in automatic stress and emotion tracking, but many respondents expected these systems to provide objective measurements for their emotional experiences. A second study examined how algorithmic sensor feedback influences emotional self-judgments, by comparing three system framings of physiological ElectroDermal Activity data (EDA): Positive ("alert and engaged"), Negative ("stressed"), and Control (no frame) in a mixed-methods study with 64 participants. Despite users reporting strategies to test system outputs, users still deferred to feedback and their perceived emotions were significantly influenced by feedback frames. Some users overrode personal judgments, believing the system had access to privileged information about their emotions. Based on these findings, we explore design implications for personal informatics including risks of users trusting systems that seemingly "unlock" hidden aspects of the self. We propose design approaches that provide opportunities for future emotion-monitoring systems to exploit these framing effects, and for users to more actively construe emotional states.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**;

Additional Key Words and Phrases: Algorithms; Deference; Trust; Personal Informatics; Health; Emotion Regulation; Emotion; Affective Computing; Feedback; Framing; Anxiety; Stress

ACM Reference Format:
Victoria Hollis, Alon Pekurovsky, Eunika Wu, and Steve Whittaker. 2018. On Being Told How We Feel: How Algorithmic Sensor Feedback Influences Emotion Perception. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 114 (September 2018), 31 pages. https://doi.org/10.1145/3264924

1 INTRODUCTION

Algorithms and sensors are now widely deployed in many aspects of our lives and recent research has begun to examine how people interpret and trust algorithmic output [16, 22, 26, 46, 70, 74, 77]. That work shows that people often have quite rudimentary understandings of algorithmic operation. Furthermore, efforts to make algorithms more transparent do not necessarily improve user understanding [26, 46, 70]. Research on algorithms

Authors' addresses: Victoria Hollis, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA, 95064, United States, hollis@ucsc.edu; Alon Pekurovsky, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA, 95064, United States, apekurov@ucsc.edu; Eunika Wu, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA, 95064, United States, ecwu@ucsc.edu; Steve Whittaker, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA, 95064, United States, swhittak@ucsc.edu.

2474-9567/2018/9-ART114 $15.00
https://doi.org/10.1145/3264924

that interpret highly personal data notes a further phenomenon: overriding one's own interpretation and instead trusting system outputs even when these conflict with users' own interpretations [74]. Such deference occurs for highly personal aspects of self-knowledge such as personality. Warshaw et al. [74] observed user reactions to an algorithm that derived 'personality profiles' from social media posts. Some users described the profiles as 'creepily accurate', while others were resistant to correcting algorithm outputs even when its characterization conflicted with their own expectations. We expand this work by characterizing whether and how users are influenced by algorithmic interpretations of their own real-time emotional states in the context of a challenging cognitive task. Each of us has privileged access to our own thoughts and affect, making it seem intuitively unlikely that users would believe a system would have greater insight than themselves into how they think or feel.

It is important that we understanding possible user deference when evaluating algorithms, as our everyday decisions are increasingly informed by inferential systems [65, 74]. We use algorithms daily for information seeking, purchasing decisions, navigation, and so forth. More recently, we are also seeing user-facing algorithms in systems to analyze personal health and well-being, e.g. recommending tailored exercises, diet choices or even providing medical diagnoses [58, 71]. New personal informatics work has begun to explore attitudes to complex personal data about physical health (e.g., [9, 40, 77]), but less attention has been paid to systems for emotion regulation. However, this is a key area: mental health is prevalent societal problem [42] and the economic burden of stress is estimated to be $300 billion dollars annually for the U.S. alone [61].

Recent qualitative studies of emotion sensing systems also indicate that users lack insight into algorithmic capabilities and are over-accepting of system interpretations of personal affective data. Snyder et al. [68] designed an ambient mood awareness system, observing that some participants deferred to system interpretations, where they: "*abdicated the role of self awareness to the system, seeming to expect (and trust) the [system] to tell them how they were feeling*". Similarly, users can be over-trusting of a system's emotional analyses of past personal experiences, sometimes judging random system interpretations to be accurate [70]. While prior work has noted individual examples of participants deferring to machine-based interpretations of affect, we directly test these effects, identifying direct consequences of algorithmic interpretations on people's emotion perception. Furthermore, we assess how the perceived valence of an experience is affected by different outputs of the same algorithm.

Understanding exactly how algorithms influence emotion perception is important as such systems become widely deployed (e.g., Apple Watch applications or the Samsung S Health application). Determining these user perceptions is a critical issue as commercial and research deployments of 'emotion sensing' algorithms vary widely in the interpretations they assign to highly similar emotion-relevant data [30, 31, 44, 48, 63]. For example similar physiological skin conductance ElectroDermal Activity (EDA) data is interpreted by some commercial systems as a positive indicator of mood (e.g., "excited") [23, 64], and in others as negative ("stressed") [53]. Related effects are demonstrated in text processing of sentiment, where various algorithms can reach different interpretations when processing the same corpus of affective data [59, 78]. Some research systems take a less definitive approach, instead encouraging users to actively appraise system feedback [10, 50, 68]. However, many commercial systems frame their feedback as precise and accurate. This makes it important to document whether and how people defer to an algorithm that seems to definitively interpret their emotional states.

This paper presents two studies. First, a survey to explore people's interests in, and expectations for, systems that track stress and emotions. Second, an in-lab, mixed methods study to examine how people understand and evaluate different algorithmic feedback about their personal emotional data, exploring potential design implications. We examine how people interpret such emotional data, directly comparing positive and negative

system feedback framings of EDA data, while conducting a moderately demanding cognitive task. We address the following research questions:

- RQ1: Do people want systems that automatically infer stress and emotion? What prior knowledge do they have of such systems, why do they want to adopt such tools, and what are their expectations for these systems? (Study 1)
- RQ2: How are participants' perceptions of their emotions influenced by seeing different outputs of the same underlying algorithm? Do they defer to system feedback? (Study 2)
- RQ3: What types of strategies do participants have for interpreting outputs of emotion-sensing systems? Do participants consider system feedback to be accurate, and do perceptions of accuracy differ depending on how the algorithm is described? (Study 2)

These are critical questions as complex systems are increasingly involved in many personal aspects of our lives and assessments of our well-being.

2 CONTRIBUTION

In Study 1, survey results show that young adults are highly interested in automatic stress and emotion tracking. Nevertheless, they have little awareness about existing tools. Furthermore, their motivations to use such systems suggest that they perceive systems as potentially offering objective, accurate interpretations of personal experiences.

Study 2 directly examines how users interpret system outputs of an emotion sensing device, and the consequences on their perceived emotions and stress. While users describe strategies to actively test system inferences about their real-time emotions, they nevertheless defer to system framings and are significantly influenced by algorithm feedback. Participants with negative feedback report increased symptoms of anxiety and are more negative in their usage experience. In contrast, we see some protective effects of positive feedback which reduces anxiety and improves task assessments. While prior work shows that perceived emotional intensity is influenced by bodily awareness [15], we instead show that different interpretations of the same data can have important consequences for perceived emotional *valence*. We additionally present qualitative evidence of user deference to system interpretations of their emotional states, moments of apparent insight, and concerns for these systems to have counterproductive effects.

Based on these findings, we present design implications for emotion sensing and Personal Informatics (PI) systems in general, exploring how, when, and why we might present self-relevant personal data. User deference means that new designs should exercise caution in assigning definitive meanings to emotion-sensing algorithms. We also describe new design approaches that: (1) reduce deferral effects by promoting more accurate mental models and collaborative interpretations of data, (2) strategically leverage beneficial effects to aid emotion management during challenging tasks.

3 RELATED WORK

3.1 Interpreting Complex System Outputs

There is an extensive HCI literature showing that users experience difficulty developing conceptual models of how complex systems operate, and how those systems' outputs should be interpreted [4]. For example, Norman [54] characterizes users' problems in mapping between their own actions and resulting system outputs as the 'gulf of evaluation'. More recent research on smart sensors and context-sensitive computing also documents users' difficulties in forming accurate models of how those systems function [7, 8]. There is also a large emerging

literature demonstrating that people have limited understandings of algorithmic systems and their outputs [16, 22, 26, 46, 70, 73, 74, 77]. For example [22] showed that some participants were unaware of the fact that Facebook filtered their feed, explaining away filtered posts as a lack of attention. One proposed remediation strategy is to provide explanations into how algorithms operate [4], although the benefits of these explanations are highly context-sensitive [43, 47], and do not guarantee increased user understanding.

3.2 Interpreting Personal Physical Health Data

The above research describes users' difficulties in deriving conceptual models of complex system operation. However, these problems may manifest differently in the context of personal health data, where users should have direct access to alternative information sources, such as self-observation, to test system outputs.

This has been observed in [77] where users of fitness trackers actively tested algorithm accuracy through a variety of methods. For example, users directly counted their steps to serve as ground truth for a pedometer, compared commercial heart rate monitors with medical devices or carried out ad hoc tests to determine which of their actions triggered system outputs. Although users find it difficult to form an accurate conceptual model, access to such external physical data allows them to critically evaluate sensor accuracy, in some cases discounting a system interpretation. In other cases, personal expectations lead users to challenge system accuracy [28, 40]. These studies again show that users experience difficulties in forming accurate mental models of complex systems. Nevertheless, users subject outputs to close scrutiny and testing in order to validate accuracy, sometimes overriding system interpretations in favor of personal judgments [40].

3.3 Tracking and Reflecting on Emotional States

Many people encounter difficulties in understanding and managing their emotions [27], motivating the design of new systems to improve emotion monitoring and regulation [15, 32, 48]. Three very different design approaches have been taken: manual, automatic and hybrid.

Manual systems require users to actively monitor and record affective experiences. For example, commercial systems such as Moodscope [18], prompt users to post daily mood descriptions generating 'mood graphs' that map affect to daily experiences. Some research systems take the same approach, including a reflective component to later re-evaluate reactions to those experiences [2, 32, 33]. In these manual systems, users directly appraise their emotions, making it less likely that they will defer to system interpretations. However, one disadvantage to such manual systems is that users may be resistant to record particular moods, for example if these are negative [32].

In contrast automatic systems use biometric measurement [48, 50], or linguistic content [11] to infer user affect. Many commercial systems [23, 53] and research systems are described as automatically tracking emotional states [13, 48, 50, 63]. Various studies allude to deferral effects for these automatic systems. For example, Ruckenstein [62] assessed user reactions to everyday self-monitoring of heartrate and heart rate variability data. Participants described system outputs as more 'factual' or 'credible' than their personal intuitions. Even entirely random feedback on emotions is trusted, with users generating folk theories to confirm system interpretations of their affect [70]. Other work [15] has examined the effects of false physiological feedback on anxiety. In a mock job interview, participants receiving inaccurately low heart rate feedback reported fewer anxiety symptoms compared with those receiving accurate feedback. While these findings highlight the protective effects of false feedback, they draw attention to the potential for emotion feedback systems to critically influence highly personal emotional experiences.

A third hybrid design approach combines manual and automatic interpretation. Hybrid systems take algorithmic inputs from sensor data (heartrate, skin response) but render these data in abstract ways to encourage users to actively appraise their emotions. For example, AffectAura [50] and Moodlight [68] present emotions as abstract visualizations. Despite these designs being intended to promote active interpretation, [68] report cases where users override personal interpretations of their own affect in favor of system outputs, amplifying negative emotional experiences.

In summary and in contrast to physical data, these studies of personal emotional data suggest that users tend to be less skeptical of system outputs, downgrading personal insights and deferring to system interpretations. From these examples, it's clear that system feedback can influence emotional states and that users may trust affect detection algorithms to be accurate. This leaves open important questions: first, does the way that systems frame feedback influence users' perceived emotion? And how do users derive meaning, test system accuracy, and form judgments about their own emotional states when using these systems? Although prior work has noted individual examples of participants deferring to machine-based interpretations of affect, we experimentally evaluate these effects, comparing positive and negative emotion feedback frames for the same algorithmic output. Understanding these effects is particularly important given that different commercial systems present contradictory interpretations of very similar EDA data, presenting this as indicating either positive [23, 64] or negative stress [53].

3.4 Theoretical Models of Emotion

Finally, we review theories of emotion and their relevance to system feedback. Classical theories of emotions argue for a set of basic, discrete emotional states ('fear', 'anger', 'surprise', 'happiness', 'disgust', 'sadness') that can be detected through physiological fingerprints such as heart rate, electrodermal activity (EDA), facial expression and so forth [19, 34]. More relevant for systems research, recent cognitive models describe how emotions are perceived and experienced, highlighting the role of active interpretive processes (see also [10]).

Appraisal Theories [20, 67] argue against a direct mapping between an objective experience and emotional response. Instead they emphasize the active role of the experiencer in interpreting that emotion. For example, emotional appraisals depend on context; e.g. whether an event is expected, goal relevant, goal hindering, and so forth [20]. Mendes et al. [51, 52] proposed that emotional appraisals are affected by one's perceived resources (e.g., power) against expected demands (e.g., outcome probability). Emotional demands and resource appraisals can be experimentally manipulated with direct effects on task performance and physiological markers of stress [36, 37].

Appraisal theories are directly relevant to framing effects with algorithmic 'emotion sensing' systems. For example, priming participants that stress is beneficial (i.e., eustress) promotes a challenge response, resulting in higher math scores [36] and fewer cardiovascular stress markers [37]. Eustress framings also increase participants' belief in their ability to meet task demands [52]. In contrast, negative stress framings promote opposite effects. However, these studies only examined altering expectations for a single feeling (stress) and did not evaluate framing effects in the context of deployed emotion detection systems, which we examine here.

Social Constructionist Models: While Appraisal Theories address goal-related aspects of emotions, other cognitive models examine the role of situational context on the emotion construction process. Social Constructionist Models [5] argue that rather than directly experiencing discrete emotions, individuals "*use knowledge to parse and conceptualize the bottom-up information that is sensorially given*". People use situational context and linguistic labeling to transform ambiguous sensations of affect (e.g., feeling 'bad'), into concrete emotional experiences

(e.g., feeling 'righteous indignation'). This model is relevant to emotion-focused PI systems, which can contribute contextual information and linguistic labels consequently impacting emotion perception. For example, whether a system is described as measuring 'stress' or 'alertness' should influence how that system is construed, with consequences for deployment and subsequent use.

In summary, by emphasizing the active role of the experiencer, cognitive models of emotion predict that system context may influence situational appraisals and emotion perception. We now turn to Study 1, a foundational survey to identify needs and expectations for adopting emotion monitoring systems.

4 STUDY 1: SURVEY OF STRESS TRACKING EXPECTATIONS AND NEEDS

4.1 Participants:

225 students at a large U.S. university completed a survey examining their *knowledge* of Personal Informatics (PI) systems (i.e. systems that collect personally relevant information) [45], *interests* in adopting a tool for automatic collection of and analysis of personal data, and *expectations* for using such a system. Two attention check questions were used to filter responses. From this filter, 37 responses were excluded with a final sample of 188 respondents (138 Female, 48 Male, 2 Genderqueer). Respondents were ages 18 to 31, (Mean: 19.81, SD: 1.66). Most (77.65%) had used a tool to track mood or health. The prevailing majority had used physical health trackers (e.g., Fitbit, Google Fit). Surveys were administered using Qualtrics. Young adults were the focus of this survey as they are known to have high particularly high levels of stress [1].

4.2 Procedure and Survey Design:

The survey first evaluated participants' knowledge of PI systems, assessing the range of existing commercial systems that they were aware of. Next, respondents identified types of data they would want collected about themselves automatically, with the remaining survey tailored to those types of data. We also probed their intuitions about the utility of the data such systems provide.

Knowledge of PI Systems: To probe for existing knowledge of PI systems, respondents could multi-select from a list which commercial PI systems they had heard of. The list included 19 different systems with options spanning popular automatic and manual fitness trackers (e.g., Fitbit, Jawbone, MyFitnessPal, Runkeeper), among several other specialized trackers (e.g., Mint, Rescuetime, Muse). In addition, the list included options for commercial manual and automatic emotion monitoring systems (e.g., Moodpanda, Spire, Pip). Respondents could also select an '*Other*' option to enter the names of any other self-tracking systems they were aware of or choose '*None of the above*' if they were completely unfamiliar with self-tracking technologies.

Self-Relevant Data Types: Respondents could select up to 3 aspects of themselves that they would want a tool to track automatically. These options spanned 14 possible categories, including health (e.g., calories burned, time spent sitting, dehydration), work (e.g., productivity, focus time), and affective categories (e.g., stress, mood). These options were chosen to reflect a wide range of possible personal informatics systems [14]. Respondents also indicated their strength of interest for each chosen personally-relevant data type using a 5-point scale ranging from 'Not at all Interested' (1) to 'Extremely Interested' (5).

Justifications and Expectations: They were next asked justify their choice of tracked data for each of the personal data types respondents they had selected:

- *Motivations:* An open-ended description of why they wanted automatic tracking of the selected personal data, followed by a primary, closed-class motivation. These 10 primary motivations were based on [60],

including understanding one's self, general improvement, achieving a specific goal, documenting one's life, exploring a new technology, and so forth.

- *Situational Contexts*: A multi-select list for 13 different contexts in which the respondent wants to have automatic stress or emotion tracking. These contexts included social factors (e.g., around family, friends, or alone) and specific activities (e.g., job interview, exam, etc.).
- *Expectations* : A rank ordered list of attributes respondents want represented in a system that automatically tracks aspects of one's self (e.g., accuracy, goal-setting, hiding information, and so forth).

Finally, respondents completed personality questionnaires and supplied demographics.

4.3 Results

We mainly focus on survey results for stress and mood tracking systems as they are directly relevant to Study 2. Respondents selected automatic mood and stress tracking as their top two interests for a Personal Informatics tool, yet had little knowledge of existing systems, expecting these technologies to be more objective than their own organic self-awareness. We focus here on quantitative and qualitative responses of those selecting automatic stress or mood tracking (n=130).

4.3.1 Automatic Stress and Emotion Tracking is a Top Interest. Stress is the top factor that respondents want a Personal Informatics tool to automatically track. We also included an option for automatic mood tracking and find this is the second most selected option for these respondents.

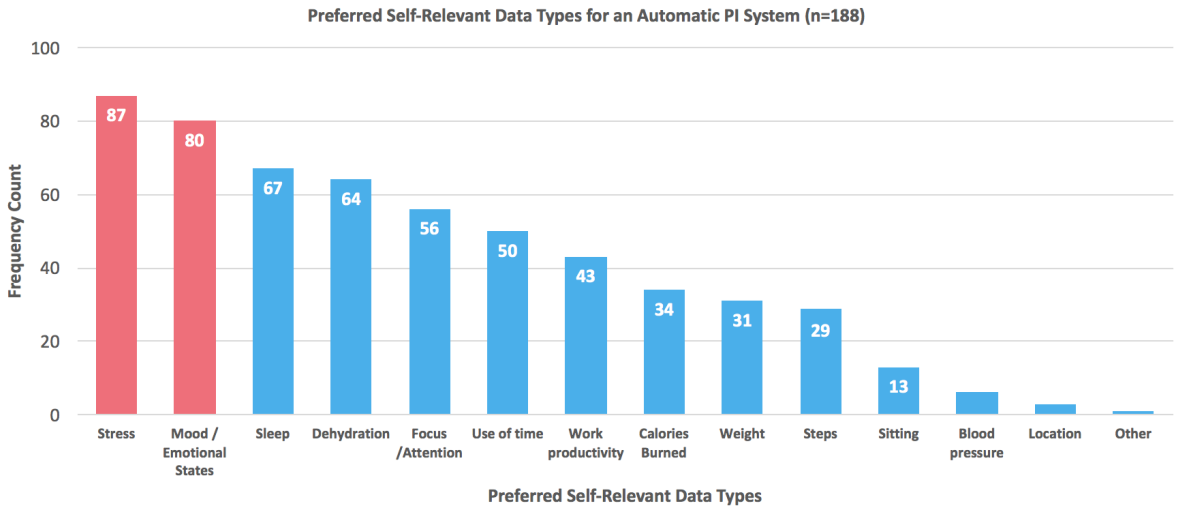

Fig. 1. Survey results from (n=188) respondents on the self-relevant data types respondents want automatically collected by a personal informatics system. Each respondent could select up to 3 choices. The top interests for automatic tracking were for stress and emotional states.

Those who wanted stress tracking were strongly interested in adopting such a tool with an average rating of 4.41 (SD: 0.75), similar to those who selected emotion tracking (Mean: 4.54, SD: 0.64). However, other questions revealed that only a single respondent could provide concrete examples of automatic stress or emotion tracking

systems. This finding is important as it shows that these young adults have a deep interest in automatic emotion monitoring systems, yet little knowledge of existing research or commercial systems from which they might base accurate mental models.

4.3.2 Expectations and Needs for Stress Tracking: Systems for Groundtruth and to Regulate Self. The data concerning *why* respondents wanted automatic stress indicated that many have an expectation these tools will serve as a groundtruth for their stress experiences, raise awareness, and help them regulate stress.

Tracking to Improve Self-Understanding: Improved self-understanding was selected as the top primary motivation for automatic stress (44.8%) and mood (61.3%) tracking. This was highly distinct from other data categories, such as calorie tracking or productivity, where self-understanding was a rare motivation (2.9% and 9.3% respectively). Respondents written descriptions were largely related to how such system feedback can serve as an objective measure for one's experiences, e.g. revealing: "*...whether I am* ***actually*** *stressed*" (S25).

This notion of objectivity was coupled with a belief that such a system would supersede one's awareness and organic monitoring abilities ("*...****We never really know*** *how stressed we are...*", S47; "*Sometimes* ***i cannot tell even my own emotions****...*", S90) or to compare device outputs to one's own subjective evaluations ("*...to compare results from the device to how I perceive my state my state of mind...*", S22). Participants also noted the consciousness-raising benefits of having access to such data: "*...****if I can physically see my stress levels, it will help me calm down*** *or take a step back...*" (S46).

Some even hypothesized that system outputs served as a reliable external standard (e.g "*...to have my beliefs* ***proved****...*", S60). System outputs were also thought to offer trustworthy mental health records that could be used when justifying behavior to others:

> "*...mental health days are a disputed issue in schools at the moment...During mental health crises we can't take a day off...****this would be something verifiable*** *and easy to track by schools and provide justification for it.*" (S36)

Tracking to Better Manage Stress and Its Repercussions on Other Aspects of Life: The second most selected motivation for tracking stress was for general improving one's life (20.7%), while this was much less often selected for automatic emotion tracking (6.3%). Here participants noted the effects of stress on other life goals, and the need to improve mental health ("*...I want to overcome my mental illness*", S27). As S69 describes, "*I know myself to have anxiety and panic attacks so* ***knowing what my stress levels are at and how those correspond with my panic attacks*** *could be useful for me*". Automatic stress or emotion trackers were also seen to potentially benefit other aspects of life such as interpersonal relations ("*...stress ends up negatively impacting a relationship.*", S59), productivity ("*...when lack of productivity is attributed to stress...*", S85), and general health ("*...better one's sleep, eating, and overall self-esteem...*" S29).

Finally, we probed respondents for *situational contexts* in which they would want automatic stress or emotion tracking. For these students, the top-selected option for a specific context was during an exam (59.8%). For automatic emotion tracking, the second most selected specific context was more broadly for when they were at school (50%). These choices were supported by free-writes with respondents noting they want automatic stress tracking because they "*...have a lot of test anxiety...*" (S9), and are "*...easily stressed out regarding academic things...*" (S79).

4.4 Conclusion

Study 1 showed that stress and emotion tracking is a top interest among students, who often selected test taking as a specific context in which they want such stress tracking. Despite this interest, we found limits in their knowledge of existing systems, as well as different motives for adopting tools for automatic stress data collection. To our surprise, we found little skepticism about the accuracy of such tools with people largely voicing expectations that such systems could serve as a ground truth for their experiences. However, these results are limited given respondents were not specifically probed about potential skepticism or concerns they have for such systems. We nonetheless see that respondents largely assumed such tools would be objective measures of emotion and had limited knowledge of existing systems to base expectations from.

Such trust in technology is potentially problematic in the context of deployed stress tracking systems where different products offer widely different interpretations of very similar data; as we have seen EDA physiological measurements are framed as signaling positive states such as 'engagement' in some products [64] and 'stress' in others [53]. Furthermore, there is an apparent risk: that many want to become aware of stress in order to better manage it, yet labeling one's experience as 'stressful' could itself prove to be counterproductive [48].

5 STUDY 2: IMPACT OF AUTOMATIC STRESS TRACKING FRAMES ON STRESS EXPERIENCES

We pursued these issues in Study 2 which examines how changes in system descriptions influence the processes of interpreting one's emotional experiences. We wanted to observe whether the beliefs identified in our survey about the advanced capabilities of stress and emotion detection systems were replicated for actual system experiences. Would users continue to trust in the 'objective standards' of such systems when they had a direct experience to compare against? We therefore built and deployed a system in the lab to test the effects of differing algorithm frames while participants took a challenging test. We compare reactions to an 'emotion sensing' system, allowing us to assess user experiences for both negative ('stressed') and positive ('alert and engaged') framings. Would users defer to system interpretations of their emotions and stress or would they be more attentive to their own feelings and experiences?

Study 2 compliments findings of Study 1 by addressing the following research questions:

- How are participants' perceptions of their emotions influenced by different descriptions of the same underlying algorithm? (RQ2)
- What types of strategies do participants have for interpreting system emotion assessments? To what extent do participants consider these outputs accurate, and do perceptions of accuracy differ based on how the algorithm is described? (RQ3)

5.1 The EmVibe System

EmVibe (See Figure 2) is a wearable system that combines an electrodermal activity (EDA) sensor to measure skin conductance, a smart watch for haptic feedback, and a custom Android application for communicating between the sensor and watch. We chose EDA, as this physiological measurement has been framed in commercial and research contexts as assessing both engagement [30, 31, 44, 64] and stress [17, 48, 53].

We acknowledge that definitions of 'emotion' and 'stress' are a controversial topic [6]. We consider 'alert and engaged' as appropriate categories for an emotional state given their presence in influential scales of emotion self-reports (e.g., [24, 75]). Furthermore, we are presenting an alternative, positive frame for stress that would be applicable to a test-taking context, making other descriptions (e.g., 'excitement') less appropriate. For the

purposes of understanding system feedback in a testing context, we considered the psychological components of stress as emotional (e.g., 'nervous', 'anxious') as determined by our use of subjective report scales (e.g. STAI), rather than physical measures of stress (e.g., blood pressure). Nonetheless, there are physiological components of stress which may not correspond to the same psychological experiences, and vice versa [20].

Fig. 2. EmVibe System. A Pip sensor [25] is held between the thumb and forefinger to measure electrodermal activity (EDA) i.e. skin conductance. A smart watch on the wrist vibrates to give haptic feedback when the sensor detects increased EDA. We tested the impact of feedback on emotion perception by changing the system description. Participants were told that feedback indicated either (1) increased 'alertness and engagement', (2) increased 'stress', or (3) were given no explanation for feedback.

5.1.1 Measuring Electrodermal Activity. EDA assesses sympathetic nervous system activity by measuring the moisture level of skin (i.e., skin conductance) [12]. In our study, EDA was measured using the commercial Pip sensor [25], which has been successfully deployed multiple times in related research [17, 49, 68] and is advertised as a stress detector. The Pip is a handheld device consisting of two gold-plated sensors, pinched between thumb and forefinger (see Figure 2).

The Pip samples EDA at 8Hz, applying a proprietary algorithm to classify EDA [25]. The Pip signal processing software bases classifications on the slope of the EDA curve, computing increases and decreases in EDA on successive windows of data. Pip gives greater weight to detected increases as these are more closely linked to sympathetic nervous system activity, while decreases may indicate thermoregulation [25]. Output from the Pip is transmitted wirelessly over Bluetooth to an Android smartphone application which then communicates EDA classifications to the smart watch. The Android application stores a logfile for each participant session including time-stamped EDA events and raw EDA data. PIP classified events as 'stressed', 'relaxed', or 'constant' depending on whether it detects increases, decreases or no changes in EDA [25]. For clarity in describing Study 2, and given the different conditions, we instead refer to these EDA events as 'increased', 'decreased', or 'constant'.

5.1.2 Presenting Haptic Feedback. Our goal was to understand participant reactions to algorithmic feedback about their emotional states. However, some of our participants were going to receive negative feedback. Prior work showed that some people respond to such feedback by attempting to avoid it [48, 68]. We therefore chose haptic vibration rather than visual feedback to reduce avoidance. When the Pip detected increased EDA, we provided feedback in the form of wrist-directed vibrations using a Motorola Moto 360 Sport smart watch. The watch display was covered so that it wouldn't serve as a visual distraction.

6 METHODS: SYSTEM INTERVENTION FOR FRAMING EMOTION ANALYTICS

Prior work shows that participants' often experience problems in interpreting algorithmic outputs, and Study 1 revealed that participants viewed emotion and mood tracking systems as likely having quite advanced interpretive capabilities. The goal of Study 2 was therefore to understand participants' reactions to different forms of algorithmic feedback about their emotional states. We wanted to know how people would interpret three different versions of the system in which outputs were framed as being either Positive 'alert and engaged', Negative 'stressed' or Neutral (no frames). We evaluated the effects of this different system feedback for exactly the same underlying Emvibe system. Participants provided qualitative and quantitative self-reported measures of affect after conducting a moderately demanding problem-solving test while receiving feedback from the EmVibe system.

6.1 Experimental System Framings

Participants were randomly allocated to one of the following conditions:

- *Negative Framing*: Received priming information on the benefits of managing one's stress. When describing the EmVibe system, we told them that the EDA vibration feedback signals "Increased activity of the sympathetic nervous system which indicates stress".
- *Positive Framing*: Received priming information on the benefits of managing one's alertness and engagement. When describing EmVibe, participants read that that the (identical) EDA vibration feedback signals "Increased activity of the sympathetic nervous system which indicates greater alertness and engagement".
- *Control*: Controls used the EmVibe and received vibration feedback for increases in EDA. However, they were given no explanation of what the system was doing, or what the (again identical) vibration feedback represented.

Participants were not told the 'EmVibe' specific system name, as we did not want this name to bias potential interpretations. However, they were told about the Pip sensor in terms of its physical construction and that it is used to measure EDA. Exact descriptions and system feedback frames are in the Appendix. These were arrived at after extensive piloting.

6.2 Timed Problem-Solving Test

The problem-solving test involved participants completing 36 questions from an abstract reasoning test, Raven's progressive matrices. Questions were from the latter half of the Raven's test (C, D, E) as these increased from moderate to relatively high difficulty. The test was administered through a laptop computer and Raven's questions were presented individually on separate web pages.

To increase challenge, a timer was displayed at the top of each question page, showing a 25 second countdown per question. If the timer reached 0 seconds, then the test website automatically transitioned to the next question. To complete a question, the participant verbalized a final answer to the researcher administering the test and the researcher gave immediate, truthful feedback on whether that answer was correct.

6.3 Instructions and Measures

6.3.1 Pre-test Materials. A demographics survey captured age, self-identified gender, university GPA, and English fluency. Participants then answered pre-test surveys to assess baseline levels of anxiety (both situational and dispositional), motivation to perform well, and appraisals of the test. Expected appraisals were collected as these are potentially influenced by primes alone [36, 37, 51, 52]. A measure of appraisals could also potentially inform explanations of mechanism, and to what extent appraisal theory could account for experimental results.

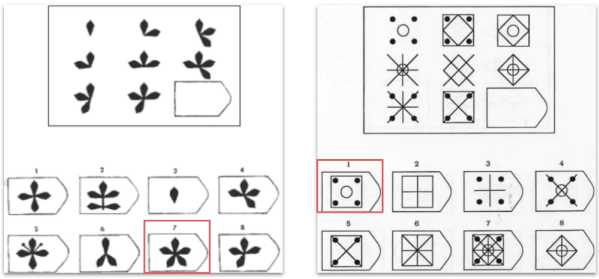

Fig. 3. Timed Exam. Participants completed three sections of a pattern recognition test (Raven's Progressive Matrices). Participants were instructed to select the answer that best completes the pattern. The leftmost image is a lower difficulty question with an answer of 7. The rightmost image is a more difficult question with an answer of 1.

State-Trait Anxiety Inventory (STAI) Survey: The STAI for Adults short-form [69] is a 20-item scale to assess mental and physical symptoms of anxiety. Responses are given on a 4-point scale with higher scores indicating greater anxiety. The State subscale (10-items) assesses the degree to which one is experiencing current symptoms of anxiety (e.g., "*I am jittery*"). The Trait subscale (10-items) assesses the degree to which one experiences anxiety symptoms in general life (e.g., "*I feel nervous and restless*"). The STAI has good internal reliability with Cronbach's alpha coefficients of 0.86-0.95. It has also been used in prior work to measure effects of false feedback on anxiety [15].

After reading the test instructions and respective framing, participants completed the following additional measures.

Motivation: We asked participants to estimate their motivation ("*How motivated are you to perform well on this test?*") with responses on a 7-point scale from 1 (Not at all motivated) to 7 (Extremely motivated). Participants may have differed in the effort they applied to the test and we wanted to control for this.

Demand and Resource Appraisals: Following prior work [51, 52], we also gathered participants' cognitive appraisals to assess their expected demands and perceived resources to cope. Six questions related to demand appraisals (e.g., this task is "*demanding*", "*distressing*", etc.) and five questions related to resource appraisals (e.g., I have the "*abilities to perform well*", "*this task is a positive challenge*", etc.). Responses are given on a 7-point scale, 1=strongly disagree to 7=strongly agree. This questionnaire is scored as a ratio of mean perceived demands to mean perceived resources. Higher scores indicate beliefs of greater perceived demands than resources.

6.3.2 Post-test Materials. After completing the problem-solving test, we re-administered the STAI-state in a post-test to determine whether framing changed currently experienced anxiety. We were also interested in participants' experiences with the system and assessments, so we included questions assessing perceived EmVibe accuracy, recall of the test-taking experience, and willingness to use EmVibe outside of a research setting.

Task Stress Rating: Participants rated on a 7-point scale (1=strongly disagree to 7=strongly agree) the extent to which they thought "*The test was stressful*".

Number of Perceived Vibrations: Outcomes could be affected by variability in perceived levels of feedback. To address this potential confound, participants answered "*Approximately how many times did you notice the band*

vibrate?".

Accuracy of Feedback: Negative Framing and Positive Framing participants answered a single item question to rate sensor feedback accuracy: "*The EDA sensor accurately detected when I was stressed*" (Negative) or "*The EDA sensor accurately detected when I was alert and engaged*" (Positive). Responses were given on a 7-point scale, 1=strongly agree to 7=strong disagree. These participants then generated a free-write explanation for how they determined system accuracy.

Test-Taking Experience Free-Write: Participants answered the following open-ended question: "*In general, please describe your experience taking the test(s) and using this technology. Was there anything you liked or disliked? Anything that surprised you or seemed fairly typical?*". This allowed us to assess qualitative differences in how participants affectively appraised the test-taking experience based upon which framing they were exposed to.

Willingness to Use: Finally, participants rated their agreement to the prompt "*I want to use a device like this to track emotion in my daily life:*" with responses given on a 7-point scale. After which, they provided a free write description of their answer. At this point, control participants were told in the survey that the sensor was simply tracking their emotions.

The final question of the post-test survey was a suspicion check question, asking participants what they thought the study was about. Finally, participants were thanked and presented with a written debrief of the study.

6.4 Participants

74 participants from a large U.S. university volunteered to join the study. 8 were removed because probes showed they had not read the study instructions, and a further 2 because they indicated in a post-test survey they had recently consumed caffeine [15]. The final sample consisted of 64 participants (16 self-identified as male, 48 as female) with an age range of 18-22 years (M: 19.5, SD: 1.4). Participants received course credit. This study was previously approved by an Institutional Review Board.

6.5 Procedure

Participants first submitted a consent form and demographics questionnaire. This was followed by a 5-minute rest period, during which a baseline EDA reading was taken using the Pip sensor. Participants then completed the pre-test anxiety scale.

Participants were randomly allocated to one of three Framings and presented with either the Negative Framing (n=24), Positive Framing (n=20), or Control (n=20). We did not anticipate conflicts between frames and participants' prior expectations as our survey participants showed little knowledge about actual emotion and stress tracking systems. After reading the relevant framing materials, each participant was given instructions for the timed test. This additionally involved multiple-choice questions to check they had read the framing and study instructions (see Appendix), followed by ratings of pre-test motivation and appraisals. We then set the participant up with EmVibe and proceeded with the timed test. After which, the participant removed EmVibe, submitted the post-test questionnaire, was thanked, and debriefed.

Study sessions were conducted by experimenters of similar age to the participants and in casual dress. Experimenters were unaware of which frame the participant had viewed. To reduce potential research authority effects,

all responses apart from test answers were collected through a laptop computer, which the experimenter did not view and no questions were answered directly to an experimenter.

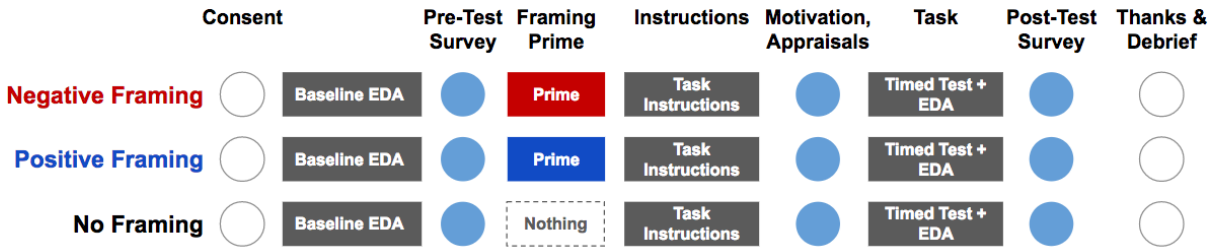

Fig. 4. Study Design. We assessed Framing effects by comparing dependent variables within participants (pre- and post-test scores on surveys) and between framings (Negative, Positive, and No Framing Control).

7 DATA ANALYSIS

7.1 Quantitative Analysis

Quantitative data was analyzed using a series of repeated measures ANOVAs for pre-post measures with group as a between-subject variable and time as a within-subject variable. Measures taken only once were analyzed through a series of one-way ANOVAs. Post hoc analyses were conducted to identify which group differences drove statistical significance.

In addition, we conducted a series of one-way ANOVAs and chi-squared tests to check for baseline differences that could alternatively explain results. These tests showed no significant differences in pre-test Trait Anxiety ($p=.622$), pre-test State Anxiety ($p=.686$), college GPA ($p=.971$), age ($p=.156$), motivation ($p=.747$), reported number of felt vibrations ($p=.231$), system accuracy ($p=.756$), willingness to use, ($p=.713$), or Gender ($p=.169$). There was also no differences between conditions in word count for participant free-writes ($p=.361$) with participants on average producing 38.44 words (SD: 19.79).

7.2 Qualitative Analysis

Quantitative analyses were supported with qualitative analyses of participant free-writes. Participant descriptions of their experiences were analyzed using open, inductive coding to identify emerging themes [72]. From 39 open codes, we found key themes relating to how EmVibe accuracy is assessed, as well as whether and when participants defer to system interpretations.

To evaluate the valence and emotionality of participants' written descriptions, their text was processed computationally using Linguistic Inquiry Word Count (LIWC) software [56]. LIWC is a widely used lexical analysis tool, which includes up to 72 different linguistic categories and has good reliability compared with human judges [55]. LIWC allows us to process the frequency of different emotional language that participants may have used to describe their experiences.

There are limitations to using only automatic approaches, which can fail to detect implicit expressions of emotion [76] so two researchers also individually coded the written participant descriptions of the test experience,

categorizing descriptions by emotional valence as 1 ("Negative"), 2 ("Neutral or Mixed Valence"), or 3 ("Positive"). We calculated inter-rated agreement of the valence ratings and an average of the ratings by the two coders was used to statistically compare the valence of free-writes across the conditions. In presenting qualitative findings and illustrative quotes, conditions are referenced as Positive Framing 'P', Negative Framing 'N', and No Framing Control 'C', along with the participant ID.

8 RESULTS

We present both the qualitative and quantitative findings in conjunction, as they relate to the main research questions. These mixed-methods results show the following:

(1) Participants actively evaluated how the system operated, focusing on which of their behaviors triggered system feedback. Some participants viewed the system as surpassing their own abilities, providing apparent insights and augmenting self-awareness (RQ3).
(2) Both positive and negative version of EmVibe were judged as moderately accurate, with some even deferring to the system when it conflicted with their personal evaluations. There were no differences in ratings of accuracy between conditions (RQ3).
(3) Perceptions were altered by system frames. The Negative description resulted in the greatest increases of anxiety, and greatest use of anxiety-related language in the post-test free-write. In contrast, Positive Framing had some benefits with those participants describing their experiences most positively (RQ2).

8.1 Strategies to Determine Accuracy

First, we observed participants' reported strategies for making sense of EmVibe outputs and, importantly, whether they considered such outputs accurate (RQ3).

8.1.1 Many Factors were Hypothesized to Trigger EmVibe: Consistent with research on physical tracking applications [77], participants reported making clear efforts to evaluate how EmVibe worked. They attempted to develop a conceptual model [54] by hypothesizing what triggered vibration feedback. Overall a broad set of mental and emotional triggers were believed to prompt feedback. Consistent with experimental system descriptions, some participants thought that feedback was triggered when they were more "*focused*" (P30), when their "*mind worked more*" (P33), or were "*stressed*" (N2). However, others imputed more advanced system capabilities, thinking that vibrations were triggered when the system "*sensed a change in me*" (N60), when "*I was questioning something*" (P19) or "*hesitated*" (P21). Some of these hypothesized triggers ("*sensed a change in me*", "*questioning something*") suggest that participants interpret the system as having extremely advanced analytic capabilities, that went beyond the descriptions that we provided.

But participants not only focused on which factors triggered EmVibe, they also evaluated whether feedback was accurate. A common evaluation method was to determine whether system feedback correlated with external information such as answer performance or question type (n=15). For example, this participant compared vibration feedback against perceived question difficulty: "*...During the hardest questions when I was struggling, I could feel the sensor buzz more frequently than the questions that I found to be easier.* " (P16). Similarly, this participant thought that vibrations were prompted by increasing cognitive challenges in the test: "*...[EmVibe] detected when I was stressed since it started vibrating as the test went further on and [became] increasingly difficult...*" (N2).

Another participant evaluated accuracy by correlating vibration feedback with perceived physical engagement:

> "*I noticed that the* ***sensor would vibrate when I came closer to the computer*** *to examine the pictures more carefully...as* ***I was processing what was going on*** *in the pictures I also* ***felt it vibrate a number of times in accuracy.***" (P34)

However, a small number of participants (n=6) were skeptical, giving accuracy ratings below 4. They considered accuracy to be low because it did not confirm their expectations:

> "*It would go off mostly after I got the answer right, and* ***it didn't go off when I would consider myself stressed like on the more complicated patterns*** *and I was running out of time.*" (N9)

8.1.2 EmVibe was Evaluated as Moderately Accurate: Following this active testing, most participants concluded that the system was reasonably accurate. Accuracy ratings (assessed in a survey question) were moderately positive (M:4.69, SD: 1.40), and significantly greater than neutral ('4') on a 7-point scale (t(41)=3.184, p=0.003). Furthermore, judgments were not different across framings. Participants in both Positive (M: 4.61, SD:1.33) and Negative Framings (M:4.75, SD:1.48) have similar accuracy ratings in a one-way ANOVA (F(1,40)=.098, p=.756). In other words, participants thought that the system was equally accurate when they were told that the same EDA algorithm was measuring that they were 'stressed' or 'engaged and alert'. Note, Control participants were not asked for accuracy judgments as they had no system description and therefore no criterion against which to evaluate outputs.

8.2 Negative Framing: Increased Anxiety and Stress

These results indicate that participants believed both versions of EmVibe to be accurate. Consequently, participants in the different framing conditions reported very different emotional and testing experiences (RQ2). When the system algorithm was described as detecting elevated 'stress' (i.e. Negative Frame), there were increased self-reported mental and physical symptoms of anxiety, and use of anxiety-related language. The same negative effects were not seen when the system was described as detecting increased 'alertness and engagement' (Positive Frame), or when no description was provided (Controls).

Negative Framing Increased Anxiety: We quantitatively tested the effects of emotional framing on self-reported physical and mental symptoms of anxiety (STAI-state survey). Participants' reactions were more negative when the algorithm was presented as assessing 'stress'. A two-way mixed ANOVA compared Anxiety scores across the 3 conditions with Time (Pre vs Post) as the within factor and Framing (Positive, Negative and No Framing Control) as the between subjects factor. Confirming our expectations that framing effects during the test would affect anxiety symptoms, there was a significant interaction of Time by Framing (F(2,61)=5.895, p=.005, partial eta squared =.162). Tukey's HSD procedure for pre-post Anxiety change scores shows that this interaction is driven by differences between Negative Framing compared with Positive Framing (p=.005) and Control (p=.050) participants. Negative Framing participants had the greatest increase in Anxiety, while Positive Framing participants had the smallest increase (see Figure 5).

Negative Framing Increased Task Stress Ratings: Negative Framing effects are also somewhat reflected in participants' perception of how stressful they considered the test. Task Stress was assessed at post-test with a 7-point scale. A one-way ANOVA for Task Stress rating showed a trending difference across Framings (F(2,61)=2.727, p=.073). Negative Framing participants having on average a higher score (M:4.63, SE:0.33), relative to Positive Framing (M: 3.45, SE:0.41), and Control (M:4.25, SE:0.36) participants.

Negative Framings Elicited More Use of Anxiety Terms to Describe Testing Experiences: In addition, we asked participants to provide a free-write description of their experience. A prompt encouraged participants to state in their own words which aspects of the testing experience they liked or disliked. We saw effects of Negative Framing in how participants characterized that test taking experience with those participants using more Anxiety-related

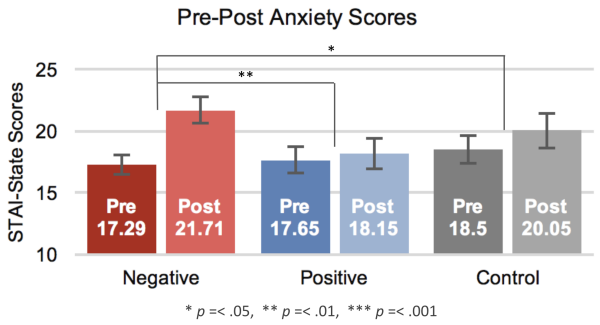

Fig. 5. Anxiety Scores (STAI). Pre vs post means and standard error for Anxiety (STAI-state) scores. Negative Framing participants had the greatest increase in anxiety from before to after the exam.

language in their descriptions.

LIWC software [56] was used to automatically analyze the language that participants used to describe their experiences. We analyzed only categories that were relevant to our research questions: positive emotion terms (e.g., 'happy', 'enjoyed'), negative emotion terms (e.g., 'hate', 'unhappy'), and anxiety-specific language (e.g., 'nervous', 'anxious'). As expected, there was a significant difference across Framings for anxiety terms with Negative Framing participants (M:2.65, SE:0.38) showing the greatest use of anxiety terms relative to the Positive Framing (M:0.56, SE:0.41), and Controls (M:0.92, SE:0.41), F(2, 61)=8.486, p=.001 (see Figure 6, Left). However, there was no difference across conditions for overall use of negative or positive language (ps .36 - .84)

Positive Framing Led to More Positive Assessments of the Testing Experience: Two researchers also individually coded the written participant descriptions of the test experience, categorizing descriptions by emotional valence as 1 ("Negative"), 2 ("Neutral or Mixed Valence"), or 3 ("Positive"). Inter-rater agreement was good (Cohen's k=.612, p<.0005) and an average of the two ratings was used to assess valence differences across framings. Again we saw Framing effects in a one-way ANOVA F(2,61)=4.001, p=.023, partial eta squared=0.12. Pairwise post hoc comparisons between Framing conditions showed this difference was driven by differences between Positive and Negative Framing (p=.019). As shown in the right hand side of Figure 6, Positive Framing participants showed greater expressions of positive affect (M:2.56, SE:0.15) than Negative Framing participants (M:2.0, SE:0.14). Control participants (M:2.18, SE:0.15) expressed intermediate levels of affect.

No Significant Difference in Demand-Resource Appraisals: One mechanism that might explain results is that the pre-test framing material can influence cognitive appraisals and consequently affect [37]. In other words, cognitive evaluations may be influenced by simply explaining the task, along with the experimental frames. To address this possibility, we compared pre-test Appraisals that were collected after reading framings for condition, but prior to the timed test. A one-way ANOVA for pre-test Appraisals showed no significant differences across the 3 framings, F(2,61)=.109, p=.897.

8.3 Positive Framing Benefits

We also wanted to know whether Positive Framing participants perceived benefits from getting feedback about when they are 'alert and engaged'. In contrast to heightened stress for Negative Framing participants, 5 Positive Framing participants described performance benefits from feedback: "*It was nice to know when I was*

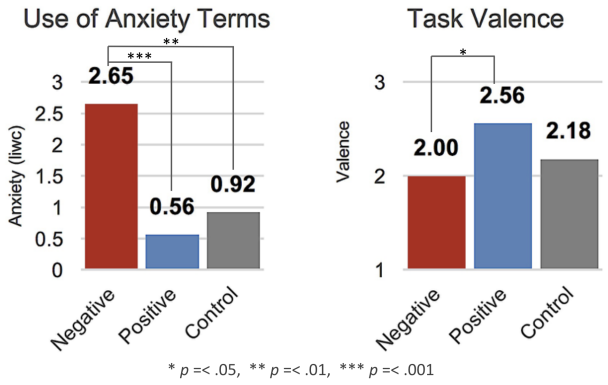

Fig. 6. Use of Anxiety language (LIWC) and qualitative coding of task valence. Negative Framing participants had the greatest use of Anxiety Language, and were least positive when describing the test-taking experience.

*engaged...**I believe I would not have done as well without it**...served to help me focus a lot harder*" (P16) or that it "*...reminded me to stay on task and motivated me*" (P21).

Importantly, these inferred benefits were specific to Positive Framing, as Negative Framing and Control participants did not describe these performance benefits. There was also discussion of expected future benefits with feedback on one's engagement and alertness being used to "*keep me being productive in class*" (P35) or to generally "*be more productive and approach situations with more understanding*" (P24). Despite identifying these potential benefits, Positive Framing participants did not actually perform better on the test ($F(2,59)=1.003$, $p=.373$).

In summary, there were strong quantitative system framing effects on emotion perception. Participants who believed that EmVibe was assessing stress (Negative Framing) were more likely to report higher levels of anxiety in post-test surveys, to use more anxiety terms in their descriptions of their testing experiences, and to be more negative in their characterizations of that experience.

8.4 Value and Perceived Insight Due to Feedback

These effects suggest that participants were directly influenced by system feedback. To explore these effects further we conducted a qualitative analysis to examine how participants evaluated system interpretations when compared against their own subjective experiences. The analysis revealed potential values and risks in emotion monitoring systems. While gaining additional information about the self could promote insight and augment self-awareness [41], several acknowledged that feedback itself could be a self-fulfilling prophecy.

8.4.1 EmVibe Provided Apparent Insights: Some participants treated system feedback as more reliable than their own intuitions about their stress and engagement levels. The following participant detected an inconsistency between their own evaluations and EmVibe's, overriding their impression in favor of the system: "*Sometimes when I felt more engaged it did not buzz me, which makes me think **maybe I'm still distracted***." (P35). Consistent with this pattern of deferring to the algorithm, EmVibe was also thought to provide direct insights into mental workload:

> "*When I had an idea of what the correct answer was, **my mind worked more and the watch vibrated.** When I had no idea, **my mind didn't work as hard and I noticed the watch didn't go off**...*" (P33)

In another case the system was initially judged to fail, but these reservations were overridden and its feedback rationalized to being seen as accurate: "*It detected stress when I did not feel stressed. Perhaps it may have been because it was* ***the minimal good stress*** *that drives someone to try and perform well on a task such as this one*" (N5).

8.4.2 System Feedback to Augment Self-Awareness: Echoing results from Study 1, system feedback also appeared valuable to participants who considered themselves less self-aware, allowing them to generate new self-insights when using the system. As N62 describes "*...it would be interesting to track the times when my emotions change,* ***it might change without a person realizing*** *or we can find out something that we didn't know before*".

Other participants thought system feedback could offer insights that exceed their own capabilities. C39 describes how emotion feedback "*...****could tell me about an emotion I don't know that I am feeling****...I feel like this isn't something that we can accurately keep track of on our own*". Again participants who felt they lack emotional self- awareness seemed more likely to find feedback especially informative: "*it* ***unlocks*** *a part of my experience that I am typically curious about*" (P20).

We did not recruit by, or ask about, mental health status. However, one participant spoke to the potential value of EmVibe for managing their mood disorder. An important consideration in deploying these systems is that participants' mental health status may affect adoption and trust:

> "*...I personally struggle with a mood disorder and, in the present moment,* ***it's difficult for me to assess my emotions and feelings accurately****. Having* ***this device would be a good indicator of how I should go about my mental health****.*" (C53).

Participants also used EmVibe feedback to confirm personal beliefs about their self-image:

> "*...when I was unsure ... or would get [an answer] wrong. it buzzed which meant I was alert or engaged.* ***I think of myself as a person who likes to do well*** *and so* ***it makes sense that when I am struggling, my mind is more engaged*** *.*" (P25).

8.5 Feedback as a Self-Fulfilling Prophecy

Participants tended to view EmVibe as providing reasonably accurate information about their internal state. However, they did not always view this information as helpful. This was particularly true for Negative Framing participants. As in prior work [48, 68], 11 participants spontaneously discussed issues with increased stress due to feedback. Rather than providing helpful insights, EmVibe seemed to exacerbate their stress.

For example, as N28 describes: "*...the biofeedback becomes* ***a self-fulfilling prophecy****...rather than cope and calm down, I just get more stressed.*" Similarly, participant N59 states "*I wasn't sure when I was feeling stress until the watch started vibrating.* ***feeling stresses*** *[sic], ended up making me feel more stressed*".

Some participants considered the device an additional source of pressure to feel less stress "*...I would want to know what stresses me out at times, but then* ***I feel as if I'm being pressured to not feel stressed****.*" (N64). This depended on system interpretations however; 2 Positive Framing participants also noted self-reinforcing effects but instead considered feedback as beneficial in making them more attentive and "*become a lot more alert*" (P31).

These results highlight the double-edged sword of analytics for PI systems. Such tools can be valuable in their potential to exceed human capabilities (e.g., sensing, statistical capabilities, memory storage). As such, they can

engender *newfound insight* in the face of expectation violations [41]. However, they also suggest the possibility of *deception* and self-fulfilling effects when participants lack alternative, trustworthy sources of information to compare system outputs against. We now summarize the findings and relate these to lines of promising future research.

9 DISCUSSION AND CONCLUSIONS

We explored different system interpretations in the context of emotion sensing systems. The first study revealed that despite strong interest in emotion and stress tracking that participants had little direct knowledge of such systems or the quality of data that they might generate. Study 2 followed up by showing that in the absence of full knowledge about how an algorithm operates, participants largely trusted EmVibe and attributed overly sophisticated abilities to the system. The result of these behaviors were significant changes in emotion perception and task assessments, dependent on the experimental system descriptions.

9.1 Overly Sophisticated System Attributions

These findings extend related work showing that users experience problems in interpreting complex personal data [9, 28, 38] and assume sophisticated capabilities of algorithms that assess affect [62, 68]. Unlike prior work which has focused on accuracy assessments for a single system version, we show no differences in perceived accuracy across both negative and positive framings. Some participants even proposed working hypotheses to explain their assessments of EmVibe's accuracy. On one hand, plausible interpretations included detecting emotional reactions to progression on the test. More strikingly, others believed that EmVibe could infer complex mental processes, such as mental workload, tackling easy versus hard questions, clarity of thought, or 'good' versus 'bad' stress.

This reveals clear issues for emotion-feedback systems. While prior studies indicate that users find it difficult to understand the meaning of data from personal systems [9, 32, 38], our results suggest a different but related problem. In the absence of concrete knowledge about how an algorithm operates, rather than discount the system, users may adopt incorrect conceptual models that defer to, and confirm system accuracy. This confirms prior work on folk theories that users develop about how algorithms operate [21, 70, 74].

9.2 Emotional Consequences of System Authority

We extend the work of [62, 68, 74] in showing that participants often defer to and trust system inferences for very personal aspects of the self. Unlike prior work, we present evidence that different framings for the same physiological data are perceived as equally accurate, with consequences for reported emotional reactions. Strikingly, EmVibe feedback was sometimes considered to surpass one's own abilities. Some participants even believed EmVibe revealed thoughts or feelings that they were unable to sense for themselves. Both studies show that those most interested in adopting stress monitoring systems expect that these tools will serve as an objective measure of their experiences and correctly improve self-understanding.

Furthermore, some individuals may be more susceptible than others to defer to system interpretations. Participants who described themselves as having lower emotional awareness readily accepted system interpretations. This might amplify the effects of framing depending on participants' trust in their own self-assessments and is a promising area for future research.

Unlike prior work which has shown that false feedback can change the magnitude of a stress response [15] or that real feedback amplifies stress [48, 68], we are the first to show that the meaning assigned to emotion

detection systems can actually influence *valence*. Our work highlights the potential for emotion detection systems to transform a negative emotional experience into a more positive one, as shown by the differences in valence of participant free-writes. While automatic emotion detection systems have often had a negative focus (e.g. stress), positive emotion feedback can have subjectively beneficial effects. This work also highlights the key role that system outputs play in the emotion perception process. However, this leaves open questions regarding specific mechanism. Study 2 extends work by [15] in that we tested whether frames alone could influence cognitive appraisals, but found no evidence for that as an explanatory mechanism. We encourage future research to further explore the predictive power of different theories for how algorithm frames and feedback influence emotion perception.

These new results suggest important design implications for PI and particularly for emotion-monitoring systems.

9.3 Design Considerations

For emotion regulation, designers need to exercise caution to anticipate self-fulfilling effects. These effects can result both from how emotion-relevant data is framed, as well as from implicit system goals about how to manage emotional states, e.g., regulation versus acceptance [29]. At the same time, results also point to unique design opportunities. While much of the current focus in affective computing is precision of algorithms, we encourage designers to consider the subjective component of mapping low level data to higher level constructs [5]. Our results raise important questions for system designers when presenting emotion feedback:

What meaning should be assigned to emotion inferences?
Many systems attempt to improve emotion regulation by increasing awareness of emotional states (e.g., [32, 48]) or by presenting false feedback to regulate negative emotion [15]. Our data suggest that, instead, systems could influence perceptions of affect, so emotional states are interpreted more positively. For example, rather than present negative, threatening states (e.g., 'stress'), feedback instead could help users construe these same states as challenges (e.g., 'engagement'). This reconstrual strategy has potentially greater long-term success, given inaccurate feedback could erode user trust [77] and accurate feedback framed as stress can exacerbate stress [48, 68]. However, such design decisions have clear ethical considerations [3].

Who should assign meaning to inferences?
This also raises the question of who should assign meaning to emotion-relevant data. Our results highlight deference effects, and one alternative design approach could encourage active appraisal by encouraging users to select from options of emotions being surfaced by the system. For example, a system might present a potential set of emotional states (e.g., 'focused', 'alert', 'excited') that could map to automatic inferences, rather than subjecting users to one definitive interpretation. Another design alternative might be to avoid emotional labels altogether by using suggestive ambient outputs [63, 68]. Such approaches support active appraisal of emotions and allow for more complex mixed-emotional states. Increased language granularity in appraising one's emotions has been associated with a range of benefits such as more successful emotion regulation [39]. A key question for future work is *how* different presentations of emotions data help or hinder emotion granularity.

Under what contexts should systems present feedback?
Seven participants spontaneously described being distracted by the vibrations. While this finding differs from [15], it draws attention to another important design consideration; we need to be judicious about when and how we provide analytic feedback. Users who are engaged in a complex intellectual task may not only be distracted

by such feedback, but furthermore, they may also lack the attentional capacity to interpret and act upon that complex feedback. This is especially important as we see that in Study 1 participants most wanted stress tracking in an exam context, yet this may be counterproductive without presenting appropriate regulation strategies that could be executed while carrying out a complex mental task.

Is self-focus always beneficial?
These findings also raise a quite fundamental question about whether systems that direct attention to the self are always beneficial. Psychological research shows that self-focusing does not necessarily promote more accurate self- knowledge [66] and can even have detrimental effects [57]. [35] showed that reflective self-monitoring with a fitness tracker actually decreased weight loss. Other work highlights that arousal feedback can increase stress [48, 68], a result we replicate here. Furthermore, [32] found that targeted monitoring of highly negative emotions decreases well-being. Future work needs to address not just whether self-awareness leads to improvements, but under what conditions heightened self-focus might be *counterproductive*, or the types of information users should selectively focus on to avoid such counterproductive effects.

9.4 Limitations

Both studies have limitations. Although there is a clear justification for focusing on students, research with more diverse populations is necessary to generalize these results. We also acknowledge that Study 1 elicited *hypothetical* adoption intentions and expectations. As increasingly more commercial systems are developed, we could see a shift in users' mental models and expectations for such systems. Furthermore, respondents were not specifically probed about concerns they may have for emotion-sensing systems. Our findings show a lack of participant responses for concerns or skepticism of such systems. These results could differ if participants are instead probed more extensively about potential negative consequences of using such systems and perceived technological feasibility.

For Study 2, the testing scenario was designed after piloting to be moderately challenging, so it is unknown whether these findings would persist in more or less demanding settings. Second, there is a question of duration of use and effects of system authority. With longer exposure and a greater variety of experiences, users may learn that system inferences do not match their own evaluations (e.g., [77]), though other work has examples of deferral in real-world deployments [62]. Finally, we cannot be certain about the source of deferral effects. Although similar results are observed with different systems and outside the lab [62], our findings could be in part due to participants using EmVibe in-lab and with a researcher present.

An additional consideration is that participants in the experimental framings were implicitly instructed to regulate their emotions. This could contribute to the feelings of "pressure" some Negative Framing participants reported. If instead we primed participants with nonjudgmental self-awareness [29], participants may have experienced less distress from the negative framings. This is an important topic for future study on deferral and authority effects of emotion-monitoring systems. Finally, although the Pip is a state of the art sensor and participants judged the algorithm to be fairly accurate, Study 2 results may have been influenced by the quality of the underlying EDA algorithm.

10 CONCLUSION

We explored deferral to algorithmic interpretations in the domain of emotion monitoring. Our results speak to a critical issue in Personal Informatics, namely the processes by which people make sense of algorithmic data. We also provide evidence showing that adoption of emotion-monitoring systems is a top interest, and that

motivations to adopt such systems differ from more traditional PI tools.

In Study 1, we show that there is a high interest among young adult respondents to adopt automatic stress- and emotion-monitoring systems. Furthermore, the top motivations to adopt these tools differed greatly from PI systems in the health and productivity space. For stress- and emotion-monitoring systems, respondents wanted to improve their self-understanding and have verifiable records of their experiences. However, they had extremely limited knowledge of existing tools for automatic emotion- or stress-monitoring.

Study 2 explored the impact of an emotion-monitoring system on the stress experiences of taking an exam. We show that despite active efforts to interpret such complex data, participants tend to be highly influenced by system feedback. Participants who were provided with negative interpretations of the same physiological information reported greater anxiety, while participants who were provided with positive interpretations were more positive about the study experience. These findings show that not only are systems capable of influencing perceived emotional intensity [15], but that such systems can also influence the emotional valence of an experience. These findings extend [15] by demonstrating system authority effects in a different research context, and providing rich, qualitative results for the processes by which users defer to system outputs for emotion.

Our work also provides design considerations for future research and systems exploring how, when, and why we might present self-relevant personal data. We encourage system designers to consider the alternative types of interpretations that could be made for emotion data and who should be the authority in such interpretations.

ACKNOWLEDGMENTS

We would like to thank our research assistants, Connor Harada, Katie Reed, Shravya Neeruganti, Ying Yan, Jodi Buddine, Precylla Ruiz, and Lindsay Nelson, for their valuable help with this research. We would also like to thank our research participants for sharing their time and thoughts with us. This research was supported by NSF grant IIS-1321102.

A APPENDIX A: SURVEY QUESTIONS FOR STUDY 1

Author's Note: Below is a subset of questions used in Study 1. Questions were tailored based on responses to Question 4 (Q4) for aspects of one's self that the respondent wanted to have automatically measured.

In this survey, we will ask you about any self-tracking technologies that you are aware of, what you would like a tool to automatically track for you, and how well you consider yourself to currently track some characteristic (e.g., the amount of sleep you get). We will also ask you some questions about how you reflect on yourself.

We recommend completing this survey on a computer.

Q1: Have you ever used a software application to track your mood or health? This includes phone apps or apps on your computer.

☐ Yes
☐ No
☐ Maybe/Don't know

Q2: Have you ever used a wearable fitness, health, or mood tracker?

☐ Yes

☐ No
☐ Maybe/Don't know

Q3: Please select all of the names of fitness trackers and/or self-tracking systems that you have heard of:
☐ Muse
☐ Fitbit or Jawbone
☐ BellaBeat, Spire or Pip
☐ Apple Watch or Withings
☐ RescueTime
☐ Daylio, Moodpanda, or MoodScope
☐ Mint
☐ MyFitnessPal
☐ Runkeeper or Runtastic
☐ Samsung S Health, Google Fit, or Apple HealthKit
☐ Other: _____
☐ None of the above

Q4: What are the *top 3* aspects of yourself that would you be interested in measuring automatically throughout your day and night?
☐ How well I sleep
☐ My stress levels
☐ My mood and/or emotional states
☐ My focus and/or attention
☐ Use of my time
☐ Work productivity
☐ Calories burned
☐ How far I walk, run or cycle
☐ Whether I am dehydrated
☐ My weight
☐ My blood pressure
☐ How long I spend sitting
☐ Tracking my location, running routes
☐ Other:_____
☐ None of the above

Author's Note: The following questions were displayed to the respondent if their response to Q4 included the selection of 'My stress levels'. For length considerations, only the stress-related questions are presented here. A similar set of questions were presented if the respondent selected 'My mood and/or emotional state' for Q4.

Q5: Rate how interested or not interested you are in using a tool that could automatically measure your stress levels:
☐ Extremely interested
☐ Very interested
☐ Moderately interested
☐ Slightly interested
☐ Not at all interested

Q6: (Optional) You selected that you are interested in automatically measuring your stress levels. Please write two or three sentences explaining why. Note: Please do not disclose any personally identifiable or sensitive in information in this field:

_____

_____

_____

Q7: What is your top motivation to track your stress?

- ☐ To achieve a goal
- ☐ To improve in this area, without a specific goal
- ☐ To document my life accurately
- ☐ To understand how this factor relates to others (for example, the impact of sleep on mood)
- ☐ To explore a new technology
- ☐ To predict how I will do in the future
- ☐ To understand myself better
- ☐ Other:_____
- ☐ Don't know/unsure
- ☐ Decline to state

Q8: Please explain your answer for the above question. Why is this your *top* motivation?

_____

_____

_____

Q9: Please select all contexts for which you would want to have automatic stress tracking:

- ☐ During leisure activities
- ☐ Consistently throughout the day, regardless of context
- ☐ When I am at work
- ☐ When I am at school
- ☐ During public speaking
- ☐ During an exam
- ☐ In a job interview
- ☐ When my mind is occupied with a task
- ☐ When I am interacting with friends
- ☐ When I am interacting with co-workers
- ☐ When I am interacting with family
- ☐ When I am alone
- ☐ Other:_____
- ☐ None of the above

Q10: Which of the following ideas about stress do you think technology should reflect?

Imagine you could use a stress tracker. These trackers can reflect different ideas about stress. For example, some technologies describe stress as a positive thing, while others give you goals to reduce it.

Please rate the following options that a stress tracker *should* reflect (toward 1) to options it should *not* reflect (toward 12).

(1) The stress tracker should help you control stress
(2) The stress tracker should help you acknowledge your stress
(3) The stress tracker should help you reduce your stress
(4) The stress tracker should help you increase your stress
(5) The stress tracker should help you find an optimal stress balance
(6) The stress tracker should help you find stress triggers
(7) The stress tracker should reveal what types of stress are good for you
(8) The stress tracker should help you have more accurate information about yourself
(9) The stress tracker should hide negative information if it won't reduce your stress
(10) The stress tracker should *not* give you specific goals to reduce stress
(11) The stress tracker *should* give you specific goals to reduce stress
(12) The stress tracker should have you set your own stress goals

Q11: Have you ever heard of the term 'Quantified Self'?
☐ Yes
☐ No
☐ Maybe/Don't know

Q12: Are you a member of the 'Quantified Self' community?
☐ Yes
☐ No
☐ Maybe/Don't know

Q14: Enter your age: _____
Q15: Enter your self-identified gender: _____
Q16: What is your occupation? _____

Author's Note: Participants were then provided with a debrief and researcher contact information

B APPENDIX B: TEST INSTRUCTIONS

In this study, you will answer 36 questions related to overall fluid intelligence. These problems may include some questions that will be potentially more challenging. Please remember that you are free to stop the study at any point and you will still receive full credit if you do so.

For each question, you will see several possible options labeled 1, 2, 3, 4, 5, 6, 7 or 8. Once you have decided your final answer:

(1) Select that answer on the screen
(2) Announce your final answer aloud to the experimenter. You cannot change your answer after announcing it.
(3) Click submit on the screen

You will get verbal feedback on whether your answer was correct or incorrect so that you can keep track of your performance.

You will have a strict time limit to answer the questions. The amount of time remaining will be displayed at the top of each page in red text.

Please try your best to answer these questions correctly and quickly.

C APPENDIX C: SYSTEM DESCRIPTION

Author's Note: Participants read the following system description with text varied based upon condition (Negative or Positive Frame). Participants who saw this content were required to answer a comprehension check question which varied by condition. No Frame Controls were not presented with any system description but still used the system and received feedback.

We are evaluating new wearable technology which gives immediate feedback on your ['high-stress'/'highly alert'] states so that you can better control them. You will hold an EDA sensor between the thumb and index finger of your non-dominant hand. You will also wear a band on your non-dominant hand and it will provide real-time vibration feedback when you are ['hitting a high-stress peak'/'highly alert']. [Positive Frame: 'A lack of vibration feedback indicates you are less alert and engaged.']

[Negative Frame] When you feel vibration feedback on your wrist, this indicates:

- That you are in a calm state
- That you are in a stressed state
- Nothing
- That you are answering questions too slowly

[Positive Frame] When you feel vibration feedback on your wrist, this indicates:

- That you are in a bored, unalert state
- That your are in an alert and engaged state
- Nothing
- That you are answering questions too slowly

D APPENDIX D: EDA DESCRIPTIONS

Author's Note: Participants were given an explanation of EDA with text varied based on condition (Positive or Negative Frame). Participants who saw this content were required to answer a comprehension check question which varied by condition. No Frame Controls were not presented with any description of EDA.

While we may not have the ability to control what events occur in our lives, we can control our ['stress response'/'alertness and engagement'] to handle demands better.

Our system uses Electro Dermal Activity (EDA) to measure changes in the sympathetic nervous system which is indicative of increased ['stress'/'alertness and engagement']. EDA has been verified as an accurate detector of ['stress'/'alertness and engagement'] in a wealth of prior research [citations varied per condition].

Today you will use a Pip device to measure EDA. The Pip uses 2 electrodes to measure the amount of electricity your skin conducts. When are you in a high stress state, activity in your Sympathetic Nervous System (SNS) increases and this increased SNS activity is reflected in an increased amount of sweat. The Pip has two electrodes and uses the skin to conduct electricity. Increased moisture causes your skin to transmit more electricity between the electrodes and results in a higher EDA reading. You can look at the diagram below to see where the two

electrodes touch the fingers.

[Participants saw an illustration of a hand with the index and forefinger highlighted.]

EDA is one of the best known external indicator of [‘stress activation’/‘alertness and engagement’] as skin is the only organ purely activated by the SNS (Boucsein, 1992). Throughout the study, we will be monitoring your [‘stress levels’/‘alertness and engagement’] through these spikes in Electro Dermal Activity (EDA).

[Participants then answered a comprehension check question]

[Negative Frame] What does Electrodermal Activity (EDA) measure?

- Increased activity in the sympathetic nervous system which indicates stress
- Decreased activity in the sympathetic nervous system which indicates stress
- Increased activity in the sympathetic nervous system which indicates relaxation

[Positive Frame] What does Electrodermal Activity (EDA) measure?

- Increased activity in the sympathetic nervous system which indicates greater alertness and engagement
- Decreased activity in the sympathetic nervous system which indicates greater alertness and engagement
- Increased activity in the sympathetic nervous system which indicates lower alertness and engagement

REFERENCES

[1] ACHA. 2014. American college health association-national college health assessment II: Undergraduate students reference group executive summary spring 2014. *Hanover, MD: American College Health Association* (2014).

[2] Phil Adams, Mashfiqui Rabbi, Tauhidur Rahman, Mark Matthews, Amy Voida, Geri Gay, Tanzeem Choudhury, and Stephen Voida. 2014. Towards personal stress informatics: Comparing minimally invasive techniques for measuring daily stress in the wild. In *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 72–79.

[3] Eytan Adar, Desney S Tan, and Jaime Teevan. 2013. Benevolent deception in human computer interaction. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1863–1872.

[4] Alper T Alan, Mike Shann, Enrico Costanza, Sarvapali D Ramchurn, and Sven Seuken. 2016. It is too hot: An in-situ study of three designs for heating. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5262–5273.

[5] Lisa Feldman Barrett. 2006. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review* 10, 1 (2006), 20–46.

[6] Lisa Feldman Barrett. 2017. *How emotions are made: The secret life of the brain.* Houghton Mifflin Harcourt.

[7] Victoria Bellotti, Maribeth Back, W Keith Edwards, Rebecca E Grinter, Austin Henderson, and Cristina Lopes. 2002. Making sense of sensing systems: five questions for designers and researchers. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 415–422.

[8] Victoria Bellotti and Keith Edwards. 2001. Intelligibility and accountability: human considerations in context-aware systems. *Human–Computer Interaction* 16, 2-4 (2001), 193–212.

[9] Frank Bentley, Konrad Tollmar, Peter Stephenson, Laura Levy, Brian Jones, Scott Robertson, Ed Price, Richard Catrambone, and Jeff Wilson. 2013. Health Mashups: Presenting statistical patterns between wellbeing data and context in natural language to promote behavior change. *ACM Transactions on Computer-Human Interaction (TOCHI)* 20, 5 (2013), 30.

[10] Kirsten Boehner, Rogério DePaula, Paul Dourish, and Phoebe Sengers. 2007. How emotion is made and measured. *International Journal of Human-Computer Studies* 65, 4 (2007), 275–291.

[11] Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *ICWSM* 11 (2011), 450–453.

[12] Wolfram Boucsein. 2012. *Electrodermal activity.* Springer Science & Business Media.

[13] Erin A Carroll, Mary Czerwinski, Asta Roseway, Ashish Kapoor, Paul Johns, Kael Rowan, and M C Schraefel. 2013. Food and mood: Just-in-time support for emotional eating. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 252–257.

[14] Eun Kyoung Choe, Nicole B Lee, Bongshin Lee, Wanda Pratt, and Julie A Kientz. 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 1143–1152.
[15] Jean Costa, Alexander T Adams, Malte F Jung, François Guimbetiere, and Tanzeem Choudhury. 2016. EmotionCheck: leveraging bodily signals and false feedback to regulate our emotions. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 758–769.
[16] Michael A DeVito, Jeremy Birnholtz, Jeffery T Hancock, Megan French, and Sunny Liu. 2018. How People Form Folk Theories of Social Media Feeds and What It Means for How We Study Self-Presentation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 120.
[17] Alison Dillon, Mark Kelly, Ian H Robertson, and Deirdre A Robertson. 2016. Smartphone applications utilizing biofeedback can aid stress reduction. *Frontiers in psychology* 7 (2016).
[18] G Drake, E Csipke, and T Wykes. 2013. Assessing your mood online: acceptability and use of Moodscope. *Psychological medicine* 43, 7 (2013), 1455–1464.
[19] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.
[20] Phoebe C Ellsworth and Klaus R Scherer. 2003. Appraisal processes in emotion. *Handbook of affective sciences* 572 (2003), V595.
[21] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First i like it, then i hide it: Folk theories of social feeds. In *Proceedings of the 2016 cHI conference on human factors in computing systems*. ACM, 2371–2382.
[22] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. I always assumed that I wasn't really that close to [her]: Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 153–162.
[23] Feel. 2017. Feel Wearable. http://www.myfeel.co/
[24] Barbara L Fredrickson, Michele M Tugade, Christian E Waugh, and Gregory R Larkin. 2003. What good are positive emotions in crisis? A prospective study of resilience and emotions following the terrorist attacks on the United States on September 11th, 2001. *Journal of personality and social psychology* 84, 2 (2003), 365.
[25] Galvanic Ltd. 2016. Pip Stress Tracker. https://thepip.com/research-partnership/
[26] Pedro García García, Enrico Costanza, Jhim Verame, Diana Nowacka, and Sarvapali D Ramchurn. 2018. Seeing (Movement) is Believing: The Effect of Motion on Perception of Automatic Systems Performance. *Human–Computer Interaction* (2018), 1–51.
[27] James Gross and Ross Thompson. 2007. *Emotion regulation: Conceptual foundations*.
[28] Shad Gross, Jeffrey Bardzell, Shaowen Bardzell, and Michael Stallings. 2017. Persuasive Anxiety: Designing and Deploying Material and Formal Explorations of Personal Tracking Devices. *Human–Computer Interaction* (2017), 1–38.
[29] Steven C Hayes, Kirk D Strosahl, and Kelly G Wilson. 1999. *Acceptance and commitment therapy: An experiential approach to behavior change*. Guilford Press.
[30] Javier Hernandez, Zicheng Liu, Geoff Hulten, Dave DeBarr, Kyle Krum, and Zhengyou Zhang. 2013. Measuring the engagement level of TV viewers. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 1–7.
[31] Javier Hernandez, 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. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 307–317.
[32] Victoria Hollis, Artie Konrad, Aaron Springer, Matthew Antoun, Christopher Antoun, Rob Martin, and Steve Whittaker. 2017. What Does All This Data Mean for My Future Mood? Actionable Analytics and Targeted Reflection for Emotional Well-Being. *Human–Computer Interaction* (2017), 1–60.
[33] Ellen Isaacs, Artie Konrad, Alan Walendowski, Thomas Lennig, Victoria Hollis, and Steve Whittaker. 2013. Echoes from the past: how technology mediated reflection improves well-being. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1071–1080.
[34] Carroll E Izard. 1971. The face of emotion. (1971).
[35] John M Jakicic, Kelliann K Davis, Renee J Rogers, Wendy C King, Marsha D Marcus, Diane Helsel, Amy D Rickman, Abdus S Wahed, and Steven H Belle. 2016. Effect of wearable technology combined with a lifestyle intervention on long-term weight loss: the IDEA randomized clinical trial. *Jama* 316, 11 (2016), 1161–1171.
[36] Jeremy P Jamieson, Wendy Berry Mendes, Erin Blackstock, and Toni Schmader. 2010. Turning the knots in your stomach into bows: Reappraising arousal improves performance on the GRE. *Journal of Experimental Social Psychology* 46, 1 (2010), 208–212.
[37] Jeremy P Jamieson, Wendy Berry Mendes, and Matthew K Nock. 2013. Improving acute stress responses: The power of reappraisal. *Current Directions in Psychological Science* 22, 1 (2013), 51–56.
[38] Simon L Jones and Ryan Kelly. 2017. Dealing with Information Overload in Multifaceted Personal Informatics Systems. *Human–Computer Interaction* (2017), 1–48.

[39] Todd B Kashdan, Patty Ferssizidis, R Lorraine Collins, and Mark Muraven. 2010. Emotion differentiation as resilience against excessive alcohol use: An ecological momentary assessment in underage social drinkers. *Psychological Science* 21, 9 (2010), 1341–1347.

[40] Matthew Kay, Dan Morris, Julie A Kientz, and Others. 2013. There's no such thing as gaining a pound: Reconsidering the bathroom scale user interface. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 401–410.

[41] Elisabeth T Kersten-van Dijk, Joyce HDM Westerink, Femke Beute, and Wijnand A IJsselsteijn. 2017. Personal informatics, self-insight, and behavior change: A critical review of current literature. *Human–Computer Interaction* 32, 5-6 (2017), 268–296.

[42] Michelle Elisabeth Kruijshaar, Jan Barendregt, Theo Vos, Ron De Graaf, Jan Spijker, and Gavin Andrews. 2005. Lifetime prevalence estimates of major depression: an indirect estimation method and a quantification of recall bias. *European journal of epidemiology* 20, 1 (2005), 103–111.

[43] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1–10.

[44] Celine Latulipe, Erin A Carroll, and Danielle Lottridge. 2011. Love, hate, arousal and engagement: exploring audience responses to performing arts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1845–1854.

[45] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 557–566.

[46] Brian Y Lim and Anind K Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing*. ACM, 195–204.

[47] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2119–2128.

[48] Diana MacLean, Asta Roseway, and Mary Czerwinski. 2013. MoodWings: a wearable biofeedback device for real-time stress intervention. In *Proceedings of the 6th international conference on PErvasive Technologies Related to Assistive Environments*. ACM, 66.

[49] Mark Matthews, Jaime Snyder, Lindsay Reynolds, Jacqueline T Chien, Adam Shih, Jonathan W Lee, and Geri Gay. 2015. Real-time representation versus response elicitation in biosensor data. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 605–608.

[50] Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. 2012. AffectAura: an intelligent system for emotional memory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 849–858.

[51] Wendy Berry Mendes, Jim Blascovich, Brenda Major, and Mark Seery. 2001. Challenge and threat responses during downward and upward social comparisons. *European Journal of Social Psychology* 31, 5 (2001), 477–497.

[52] Wendy Berry Mendes, Heather M Gray, Rodolfo Mendoza-Denton, Brenda Major, and Elissa S Epel. 2007. Why egalitarianism might be good for your health: Physiological thriving during stressful intergroup encounters. *Psychological Science* 18, 11 (2007), 991–998.

[53] Neumitra. 2016. Neumitra Stress Tracker. https://neumitra.com/

[54] Donald Norman. 1988. *The Design of Everyday Things (Originally published: The psychology of everyday things)*.

[55] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.

[56] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[57] Tom Pyszczynski, James C. Hamilton, Fred H. Herring, and Jeff Greenberg. 1989. Depression, self-focused attention, and the negative memory bias. *Journal of Personality and Social Psychology* 57 (1989), 351–357.

[58] Mashfiqui Rabbi, Min Hane Aung, Mi Zhang, and Tanzeem Choudhury. 2015. MyBehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2015), 707–718.

[59] Lena Reed, Jiaqi Wu, Shereen Oraby, Pranav Anand, and Marilyn Walker. 2017. Learning lexico-functional patterns for first-person affect. *arXiv preprint arXiv:1708.09789* (2017).

[60] John Rooksby, Mattias Rost, Alistair Morrison, and Matthew Chalmers Chalmers. 2014. Personal tracking as lived informatics. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 1163–1172.

[61] Paul J Rosch. 2001. The quandary of job stress compensation. *Health and Stress* 3, 1 (2001), 1–4.

[62] Minna Ruckenstein. 2014. Visualized and interacted life: Personal analytics and engagements with data doubles. *Societies* 4, 1 (2014), 68–84.

[63] Pedro Sanches, Elsa Kosmack Vaara, Marie Sjölinder, Claus Weymann, Kristina Höök, and Elsa Vaara. 2010. Affective Health–designing for empowerment rather than stress diagnosis. *Know thyself: monitoring and reflecting on facets of one's life at CHI 2010, Atlanta, GA, USA* (2010).

[64] Sensoree. 2017. GER Mood Sweater.

[65] Eric Siegel. 2016. *Predictive analytics: The power to predict who will click, buy, lie, or die*. Wiley Hoboken (NJ).

[66] Paul J Silvia and Guido H E Gendolla. 2001. On introspection and self-perception: Does self-focused attention enable accurate self-knowledge? *Review of General Psychology* 5, 3 (2001), 241.

[67] Craig A Smith and Leslie D Kirby. 2009. Putting appraisal in context: Toward a relational model of appraisal and emotion. *Cognition and Emotion* 23, 7 (2009), 1352–1372.
[68] Jaime Snyder, Mark Matthews, Jacqueline Chien, Pamara F Chang, Emily Sun, Saeed Abdullah, and Geri Gay. 2015. Moodlight: Exploring personal and social implications of ambient display of biosensor data. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 143–153.
[69] R.L. Spielberger, C.D., Gorsuch. 1983. State-trait anxiety inventory for adults: Manual, instrument, and scoring guide.
[70] Aaron Springer, Victoria Hollis, and Steve Whittaker. 2017. Dice in the Black Box: User Experiences with an Inscrutable Algorithm. In *AAAI Spring Symposium Series*.
[71] Stanford Medicine. 2017. Stanford, Apple to Collaborate in New Heart Study. http://med.stanford.edu/appleheartstudy.html
[72] David R Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation* 27, 2 (2006), 237–246.
[73] Kristen Vaccaro, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The Illusion of Control: Placebo Effects of Control Settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 16.
[74] Jeffrey Warshaw, Tara Matthews, Steve Whittaker, Chris Kau, Mateo Bengualid, and Barton A Smith. 2015. Can an Algorithm Know the Real You?: Understanding People's Reactions to Hyper-personal Analytics Systems. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 797–806.
[75] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063.
[76] Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation* 39, 2 (2005), 165–210.
[77] Rayoung Yang, Eunice Shin, Mark W Newman, and Mark S Ackerman. 2015. When fitness trackers don't'fit': end-user difficulties in the assessment of personal tracking device accuracy. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 623–634.
[78] Sunmoo Yoon, Faith Parsons, Kevin Sundquist, Jacob Julian, Joseph E Schwartz, Matthew M Burg, Karina W Davidson, and Keith M Diaz. 2017. Comparison of Different Algorithms for Sentiment Analysis: Psychological Stress Notes. *Studies in health technology and informatics* 245 (2017), 1292.

Received May 2018; revised July 2018; accepted September 2018