A Longitudinal Analysis of Real-World Self-report Data

Niels van Berkel[1(✉)], Sujay Shalawadi[1], Madeleine R. Evans[2], Aku Visuri[3], and Simo Hosio[3]

[1] Aalborg University, Aalborg, Denmark
{nielsvanberkel,sujaybs}@cs.aau.dk
[2] Levell, London, UK
madeleine@levell.io
[3] University of Oulu, Oulu, Finland
{aku.visuri,simo.hosio}@oulu.fi

Abstract. While self-report studies are common in Human-Computer Interaction research, few evaluations have assessed their long term use. We present a longitudinal analysis of a web-based workplace application that collects well-being assessments and offers suggestions to improve individual, team, and organisational performance. Our dataset covers 219 users. We assess their first year of application use, focusing on their usage patterns, well-being evaluations, and behaviour towards notifications. Our results highlight that the drop-off in use was the steepest in the first week (-24.2%). However, substantial breaks in usage were common and did not necessarily result in dropout. We found that latency periods of eight days or more predicted a stronger intention to drop out than stay engaged and that reminder notifications did not result in more completed self-reports but significantly prolonged the usage period. Our work strengthens findings related to high drop out rates, but also provides counter-evidence by showing that despite individuals appearing to drop-off in short-term studies, individuals can and do return to self-report applications after extensive breaks. We contribute an analysis of usage behaviour drivers in the area of technology-enabled well-being measurement, responding to the call for longer-term research to extend the growing literature on self-report studies.

Keywords: Self-report · well-being · longitudinal · diary study · experience sampling

1 Introduction

The collection and analysis of self-report data has a long history in Human-Computer Interaction (HCI), particularly in subjective well-being research. Well-known examples of self-report-based studies are diary studies [41] and experience sampling studies [2,30], in which researchers provide participants with a self-report application (desktop or mobile) or device to regularly report on the

J. Abdelnour Nocera et al. (Eds.): INTERACT 2023, LNCS 14144, pp. 611–632, 2023.
https://doi.org/10.1007/978-3-031-42286-7_34

experiences or aspect(s) of their lives that are being studied. For example, self-report applications have been utilised to gather insightful and valuable data as to what drives the states of e.g. happiness, calm, pain, or peacefulness [22], as well as helped HCI researchers to understand aspects of digital interactions, such as usage patterns [19] and user experiences [8] among many others. In recent years, the Quantified Self (QS) movement has popularised the value of self-report data for personal use [10,29,33], driving popular interest in self-tracking applications [31]. Much in the manner that academic well-being researchers seek to pair self-report data on affective experience with contextual data such as activity logs, QS users seek to review the logs of self-report data for health, performance, or wider well-being insights that they can then use to improve their lives [29,32,46].

Prior work has highlighted the typically limited duration of self-report studies as a limitation in identifying persistent changes in use, self-reported affect, and real-world outcomes over time [17,24]. Kjaerup et al. analysed longitudinal studies published in CHI between 1982–2019 and found only 56 studies with a duration of more than one month [27]. A recent review by Caraban et al. on nudging technologies, *i.e.* technology that promotes subtle behaviour changes, finds that just 18 out of the analysed 50 studies had a duration longer than a day, with seven studies being over one month in duration [7]. The lack of longitudinal data in such studies can result in skewed insights into participant motivation, engagement, usage, as well as real-world experience and outcomes patterns, in particular in relation to the long-term usage of self-report applications. A lack of longitudinal research data can also lead researchers to draw unreliable conclusions. Bias due to the characteristics of individuals that participate in academic studies is another major concern, as individuals that opt to participate in research evaluations may not necessarily behave similarly to average users along the parameters of motivation and usage behaviour, limiting the insights drawn to that sample rather than to a wider population [35]. This 'participation bias' is a widely recognised research participation effect, perhaps most well-known to the reader in the context of course evaluations. In such evaluations, motivation to take part is highest among those with strong positive or negative sentiments toward the stated potential outcome or topic of the study at hand—thereby creating an effect of self-selection among the studied population [13].

In this paper we present an analysis of one year of usage data of a self-report application designed to enable individual users, teams and companies to track and improve experienced well-being in the context of work. Specifically, our application offers a web-based platform and is designed to be integrated in a professional context to inform and support employees and team leaders. Users of the application can track four dimensions of their subjective well-being, add and share ideas to improve well-being in the context of work, and report 'blockers' they experience in work or home life which negatively impact their subjective well-being, and thus their health, productivity, or job satisfaction. The application provides users with a historical overview of these collected self-reports and an opportunity to safely share their data with their team, generating an aggregated and anonymised well-being data stream (group-level summary statistics and trends) that is visible on shared team and company-level dashboards.

To improve our understanding of the long-term use of well-being applications truly 'in the wild' and outside the context of a controlled academic study, we present an analysis of one year of application use data. We deployed the application to two organizations, one in the recruitment services industry and the other in the healthcare services industry. Our analysis spans a total of 219 users, is informed by prior research findings within the HCI community, and allows us to quantify changes in application use, well-being, and configuration settings in a longitudinal setting. Concretely, our work aims to answer the following research questions:

- How does usage and non-usage behaviour change over time?
- How do self-reports of both well-being dimensions and challenges fluctuate as the result of temporal aspects such as time of day or day of week?
- Using a long-term data set, can we identify new insights as to what types of usage behaviour may predict drop out vs. long-term use?

Our analysis shows that the steepest drop in usage over an entire year of use is already found in the first week of use, totalling 24.2% of our entire user sample. In contrast to prior work analysing shorter usage periods, we find that users do return to use the application even after week-long breaks in application use. We compare and contrast our results with earlier findings on *e.g.*, dropout rates, collection of self-report data, and well-being evaluations, contributing to the HCI literature by demonstrating the new insights that can be obtained by taking a long-term view on user behaviour.

2 Background

A recent review by Sanches et al. highlights the increasing interest within the HCI community towards the study and support of affective health [43]. Workshops organised in relation to the topic of mental health and well-being, for example at DIS 2020 [44], highlight that many questions concerning the methods and tools to study well-being are still unanswered. We motivate our work and analysis by building on prior work focused on subjective well-being measurement and understanding as the research goal and longitudinal research as a method and form of data collection.

2.1 Collecting Longitudinal Well-Being Insights

The HCI community has long shown an interest in studying user experience and other aspects of interaction with technology over longer periods of time. Wilson highlights the challenges often encountered in longitudinal research, "*Longitudinal studies, which involve repeated sessions spread out over multiple days [...] are expensive and time-consuming. As a result, many interactive sensing-based systems come with few convincing quantitative user studies that prove their utility.*" [50]. One of the earliest reviews on longitudinal HCI research analysed the

research methods employed in the mobile HCI community [26]. Among the 144 published research papers considered, 79% of them used some form of field study as a validation scheme. However, Kjeldskov et al. comment that these studies lack the depth and duration to provide an accurate picture of mobile devices and their relationships to real-world scenarios. As aforementioned, a recent review of longitudinal studies published at CHI found only 56 studies that extend over one month in duration.

A number of studies have tried to overcome the challenges of conducting longitudinal studies in the use of self-tracking experience data for well-being research. For example, Isaacs et al. demonstrate a smartphone application called *Echo* which supports technology-mediated reflection, studying usage for durations of both one month and four years [20]. Using the data from the initial one month, Issac et al. showed that reflections improved actions and lessons learnt for future behaviour. Using data collected over four years, the researchers discovered new insights, including the surprising attenuation of affect bias (*i.e.*, observing that negative emotions associated with past reflective moments faded quicker than positive emotions). The 'Tesserae Project' is another recent example of a longitudinal study with data collection covering up to one year [34]. In this project, participants' use of social media was logged and stored by the researchers to obtain insights into workplace performance—allowing for fully passive participation by the study participants. The authors state that the high retention rates of the study (95% of participants are still enrolled over halfway through the study duration), are due to strict privacy efforts, compensation schemes, unobtrusive nature of the data collection, and the perceived potential value of the study outcomes for the participants [34]. In contrast to the passive collection of data found in the Tesserae Project, we are specifically interested in measuring subjective well-being of individuals in work which, by definition, requires collecting active data contributions. Active data contributions, as for example found in experience sampling-based studies in which participants provide multiple contributions throughout the day [2,30], allow for the collection of insights into the participants' experiences, thoughts, and feelings in a way that is not possible to capture through researcher observation, wearable data collection, or passive monitoring. Furthermore, the literature suggests that the moment of reflection generated by active tracking can provide benefits: prompting the individual to more clearly attend to and evaluate their in-work experiences, leading to the potential for performance-improving and self-corrective behaviour off the back of immediate insight generated by the act of reflection itself [45].

One prominent example of such a study is the 'StudentLife' study [48]. In this study, a total of 48 participants regularly reported on various aspects of their life (including dimensions of experienced well-being) over ten weeks. Combined with passive sensing on mobile devices, the researchers uncover several correlations between the active and passive sensing data. For example, sleep duration, conversation frequency, and co-location with others all correlate significantly with the PHQ-9 depression scale. Furthermore, the authors highlight how the active data contributions on participant mood and other variables provide insight into

a trend among participants, with positive affect and low stress levels at the start of semester followed by increased stress and a drop in positive affect towards the end of the semester due to increased workload [48]. Such insights can only be uncovered through deployments that cover an extensive period. Another example is found in the work by Fritz et al., who study long term usage of activity-tracking devices in the context of exercise [14]. Through a study involving 30 participants, the researchers investigated usage behaviour across periods ranging from 3 to 54 months. Results highlighted how obtaining a long streak of tracking exercise positively affected participants' experienced well-being and acted as an accountability tool in completing exercises and earning digital rewards through gamification. Finally, within the domain of digital health Yeager et al. present a two-week study of an internet intervention for trauma recovery [51]. While the intervention was found to be effective across a two week period, the researchers highlight the absence of longitudinal studies as a limiting factor in designing and evaluating the effect of self-guided intervention during patient relapse.

Within the broader HCI community, various researchers have explicitly expressed the need for longitudinal studies, arguing that short deployments do not provide insight into long-term engagement, data patterns and real-world outcomes, and possibly conceal a novelty effect driving initial engagement with new technologies [17,24]. Following this call for more longitudinal studies and deeper insights into the drivers of technology usage and patterns in user reported data over time [24,27], we set out to analyse usage behaviour and user subjective (experienced) well-being data as collected by a self-report application deployed in a real-world context.

2.2 Self-tracking Applications

Self-tracking, the process of "*turning everyday experience into data*" [37], has expanded from physical activity tracking (*e.g.*, step counters) towards the goal of capturing and presenting a holistic overview of a person's physical condition, mental state, or overall well-being. Those using Quantified Self applications primarily use self-tracking to gain knowledge about themselves, such as their behaviours, affect patterns, and habits, which allows them to derive meaningful insights to incorporate positive changes [4,10]. Schön distinguishes between 'reflection in action', which takes place during the activity, and 'reflection on action', which occurs after the activity with the user analysing the captured data [45]. With reference to the domain of HCI, self-report applications typically focus on 'reflection on action' by providing users with the possibility to observe changes over time.

Whether as a study participant or to collect data for one's own interests, adherence to self-tracking is a widely recognised challenge. Several papers provide suggestions to reduce end-user strain in logging data, with the goal of reducing dropout rates. For example, Zhang et al. present 'unlock journaling', in which users can log mood data while unlocking their smartphone [52]. An evaluation of this system highlights that the presented technique of unlock journaling was experienced as less intrusive than a traditional notification-driven approach while

resulting in a higher frequency of data journaling. A recent study by Cherubini et al. on daily step counts showed a detrimental effect on long term engagement after notification frequency and monetary compensation was increased, as participants' intrinsic motivation decreased as a result [9]. Another method of integrating self-tracking into everyday life can rely on the principle of affordances (properties of objects which show users the actions they can take) that users have developed during their lifetime to make the data acquisition seamless. Rapp and Cena suggest that everyday physical objects (*e.g.*, bracelets, t-shirts) are highly suitable for self-tracking as they integrate easily into the day-to-day activities of the wearer [40]. Kim et al. present a self-tracking application that allows users to more easily interact with visualisations of their self-tracking data in order to make the collection of data itself more meaningful [25].

With most affect tracking applications focused on the individual, the study of affect tracking at a group level has been relatively sparse. A notable exception to this research gap is the work by Pelayo et al. [42]. In their study, Pelayo et al. collaboratively engaged with call centres to assess whether mood self-tracking can improve work performance, emotional awareness, and team communication. The authors conducted a four-week study in which participants collected mood data through a desktop application. Pelayo et al. identify that participants generally had a positive attitude towards using the application, primarily based on the hope that interpretation of the data by the user's management will improve the working conditions.

Among the spectrum of self-tracking applications, long-term usage is a particularly common problem. Blandford, in discussing the challenges and opportunities in HCI for health and well-being, suggests that meaningful steps forward can be made by collaborating with technology developers in evaluating technologies [6]. Such an approach can empower users to reflect on their data and eventually improve data quality and reduce user effort to acquire self tracking data [36]. In this paper, we set out to do precisely this. Through a collaboration with a well-being measurement start-up, we generate a year-long stream of data comprised of individual subjective well-being self-reports as well as usage data. This allows us to expand the HCI literature with a substantially longer-than-average study length, from which we derive novel insights into patterns in individual well-being as well as into the patterns and drivers of long-term user adherence to self-report applications.

3 Method

Our analysis is based on a collaboration with *Anonymised Company*. *Anonymised Company* offers an application for regular assessments of experienced affective well-being across a company by enabling individual employees to report their experienced well-being, showing individual employees their own data and insights, and offering aggregate team- and company-level insights through shared dashboards. Users of the application can report their well-being levels (detailed below, see Fig. 1), anonymously capture and share specific 'blockers' that may

stagnate their progress at work, or anonymously share ideas with their team or company management (Fig. 2). Blockers can represent a variety of issues that can arise at a workplace or in an individual's personal life, such as work relationships (e.g., difficulties in interacting with management), work environment (e.g., interruptions, air quality), or struggles with health, sleep, or exercise. In this analysis we focus on engagement with the application to actively input self-reports, the reported well-being and blockers data itself, as well as a potential moderator of short and long-term usage. *Anonymised Company* offers the same features through both a mobile and web-based application that allow for the exact same data collection input. However, given the recent release of the mobile application, the large majority of the collected data originates from web-based input. We therefore solely focus on the analysis of data from the web-based application.

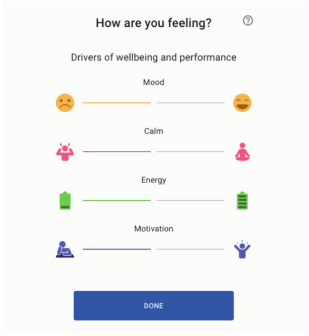

Fig. 1. Self-report component of the online application.

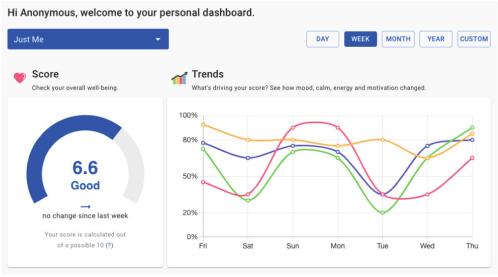

Fig. 2. Dashboard, displaying changes in self-reported well-being values over time (day, week, month, year).

3.1 Participant Recruitment

The target population for our study was the entire population of two organizations, including all employees, managers, and leadership of the firms. All participants were sent email-based invitations inviting them to sign-up, along with contextual information as to the purpose and potential benefit for them, along with information designed to inform individuals of how their data would or would not be used, and the expected time-cost of the self-report 'check-in'. The sample assessed in this paper reflects the characteristics of the individuals who received the email and decided to download and subsequently use the application at least once.

We introduced the application to each organization through a predesigned roll-out process. This process started with a kick-off and induction for the leadership team, followed by the individuals in each team, consisting of specific training on the overarching problem (*i.e.*, poor or varied employee well-being

that impacts performance at work), the application's well-being measurement concept, the features of the application, and how the data would be analysed and used. Emphasis was placed on the collective and participatory nature of the well-being tracking for the joint benefit of both individuals and the group. Furthermore, we highlighted the fact that the individual user was in control of how their data was shared, something reinforced by the users' ability to see their data aggregates on the team and company dashboards. In all cases it was stressed to users that participation was voluntary, and they were free to cease usage if they wished to do so. In addition, in both organization A and B, team leaders and company leaders were trained to use the dashboards to identify how well-being has changed, respond to blockers submitted by employees within the application, and use the data and insights to inform their organisational strategies.

Participants were not instructed or required to complete any minimum number of self-reports, and were not compensated for their participation, thus limiting the bias that could come from extrinsic incentives or a sense of obligation. Within the application, participants could provide either a personal well-being check-in, one or multiple blockers, or a combination of both.

Participant Anonymity. Our approach to participant anonymity was developed in consultation with early platform users during beta tests (prior to the study period). Due to the anonymous nature of the platform's data collection and data use, individuals were not asked to provide information about their age, gender, work experience, prior use of well-being tracking or technology, self-efficacy, organisational roles, or other personal information. Further, the leaders of the organizations in which the application is deployed were kept unaware of which individuals use or do not use the application. Critically, team dashboards were accessible to both team leaders as well as all members of the team. By opening up the team dashboards to every member of the team, we sought to head-on address issues and concerns with anonymity. By viewing the dashboards themselves, individuals could see what was meant by the analysis of the team data and presentation 'on average and in aggregate', and also see where their blockers and ideas feedback appeared, and therefore make their own judgement about whether the structure of these features in the app afforded them sufficient flexibility in anonymity. Our study aligns with our University's regulation on research ethics and data management.

3.2 Dataset

The specific variables represented in the dataset are as follows;

– **Well-being check-ins**. Self-reported experiences on the four dimensions 'Mood', 'Calm', 'Energy', and 'Motivation', as captured through four horizontal sliders in response to the question 'How are you feeling?'. The interface allows the user to independently capture positive versus negative states, with a single line capturing the conjoin of a 0–50 positive affect scale (neutral

to the right) and an 0–50 negative affect scale (neutral to the left). For all sliders, the default position is in the middle 'neutral' state.
- **Blockers**. Self-reported issues negatively impacting well-being. Categorised as either personal or work related.
- **Timestamp of self-reports**. For each self-report on well-being or blockers, a timestamp is recorded. Reporting of well-being check-ins and blockers can be completed independently from one another.
- **Reminder notification configuration**. The application offers users the ability to send up to one reminder email per day at a timepoint set by the user. The email consists of one question 'How did it go today?' and suggest to 'check in', which takes the user directly to the self-report screen in the application. Further, users are able to select the days at which they receive a reminder—ranging anywhere from zero (no reminders at all) to seven days.

In order to allow for a longitudinal analysis of the presented results, we exclude any user whose account was created less than one year before the last available data dump (July 2021). Following this exclusion criteria, we conduct our analysis on the first year of application usage of the remaining total of 219 individual users.

4 Results

Our analysis focuses on an assessment on an analysis of long-term usage behaviour, a review of the self-reported well-being check-ins and blockers, and an investigation into user preferences into reminder notifications and their effect on completed self-reports. In order to ensure comparability in our analysis, we solely consider the first year of data that we have available on each user of the application. As our dataset consists of users that commence usage of the application at different times, we ensure a relative calculation of usage duration (*i.e.*, we consider a user's first day of use as the starting point of application usage). This analysed dataset consists of a total of 2968 completed self-reports and 515 blockers. The data are provided by a total of 219 unique users, 192 of which work at organization A and 27 users which work at Organization B.

4.1 Application Usage

We first present the overall drop-off rate among users over a one year period in Fig. 3. Here, we define the drop-off week as the last week in which a check-in (*i.e.*, a full well-being report) is provided following their initial check-in. Overall, we find the steepest drop-off of users within the first week of usage (dropping 24.2% of users). A total of 43 users, 19.6%, only provide one check-in. The number of users dropping out increases throughout the year, with the increase in dropouts becoming more gradual following the initial month. The distribution of the users' check-ins over time is shown in Fig. 4 (day of the week) and Fig. 9 (time of day). The fact that most check-ins are on weekdays rather than weekends reflects the

strong focus of the application on the workplace context. Assessing the time of check-in, we observe that most users check in at the end of their (work)day or later in the evening.

We furthermore assess whether and how usage of the application changes over time. Figure 5 shows the average number of check-ins made and blockers reported among active users of the application per week. This highlights that the average number of weekly check-ins decreases rapidly from an average of 1.7 in the first week of use to 1.0 in one month of usage, after which the decrease slows down and stabilises around 0.3 check-ins per week. The visualisation of Fig. 5 furthermore highlights that the most long-term users (over 40 weeks of use) provide a higher average of check-ins per week, hovering around 0.75.

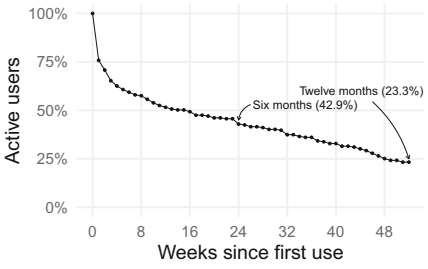

Fig. 3. Overall drop-off of rate across a one-year period.

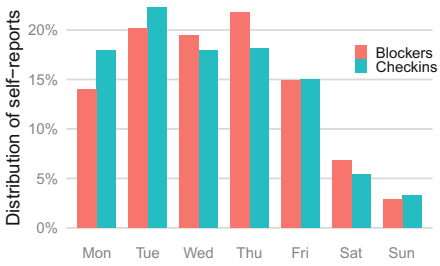

Fig. 4. Distribution of check-ins and blockers across the week.

To better understand usage differences between short-term and long-term users of the application, we divide users into quartiles as based on their dropout week. We report these characteristics in Table 1. Our overview on these usage characteristics highlights that users belonging to Q4 (those with the longest usage duration) not only have the longest usage duration but also provide the highest number of check-ins (an average of 30 check-ins over the analysed usage duration). The information on median breaks among users (*i.e.*, days in-between two provided self-reports) highlights that extended breaks in usage of the application are common. Users with the longest usage duration have a median break in usage duration of three weeks.

Breaks in application usage may be indicative of a user quitting an application. Being able to predict which user will return to an application would be worthwhile information for application developers, as users can be reminded or incentivised to continue their usage of an application. Therefore, we next assess whether a relationship exists between gaps in usage duration and a user's final application usage. Figure 6 shows the distribution of usage gaps in the days leading up to a subsequent usage of the application (*i.e.*, non-final application usage) or the final usage of the application. We find a median value of a two-day break for a check-in that is eventually followed by a next check-in. For a check-in

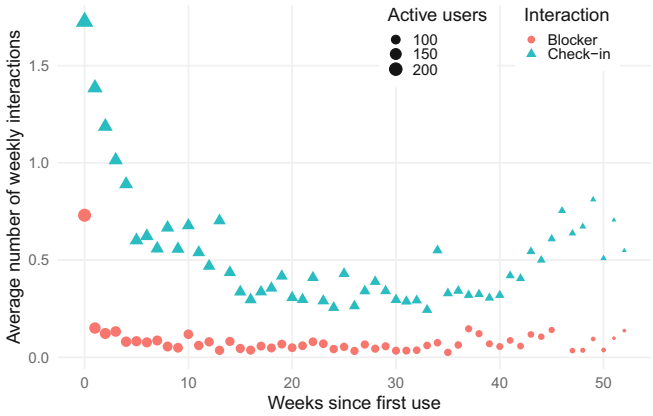

Fig. 5. Average number of interactions among active users.

Table 1. Usage characteristics as grouped by quartiles of usage duration.

Quartile	Weeks till dropout	Avg no. check-ins	Median break (days)
Q1	0–0	1.4 (SD = 0.6)	0
Q2	0–10	5.9 (SD = 4.7)	4.5
Q3	11–41	17.1 (SD = 14.5)	13.7
Q4	43–52>	30.1 (SD = 32.7)	21.1
All	0–52>	13.6 (SD = 21.0)	6.5

which is not followed by any further usage (*i.e.*, the final check-in), we find a median break in usage of five days leading up to this final check-in. A Mann Whitney U (non-parametric equivalent to the two sample t-test) indicates that the distributions between the two groups differed significantly (Mann-Whitney U = 248312, n1 = 2749, n2 = 219, $p < 0.01$ two-tailed). As indicated in Fig. 6, after a break in participant usage of eight days, the likelihood of the participant quitting usage of the application is larger than future usage of the application.

4.2 Well-Being and Challenges

On each check-in, users reported their mood, calmness, energy, and motivation levels on a scale from −50–50 (see Fig. 1), translated in this paper to 0–100 for analysis and reporting purposes. Assessing fluctuations in participant well-being, we identify the start of the working day (08:00–12:00) and end of the working day (16:00–18:00) as the timeslots with the most positive well-being check-ins reported. Well-being check-ins are consistently lowest in the early afternoon (13:00–15:00) and dip again toward the late evening (19:00–22:00).

Next, we analyse the blockers as reported by application users. Across the dataset, a total of 515 blockers were reported. On average, users report one

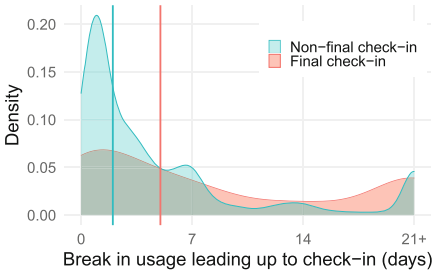

Fig. 6. Gaps in usage prior to non-final and final check-ins. Vertical lines indicate median breaks in usage.

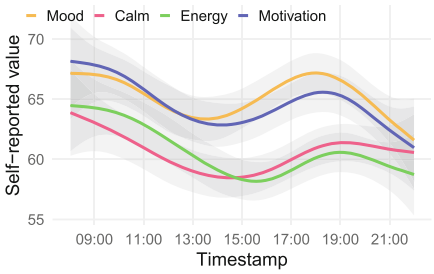

Fig. 7. Fluctuations in well-being values throughout the day.

blocker for every 5.8 well-being reports. Assessed by category, roughly 60% of the blockers are work-related blockers (314, of which 180 on work and resources and 70 on work processes, among smaller categories) and 40% of blockers are of personal nature (201 blockers). Figure 4 shows the frequency of blockers per day of the week. Unsurprisingly, and in line with the well-being reports, we find that users are reporting the fewest number of blockers on the weekends. Across the working days, we find the least number of blockers to be reported on Monday.

4.3 Reminder Notifications

As a final element of analysis, we assess user preferences towards reminder notifications and their effect on completed self-reports. The application provides users the opportunity to change the default notification schedule, which is set to trigger at 08:00 at every workday (Mon–Fri). Our dataset contains information solely on the user's final configuration, *i.e.* any changes throughout use are not considered in this description. The majority of users change away from the default time of notification (77.6%). However, significantly fewer users change the default days of notification (37.0%). In changing the reminder notification time, participants primarily shift to a later time in the day, either towards the end of the working day or later in the evening (see Fig. 8 for an overview). In their preference of notification days, next to the default option of Mon–Fri (63.0%), the most popular option is to disable notifications altogether (23.3%), followed by a long tail of combinations with fewer notification days. Contrasting the time at which reminder notifications arrive (Fig. 8) with the time of check-ins being completed (Fig. 9), we observe a temporal correlation between the reminders and the actual check-ins—especially when disregarding the default reminder time.

Perhaps the most critical in terms of reminder notifications is whether they help in the collection of self-report data. Our dataset allows us to assess both whether users without notifications provide fewer self-reports, and whether they are more prone to stop using the application altogether. We calculate the average number of submitted self-reports among those with notifications enabled and

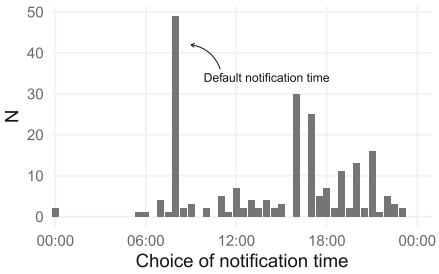

Fig. 8. Distribution of preferred reminder time.

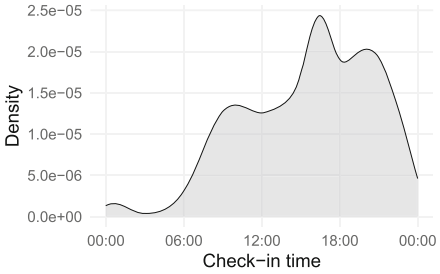

Fig. 9. Distribution of check-in times across time of day.

those with notifications disabled. Users with notifications enabled completed an average of 13.4 self-reports, with users without notifications completing an average of 14.1 self-reports. A t-test did not find a significant difference between these two groups ($t(106.39) = 0.234$, $p = 0.815$). In terms of usage duration, we find a significant difference between those with notifications enabled (average of 21.3 weeks of active usage) as compared to those with notifications disabled (an average of 13.1 weeks), $t(106.58) = 2.957$, $p = 0.004$. We visualise this difference in Fig. 10. While this highlights that participants with notifications enabled make use of the application for a significantly longer period of time, we are unable to establish a causal relationship (*i.e.*, reminder notifications could lead to sustained usage, or those that disabled notifications were simply less interested from the beginning and therefore disabled notifications early on).

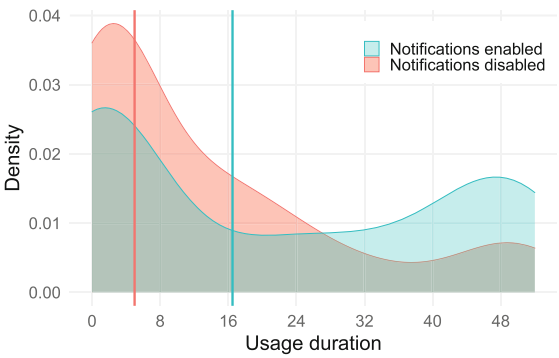

Fig. 10. Distribution of usage duration as split between those with notifications disabled and enabled.

5 Discussion

Through the presented analysis we aim to provide insights into the use of well-being oriented self-report application across the first year of usage. While our analysis does not follow a strict recruitment and usage protocol as typically seen in laboratory studies, we believe this is compensated by the ecological validity of the study. It is impossible to study these deployments in a team or organisational environment without actually deploying them in the real-world, the results of which we present in this paper. Well-being has long been a topic of interest within the broader HCI community [6], as well as one of the main topics of interest within Quantified Self research [18,28]. Our data on participants' self-reported well-being levels align closely with the existing literature. Widely reported phenomena of shifts in well-being, including high well-being values in the morning, followed by a dip in the early afternoon, and a subsequent increase after working-hours [15], are closely matched (see Fig. 7). Our findings further indicate a close alignment between the self-reported well-being variables of mood, calmness, energy, and motivation when aggregated across users, albeit with different intra-daily fluctuations in these. Given the large overlap in construct measured, the collection of these four well-being parameters may be an unnecessary burden to participants. Especially in the light of widely reported participant strain in Experience Sampling-based studies, potentially resulting in fewer completed self-reports or earlier dropout rates [3,12].

5.1 Considering Dropout

Within the HCI community, and among Quantified Self researchers in particular, attrition of users (or participants in the context of a study) is a well-known phenomenon [1,21,38]. For example, Gouveia et al. found that only 14% of the users of their app-based activity tracker used the application for longer than two weeks, despite implementing a number of literature-based design strategies to promote long-term engagement [16]. Such low numbers are no outliers, with application usage often being highly exploratory and quickly removed once installed [39]. Comparison of dropout rates between different deployments is, however, challenging, as a number of different factors are likely to influence participants' usage commitments. We provide three possible explanations for our relatively high level of observed engagement over a long period of time.

First, our analysis is based on a deployment within the context of two organizations and embedded into their infrastructure. The analysed application allows teams to assess and monitor their well-being levels anonymously and in aggregate, share blockers and ideas with their teams. This sets the application apart from other well-being tracking applications, as the end user (*i.e.*, employee) is aware that their contribution is not solely of use to themselves but might be picked up as part of a larger trend within the organisation. This highlights the potential positive impact on engagement rates of operationalising (well-being) self-reports beyond the scale of the individual. Along this line, Berrocal et al. studied 'Peer-ceived Momentary Assessment', in which an individual's

self-reports are further complemented with peer-reports as provided by family members or friends [5]. The findings from Berrocal et al.'s study highlight how the assessments provided by peers align with self-reports across a number of constructs (observable behaviours and states). The involvement of peers, especially in an inherently hierarchical setting such as the workplace, may add additional pressure on individuals to join in participating in active self-reported data collection. While the current study focused solely on well-being reports as perceived by the user, future work may explore how the collection of well-being in the context of work can affect workplace dynamics and support the optimisation of work for experienced well-being.

Second, in addition to collecting self-report data, our application also provides users with a dashboard through which they can observe changes in their well-being reports across different timespans (day, week, month, year). This allows users to generate personal value from the application over the long term, by deriving their own insights. Furthermore, the application provides a way for users to respond to blockers and ideas shared within the app, which may lead to user expectations that self-reporting will generate a novel response. This aligns with earlier recommendations by Gouveia et al. to maintain engagement with users [16]. While Gouveia et al.'s work focused primarily on the use of activity trackers, their suggestion to sustain engagement by providing a novel message in response may be applicable also in the scenario of self-report studies.

Third, our evaluation indicates that users can and do return to using the application even after extended periods of dropout. This behaviour is not captured in deployments restrained to a shorter time period.

5.2 Check-in Settings and Behaviour

While our results show that having notifications enabled did not result in more completed self-reports, they do highlight that those with notifications enabled are more likely to use the application for a longer period of time. Prior work suggests that users rarely change the default notification settings of applications [49], although it does not provide any concrete numbers on how many users do change away from system defaults. Our results highlight that, although the default notification time is the most popular time for notifications (see Fig. 8), a large majority of 78% of our users did change away from the default notification time. Contrasting the insights on default notification time in Fig. 8 with the distribution of check-in times across the day in Fig. 9 furthermore highlights that while 08:00 is the most common time for notifications to arrive, it is not the most common time for users to provide their self-reports. Instead, we see an obvious temporal correlation between participant check-in time and their choice of notification time among those that have adjusted the default notification time. This indicates that, at least for studies in which one daily notification is common (*e.g.*, diary studies), allowing participants to customise the time of incoming notification could increase participant responsiveness. While this increased flexibility in study design was already suggested in 2003 by Consolvo and Walker [11], it is

still rather uncommon to allow study participants to choose their time of notification in diary/Experience Sampling Method studies. Whether or not providing a default notification time in the first place, as compared to forcing users to pick a time of their own, affects subsequent response rate and response time to notifications is an interesting avenue for future work.

Our results highlight that active users provide less than one well-being self-report on average per week. The reporting of blockers is rare, with blockers making up close to 15% of the entire dataset. How do we interpret these numbers in light of the aforementioned discussion on attrition and dropout rates? A relevant study on well-being in the workplace was conducted by Pelayo et al. [42], who investigated mood self-tracking in call centres over four weeks across 71 participants. Their results highlight an average of 17 moods reported per team member, or just over four per week. The setup of this study relied on regular meetings with team managers and coaches to motivate and support managers and coaches of the call centre teams in the evaluation. This stands in stark contrast with our study, in which we minimised research contact with the application's users and therefore obtain a better understanding of the users' natural behaviour and interaction with the application.

5.3 Implications for Research

By investigating the long-term usage of a self-report application deployed in a real-world company environment, we uncovered novel insights into using self-report applications. Several of our findings contradict existing work or raise new questions regarding the use of self-tracking applications. Simultaneously, we acknowledge the limitations of interpreting results from individual deployments. As such, we present the following implications primarily as an opportunity for future research to better inform researchers and practitioners of real-world application usage.

Long-term Deployments Reveal Varying Usage Patterns. Contrary to our expectations, we found that users that had seemingly given up on using the application returned to use the application after relatively long usage breaks (up to three weeks for the top 25% of long-term users – Table 1). This behaviour would not be captured during typical study deployments and may require the research community to revisit some of the existing beliefs regarding long-term application usage. One factor that may cause such behaviour is that our users are surrounded by individuals who still retain the use of the tracking application, thus reminding of the option to continue their self-tracking.

Usage Patterns may Indicate Future Application Abandonment. Our results highlight a significant difference in the length of usage breaks between continuing and final application usage, indicating that the length of usage breaks can be used to predict future abandonment (see Fig. 6). While additional investigation of this phenomenon is required to confirm the generalisability of this

finding, the possible implications are numerous. Applications within ubiquitous computing, Quantified Self, and self-report studies, among others, could directly benefit from knowing when users may quit – providing the opportunity to incentivise continued application usage or support alternative forms of usage [47]. We, therefore, consider further identification and the possible prediction of future application abandonment as a valuable opportunity for the research community.

Industry Collaboration can Produce Rich Longitudinal Insights. While *in situ* evaluations are often recognised for their high external validity, they have also been described as "*costly and often technically challenging*" [23]. This may, in part, explain the relatively low number of longitudinal deployments that have been conducted within our research community [27]. While not a silver bullet to resolve these challenges, we highlight that collaboration with industry provides opportunities for data collection that overcome some of the constraints academic researchers face. Furthermore, we stress that such collaborations should not impact researcher impartiality. The limitations of assessing such datasets, such as sample bias, should be explicit in the published work.

5.4 Limitations

We highlight several limitations in our work that should be considered when interpreting the presented results. First, our analysis focuses on one specific application and may therefore not generalise to other well-being oriented self-report applications. While the primary elements of the application closely align with existing applications in Quantified Self and self-report research (*e.g.*, reminders, logbook of past activities), individual aspects of an application may significantly affect end-user motivation and willingness to provide data input [16]. Second, users of the application may have left the organisation during data collection. As such, they have stopped using the application as a result of their employment situation rather than other factors that might affect application usage. We are unable to incorporate this aspect into our analysis due to the anonymity of user data. While we expect the effect to be relatively limited, this may have resulted in a slight overestimation of the users' dropout rate. Finally, our analysis is focused on the deployment across two individual organizations. Given this limited sample of organizations, we are unable to assess the effect that different organisational cultures or leadership advocacy patterns have on the usage of such an application. We suggest this as a potential avenue for future research.

6 Conclusion

We present an analysis of one year of well-being self-report data. This longitudinal analysis, involving over two hundred individual users, provides insights into real-world usage behaviour. Our findings highlight that while dropout rates of self-report applications are high, we also find that, contrary to prior work, a

short-term break in usage does not necessarily signal a dropout, with a cohort of users returning to the application in the long term after extended intervals in use. In terms of well-being values, we find the application's well-being reports are consistent with prior findings regarding the changing patterns of well-being self-reports during the day at various times. Regarding notifications, we find that a large majority of users change their default notification time and that those with notifications enabled have a significantly longer average usage duration compared to those with notifications disabled. Based on our findings, we highlight recommendations for future studies involving active participant contributions, including the use of contextual feedback features to increase long-term engagement rates; the impact of presenting not only individual but also the self-report data of peers; and the effect of organisational culture on long-term self-report behaviour and retention.

References

1. Attig, C., Franke, T.: Abandonment of personal quantification: A review and empirical study investigating reasons for wearable activity tracking attrition. Comput. Hum. Behav. **102**, 223–237 (2020). https://doi.org/10.1016/j.chb.2019.08.025
2. van Berkel, N., Ferreira, D., Kostakos, V.: The experience sampling method on mobile devices. ACM Comput. Surv. **50**(6), 1–40 (2017). https://doi.org/10.1145/3123988
3. van Berkel, N., Kostakos, V.: Recommendations for conducting longitudinal experience sampling studies. In: Karapanos, E., Gerken, J., Kjeldskov, J., Skov, M.B. (eds.) Advances in Longitudinal HCI Research. HIS, pp. 59–78. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67322-2_4
4. van Berkel, N., Luo, C., Ferreira, D., Goncalves, J., Kostakos, V.: The curse of quantified-self: an endless quest for answers. In: Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers, pp. 973–978. UbiComp/ISWC 2015 Adjunct (2015). https://doi.org/10.1145/2800835.2800946
5. Berrocal, A., Concepcion, W., De Dominicis, S., Wac, K.: Complementing human behavior assessment by leveraging personal ubiquitous devices and social links: an evaluation of the peer-ceived momentary assessment method. JMIR Mhealth Uhealth **8**(8), e15947 (2020). https://doi.org/10.2196/15947
6. Blandford, A.: HCI for health and wellbeing: challenges and opportunities. Int. J. Hum Comput Stud. **131**, 41–51 (2019). https://doi.org/10.1016/j.ijhcs.2019.06.007
7. Caraban, A., Karapanos, E., Gonçalves, D., Campos, P.: 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. In: Proceeding of the 2019 CHI Conference, pp. 1–15 (2019). https://doi.org/10.1145/3290605.3300733
8. Chalhoub, G., Kraemer, M.J., Nthala, N., Flechais, I.: "it did not give me an option to decline": a longitudinal analysis of the user experience of security and privacy in smart home products. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (2021). https://doi.org/10.1145/3411764.3445691

9. Cherubini, M., Villalobos-Zuñiga, G., Boldi, M.O., Bonazzi, R.: The unexpected downside of paying or sending messages to people to make them walk: Comparing tangible rewards and motivational messages to improve physical activity. ACM Trans. Comput.-Hum. Interact. **27**(2), 1–44 (2020). https://doi.org/10.1145/3365665
10. Choe, E.K., Lee, N.B., Lee, B., Pratt, W., Kientz, J.A.: Understanding quantified-selfers' practices in collecting and exploring personal data. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1143–1152 (2014). https://doi.org/10.1145/2556288.2557372
11. Consolvo, S., Walker, M.: Using the experience sampling method to evaluate ubicomp applications. IEEE Pervasive Comput. **2**(2), 24–31 (2003). https://doi.org/10.1109/MPRV.2003.1203750
12. Eisele, G., et al.: The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. Assessment (2020). https://doi.org/10.1177/1073191120957102
13. Fowler, F.: Survey Research Methods. Thousand Oaks, California, 4th edn. (2009). https://doi.org/10.4135/9781452230184
14. Fritz, T., Huang, E.M., Murphy, G.C., Zimmermann, T.: Persuasive technology in the real world: A study of long-term use of activity sensing devices for fitness. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 487–496. CHI 2014 (2014). https://doi.org/10.1145/2556288.2557383
15. Golder, S.A., Macy, M.W.: Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science **333**(6051), 1878–1881 (2011). https://doi.org/10.1126/science.1202775
16. Gouveia, R., Karapanos, E., Hassenzahl, M.: How do we engage with activity trackers? a longitudinal study of habito. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 1305–1316. UbiComp 2015 (2015). https://doi.org/10.1145/2750858.2804290
17. Grudin, J.: Why CSCW applications fail: Problems in the design and evaluation of organizational interfaces. In: Proceedings of the 1988 ACM Conference on Computer-Supported Cooperative Work, pp. 85–93. CSCW 1988 (1988). https://doi.org/10.1145/62266.62273
18. Hollis, V., Konrad, A., Whittaker, S.: Change of Heart: Emotion Tracking to Promote Behavior Change. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 2643–2652 (2015). https://doi.org/10.1145/2702123.2702196
19. Horvitz, E., Koch, P., Apacible, J.: Busybody: Creating and fielding personalized models of the cost of interruption. In: Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work, pp. 507–510. CSCW 2004 (2004). https://doi.org/10.1145/1031607.1031690
20. Isaacs, E., Konrad, A., Walendowski, A., Lennig, T., Hollis, V., Whittaker, S.: Echoes from the past: How technology mediated reflection improves well-being. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1071–1080. CHI 2013 (2013). https://doi.org/10.1145/2470654.2466137
21. Jun, E., Hsieh, G., Reinecke, K.: Types of motivation affect study selection, attention, and dropouts in online experiments. Proc. ACM Hum.-Comput. Interact. **1**(CSCW), 1–15 (2017). https://doi.org/10.1145/3134691
22. Kahneman, D., Krueger, A.B.: Developments in the measurement of subjective well-being. J. Econ. Perspect. **20**(1), 3–24 (2006). https://doi.org/10.1257/089533006776526030

23. Kaptein, M.: Experiments, longitudinal studies, and sequential experimentation: how using "intermediate" results can help design experiments. In: Karapanos, E., Gerken, J., Kjeldskov, J., Skov, M.B. (eds.) Advances in Longitudinal HCI Research. HIS, pp. 121–149. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67322-2_7
24. Karapanos, E., Zimmerman, J., Forlizzi, J., Martens, J.B.: User experience over time: An initial framework. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 729–738 (2009). https://doi.org/10.1145/1518701.1518814
25. Kim, Y.H., Lee, B., Srinivasan, A., Choe, E.K.: Data@hand: Fostering visual exploration of personal data on smartphones leveraging speech and touch interaction. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI 2021 (2021). https://doi.org/10.1145/3411764.3445421
26. Kjeldskov, J., Paay, J.: A longitudinal review of mobile HCI research methods. In: Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 69–78 (2012). https://doi.org/10.1145/2371574.2371586
27. Kjærup, M., Skov, M.B., Nielsen, P.A., Kjeldskov, J., Gerken, J., Reiterer, H.: Longitudinal studies in HCI research: a review of CHI publications from 1982–2019. In: Karapanos, E., Gerken, J., Kjeldskov, J., Skov, M.B. (eds.) Advances in Longitudinal HCI Research. HIS, pp. 11–39. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67322-2_2
28. Klasnja, P., Consolvo, S., Pratt, W.: How to Evaluate Technologies for Health Behavior Change in HCI Research. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 3063–3072 (2011). https://doi.org/10.1145/1978942.1979396
29. Kuosmanen, E., et al.: How does sleep tracking influence your life? experiences from a longitudinal field study with a wearable ring. Proc. ACM Hum.-Comput. Interact. **6**(MHCI), 1–19 (2022). https://doi.org/10.1145/3546720
30. Larson, R., Csikszentmihalyi, M.: The experience sampling method. In: Flow and the Foundations of Positive Psychology, pp. 21–34. Springer, Dordrecht (2014). https://doi.org/10.1007/978-94-017-9088-8_2
31. Lee, J.H., Schroeder, J., Epstein, D.A.: Understanding and supporting self-tracking app selection. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **5**(4), 1–25 (2021). https://doi.org/10.1145/3494980
32. Lukoff, K., Yu, C., Kientz, J., Hiniker, A.: What makes smartphone use meaningful or meaningless? Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **2**(1), 1–26 (2018). https://doi.org/10.1145/3191754
33. Lupton, D.: The Quantified Self. Wiley (2016)
34. Mattingly, S.M., et al.: The tesserae project: Large-scale, longitudinal, in situ, multimodal sensing of information workers. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–8 (2019). https://doi.org/10.1145/3290607.3299041
35. McCambridge, J., Kypri, K., Elbourne, D.: Research participation effects: a skeleton in the methodological cupboard. J. Clin. Epidemiol. **67**(8), 845–849 (2014). https://doi.org/10.1016/j.jclinepi.2014.03.002
36. Meyer, J., Simske, S., Siek, K.A., Gurrin, C.G., Hermens, H.: Beyond quantified self: data for wellbeing. In: CHI 2014 Extended Abstracts on Human Factors in Computing Systems, pp. 95–98. CHI EA 2014 (2014). https://doi.org/10.1145/2559206.2560469

37. Neff, G., Nafus, D.: Self-tracking. MIT Press (2016)
38. Rabbi, M., Li, K., Yan, H.Y., Hall, K., Klasnja, P., Murphy, S.: ReVibe: a context-assisted evening recall approach to improve self-report adherence. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **3**(4), 1–27 (2019). https://doi.org/10.1145/3369806
39. Rahmati, A., Tossell, C., Shepard, C., Kortum, P., Zhong, L.: Exploring iphone usage: the influence of socioeconomic differences on smartphone adoption, usage and usability. In: Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 11–20. MobileHCI 2012 (2012). https://doi.org/10.1145/2371574.2371577
40. Rapp, A., Cena, F.: Affordances for self-tracking wearable devices. In: Proceedings of the 2015 ACM International Symposium on Wearable Computers, pp. 141–142. ISWC 2015 (2015). https://doi.org/10.1145/2802083.2802090
41. Rieman, J.: The diary study: A workplace-oriented research tool to guide laboratory efforts. In: Proceedings of the INTERACT 1993 and CHI 1993 Conference on Human Factors in Computing Systems, pp. 321–326. CHI 1993 (1993). https://doi.org/10.1145/169059.169255
42. Rivera-Pelayo, V., Fessl, A., Müller, L., Pammer, V.: Introducing mood self-tracking at work: empirical insights from call centers. ACM Trans. Comput.-Hum. Interact. **24**(1), 1–28 (2017). https://doi.org/10.1145/3014058
43. Sanches, P., et al.: HCI and Affective Health: Taking Stock of a Decade of Studies and Charting Future Research Directions. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2019), pp. 1–17 (2019). https://doi.org/10.1145/3290605.3300475
44. Sas, C., Höök, K., Doherty, G., Sanches, P., Leufkens, T., Westerink, J.: Mental wellbeing: future agenda drawing from design, HCI and big data. In: Companion Publication of the 2020 ACM Designing Interactive Systems Conference, pp. 425–428. DIS 2020 Companion (2020). https://doi.org/10.1145/3393914.3395920
45. Schön, D.A.: The reflective practitioner: How professionals think in action. Routledge (2017)
46. Thieme, A., Wallace, J., Meyer, T.D., Olivier, P.: Designing for mental wellbeing: towards a more holistic approach in the treatment and prevention of mental illness. In: Proceedings of the 2015 British HCI Conference, pp. 1–10. British HCI 2015 (2015). https://doi.org/10.1145/2783446.2783586
47. Visuri, A., van Berkel, N., Goncalves, J., Rawassizadeh, R., Ferreira, D., Kostakos, V.: Understanding usage style transformation during long-term smartwatch use. Pers. Ubiquit. Comput. **25**(3), 535–549 (2021). https://doi.org/10.1007/s00779-020-01511-2
48. Wang, R., et al.: StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 3–14 (2014). https://doi.org/10.1145/2632048.2632054
49. Westermann, T., Möller, S., Wechsung, I.: Assessing the relationship between technical affinity, stress and notifications on smartphones. In: Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct, pp. 652–659. MobileHCI 2015 (2015). https://doi.org/10.1145/2786567.2793684
50. Wilson, A.: Sensor- and Recognition-Based Input for Interaction, chap. 10, pp. 153–176. CRC Press (2009). https://doi.org/10.1201/b10368-13

51. Yeager, C.M., Shoji, K., Luszczynska, A., Benight, C.C.: Engagement with a trauma recovery internet intervention explained with the health action process approach (HAPA): longitudinal study. JMIR Ment. Health **5**(2), e9449 (2018)
52. Zhang, X., Pina, L.R., Fogarty, J.: Examining unlock journaling with diaries and reminders for in situ self-report in health and wellness. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 5658–5664 (2016). https://doi.org/10.1145/2858036.2858360