Towards Improving Emotion Self-report Collection using Self-reflection

Surjya Ghosh
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
Surjya.Ghosh@cwi.nl

Bivas Mitra
IIT Kharagpur
West Bengal, India
bivas@cse.iitkgp.ac.in

Pradipta De
Georgia Southern University
Satesboro, USA
pde@georgiasouthern.edu

CHI '20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.3383019

Abstract

In an Experience Sampling Method (ESM) based emotion self-report collection study, engaging participants for a long period is challenging due to the repetitiveness of answering self-report probes. This often impacts the self-report collection as participants dropout in between or respond with arbitrary responses. *Self-reflection* (or commonly known as analyzing past activities to operate more efficiently in the future) has been effectively used to engage participants in logging physical, behavioral, or psychological data for Quantified Self (QS) studies. This motivates us to apply self-reflection to improve the emotion self-report collection procedure. We design, develop, and deploy a self-reflection interface and augment it with a smartphone keyboard-based emotion self-report collection application. The interface provides feedback to the users regarding the relation between typing behavior and self-reported emotions. We validate the proposed approach using a between-subject study, where one group (*control group*) is not exposed to the self-reflection interface and the other group (*study group*) is exposed to it. Our initial results demonstrate that using self-reflection it is possible to engage the participants in the long-term and collect more self-reports.

Author Keywords

Self-reflection; Experience Sampling Method (ESM); Human emotion; Self-report; Smartphone

CCS Concepts

•Human-centered computing → Human computer interaction (HCI); *User interface design;*

Introduction

One of the key tasks in affective computing is to develop emotion detection models. Often these models are trained using self-reported emotion labels collected from an Experience Sampling Method (ESM) based study. However, responding to a large number of self-report probes in a long-term study induces survey fatigue and disengages participants, which can compel users to skip answering the probes, to respond with arbitrary responses, or even withdraw from the studies in the middle [18, 21, 19, 25]. All of these can adversely impact the self-report quality and therefore influence the efficiency of the developed model. Hence, ways to improve engagement in emotion self-report collection studies are essential.

In the existing literature, multiple approaches are practiced to improve the self-report collection procedure. One of the commonly adopted approaches is based on devising smart ESM strategies, which identify the opportune probing moments, minimizing the user interruption. This approach leverages on different on-device sensors to infer user context and issue the probes at favorable moments, when the high-quality self-reports can be elicited [28, 9, 20, 1, 7, 10]. Additionally, attempts have been made to improve the participant retention rate and data quality based on incentives, which can be in different forms like monetary, providing community benefit, increasing reputation, rewarding participants [17, 11, 6]. Recently, different gamification strategies like unlocking higher levels of the game, rewarding points for completing tasks have been proposed to engage users during ESM driven data collection [8, 26].

Another potential approach of engaging participants (or increasing compliance) in an ESM-based study can be providing insights based on the data recorded during the study to support decision making, monitor behavior changes and so on [12, 16, 15, 5, 4]. In psychology, this process of analyzing past behavior (or actions) to operate more efficiently in the future is commonly termed as *self-reflection* [3, 14]. For example, in a daily-diet tracking application, the participants may feel to record food intake more carefully, as the application provides feedback (based on the recorded details) in terms of calorie consumption and allows them to reflect on developing better food habits. In today's world, self-reflection is also one of the key driving forces behind the quantified self movement [22], which enables users to track their own activities (or behavior) and obtain insights based on the physical, psychological and behavioral data collected over time [13, 2, 29, 27, 16]. In the same vein, in a long-term, ESM-based emotion self-report collection study, enabling users to reflect in terms of emotion variation, suitable triggers, etc. can motivate them to record high-quality emotion self-reports.

In this paper, we propose a self-reflection interface, which can be augmented with an emotion self-report collection app. This interface leverages the data collected for emotion detection and provides feedback to the participants during the study. We design and implement the interface as an Android application. We augment it with an emotion self-report collection app that traces keyboard interactions on a smartphone and collects self-reports. This enables the interface to access the keyboard interaction patterns and emotion self-report details to provide feedback in terms of typing behavior, the relationship between typing parameters (like typing speed, typing error, emoji usage) and emotion. For example, it allows users to reflect on questions like *"Do I make more typing error when I am happy?"*. We validate

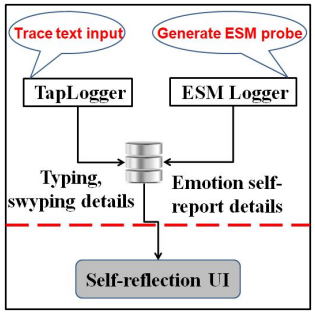


Figure 1: *EmoReflect* architecture. It traces the keyboard interactions, collects emotion self-reports and enables self-reflection.

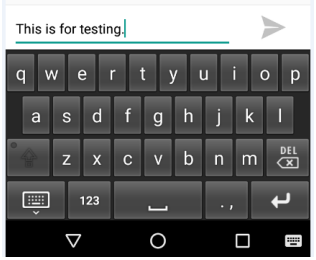


Figure 2: *EmoReflect* keyboard

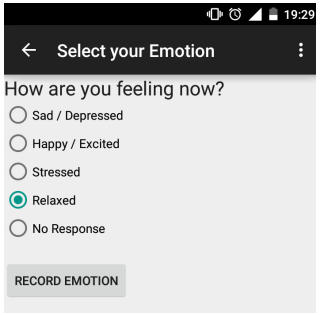


Figure 3: Self-report UI

the proposed interface using a 6-week in-the-wild study involving 20 participants. We perform a between-subjects study, in which half of the participants are exposed to the self-reflection interface (*study group*) and the other half is not exposed to it (*control group*). Analysis of the collected dataset reveals that on average the participants in the study group report 50% more emotion self-reports than the participants of the control group. Moreover, the average participation duration for the study group also increases by 74%. Although these early findings demonstrate the potential of self-reflection to improve self-report collection, in the future, we plan to investigate the quality of self-reports in detail.

Self-reflection Platform Design

We design and implement the self-reflection platform as an Android application *EmoReflect*. We enable self-reflection based on the user's keyboard interaction and emotion self-reports. Hence, *EmoReflect* consists of the following components - *TapLogger* traces the keyboard interaction pattern, *ESMLogger* generates and collects the emotion self-reports, while *Self-reflectionUI* supports the self-reflection based on collected typing data and emotion self-reports. We show the architecture of the *EmoReflect* in Figure 1.

TapLogger: Trace Keyboard Interaction
We develop an instrumented QWERTY keyboard using Android Input Method Editor (IME) facility as part of *EmoReflect*. We show the keyboard interface in Figure 2. We record the timestamp, associated application name, any non-alphanumeric character typed, pressure and speed during every touch interaction. To ensure user privacy, we do not store any alphanumeric characters.

ESMLogger: Collect Emotion Self-reports
We collect the emotion self-reports via ESM probes. We define the time spent by the user on a single application without changing it as a *session*. The ESM probes are triggered once the user completes text entry in a session. We collect four types of self-reports (*happy, sad, stressed, relaxed*) using ESM probes (Figure 3). We select these discrete emotions as their valence-arousal representation is unambiguous on the Circumplex plane [23]. The user can also skip self-reporting by selecting *No Response* in the UI (Figure 3).

Self-reflection UI: Reflect on Keystroke and Self-report Details
In this prototype, we implement two types of feedback (for simplicity). These insights come from answering two types of questions - `(Cat1)` Does emotion state significantly influence different typing characteristics? and `(Cat2)` Does the frequency of different emotion self-reports significantly vary across different time-period? We frame these questions as hypotheses and validate them using statistical tests. We show the self-reflection UI in Figure 4. We summarize each question category, the independent, dependent variables and the corresponding null hypothesis in Table 1.

For `Cat1` questions, we validate the hypothesis if emotions significantly influence different keystroke parameters. Here, the independent variable (cause or trigger) is the emotion state and the dependent variables are the keystroke parameters (*typing speed, typing error, session duration, emoji usage*). At a time, the user can select one dependent variable and check if it varies significantly across emotion states. Depending on the number of the emotion states and the distribution of the dependent variable (e.g. normal), the suitable statistical test is selected for hypothesis validation. We show two such examples (from a sample user) in Figure 5. In the left-hand figure, we show that the typing speed (mean elapsed time between two consecutive keypress events) is significantly ($p < 0.05$) lower in *sad*, while in the right-hand figure, we show that emoji usage is significantly

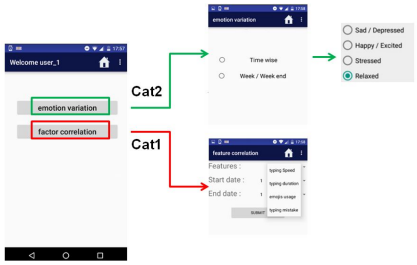


Figure 4: Self-reflection UI displaying both types of questions

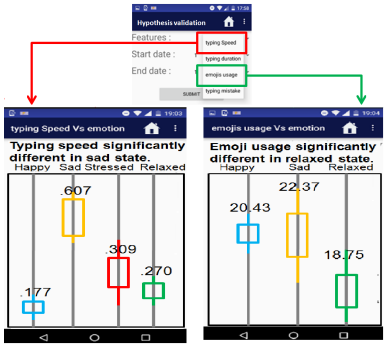


Figure 5: `Cat1` questions result to show if a keystroke parameter is significantly ($p < 0.05$) different across different emotions. The left-side figure shows that typing speed is significantly lower in *sad* state. The right-side figure shows that emoji usage is significantly lower when *relaxed*.

($p < 0.05$) lower when *relaxed*. In both cases, we apply one-way ANOVA [24] as we have more than 2 groups and the underlying distribution is normal.

Question category	Variable type	Variable name	Variable description	Null hypothesis
Cat1	Independent	Emotion states	happy, sad, stressed, relaxed	
	Dependent	Typing speed	Mean of all elapsed times between two successive key press events present in a session	H0: Typing speed does not vary significantly across emotions
		Typing error	Number of times backspace and delete keys are pressed in a text entry session	H0: Typing error does not vary significantly across emotions
		Session duration	Duration of the text entry session	H0: Session duration does not vary significantly across emotions
		Emoji usage	Number of special characters used in a text entry session	H0: Emoji usage does not vary significantly across emotions
Cat2	Independent	Time period	Different time period - time of day, weekday-weekend	
	Dependent	Happy frequency	Frequency of happy state	H0: Amount of happy self-report (%) does not vary significantly across selected time-period
		Sad frequency	Frequency of sad state	H0: Amount of sad self-report (%) does not vary significantly across selected time-period
		Stressed frequency	Frequency of stressed state	H0: Amount of stressed self-report (%) does not vary significantly across selected time-period
		Relaxed frequency	Frequency of relaxed state	H0: Amount of relaxed self-report (%) does not vary significantly across selected time-period

Table 1: Different question categories and the corresponding dependent and independent variables implemented in this prototype. The framed null hypothesis is shown in the last column.

For `Cat2` questions, we validate the hypothesis if the frequency ($\%$) of different emotions varies across the time-period. In this case, the independent variable is the time-period (i.e. weekday/weekend, time of day) and the dependent variable is the emotion frequency (*happy, sad, stressed, relaxed*). We compute the percentage of the selected emotion states for different time-period (e.g. in weekday and at the weekend). We perform Chi-square tests [24] to verify if the frequency distribution of the selected emotion varies significantly across the selected time-period.

Study Design

We design a between-subject study to measure the effectiveness of the self-reflection interface. We implement two variants of *EmoReflect* - (1) SR, which implements the self-reflection interface and (2) No-SR, which does not implement the same. These two variants of *EmoReflect* are used to perform the study, where we assigned half of the participants to use each of the prototypes.

Participants

We recruit 20 university students (16 males, 4 females, aged between $22 - 35$ years) for the field study. We split them into two groups, each containing 10 participants (8 male, 2 female). We install the No-SR variant (with no self-reflection capability) of *EmoReflect* in the smartphone of the participants of one group and denote the group as No-SR (*control group*). The SR variant (with self-reflection capability) is installed in the smartphones of the participants of the SR group (*study group*). We carry out the study during the same time-period (duration 6-weeks) to minimize external discrepancy.

Instructions to the Participants

We instruct the participants in each group to use the app for 6 weeks. They are asked to select the *EmoReflect* key-

Parameter	No-SR	SR
No. of participants	10 (8 M, 2 F)	10 (8 M, 2 F)
Avg. age	27.1 (sd. 3.48)	28.2 (sd. 3.65)
Total typing dur. (in Hr.)	54.75	48.07
Avg. typing session per day	72.4	70.0
Happy session (%)	15	14
Sad session(%)	11	4
Stressed session (%)	19	42
Relaxed session (%)	55	40

Table 2: Dataset details. Comparatively large number of *stressed* emotion in SR group is attributed to the ratings provided by two participants in the group, as they responded large number of probes as *stressed*.

board as the default keyboard. We inform the participants that when they switch from an application after completing text entry, they may receive a survey questionnaire as a pop-up, where they can record their emotions. We also inform the participants that they can skip self-reporting by selecting the *No Response* option in the pop-up. Additionally, the participants in the SR group are informed about the self-reflection facility, where they can find insights in terms of typing behavior and emotion variation. They are also told that the feedback is provided in terms of Q&A based on the data recorded during the study. A demonstration is given on how to use the self-reflection interface to find the answer to their questions.

Data Analysis

In this section, we discuss the dataset collected during the study. During the data collection period, we have collected 747,103 keypress events spanning across 103 hours of typing. We summarize the dataset for both the groups in Table 2.

Demographics and Typing Data Volume

The average age of the participants in No-SR and SR group is 27.1 years and 28.2 years respectively - the variation is not significant according to t-test (df = 9, t-stat = 0.79, p-value = 0.447). Similarly, according to t-test, the amount of time spent (df = 9, t-stat = 0.66, p-value = 0.527) and the average number of sessions recorded per day (df = 9, t-stat = 0.06, p-value = 0.956) by the users of No-SR and SR group is not significantly different.

Distribution of different Emotions

We also record the frequency distribution of different emotion self-reports (*happy, sad, stressed, relaxed*) from both the groups in Table 2. We observe that there is no significant difference in the frequency distribution of *happy* states between the participants from each group (df = 9, t-stat = 1.14, p-value = 0.281). The same is observed for *sad* (df = 9, t-stat = 1.49, p-value = 0.170), and *relaxed* (df = 9, t-stat = 0.41, p-value = 0.689) states. However, we observe a major difference in the distribution of the *stressed* state. We investigate further and identify that two participants in the SR group were highly *stressed* and almost 3 times more (than average) sessions were tagged with *stressed* emotion for them.

In summary, the collected dataset from both the groups are comparable in terms of typing duration, average daily session recorded. They are also found to be similar in terms of the frequency distribution of emotion. We use this dataset for evaluation.

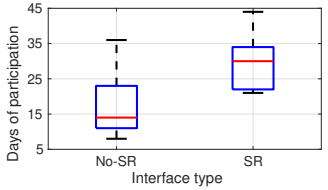
Evaluation

In this section, we evaluate the effectiveness of the self-reflection interface in engaging the participants during the study. We measure the engagement in terms of participation days and the number of self-report probes. In specific, we use the following metrics to measure the improvement in the self-report collection.

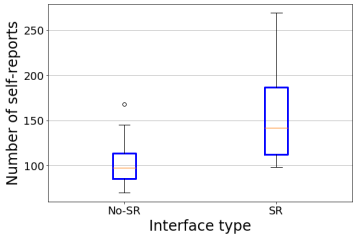
(a) Days of Participation: Although the study was performed for 6 weeks with each group, we measure the effective days of participation. It is defined as the total number of days in which users have answered the emotion self-reports (including *No Response*). On other days, there is no self-report probe generated (or attended) by the participants.

(b) Number of Self-report Probes: We measure the number of self-report probes generated in each of the variants.

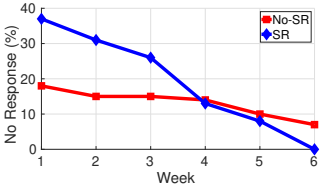
(c) Rate of Skipped Responses: We measure the percentage of self-reports skipped (recorded as *No Response*) in every week.



(a) Days of participation



(b) Number of self-reports



(c) Amount of *No Response*

Figure 6: Comparing engagement for both the groups (No-SR, SR). (a) Average number of participation days in SR group is found to be significantly ($p < 0.05$) higher. (b) The number of self-report probes obtained for each participant of the SR group is found to be significantly ($p < 0.05$) higher. (c) Amount of *No Response* probes in SR group reduces with time and drops below the same of the No-SR group from week 4.

Days of Participation

We compare the days of participation in Figure 6a. For the No-SR group, the average number of participation days is 17.0, while the same for the SR group is 29.6. We obtain an improvement of 74% in average days of participation for the SR group with respect to the No-SR group. The difference is significant at $p < 0.05$ (measured using Welch's unequal variances t-test (df = 9, t-stat = 4.94, p-value = 0.000).

Number of Self-report Probes

We also compare the number of self-report probes for both the variants. We observe that the participants of the SR group encounter a significantly high number of probes than those of the No-SR group using Welch's unequal variance t-test (df = 9, t-stat = 2.14, p-value = 0.030). The average number of probes responded by the No-SR group participants is 106.2, while the same for the SR group is 159.4 (50% more than that of the No-SR group). We show the comparison of self-report probes in Figure 6b. Our analysis reveals that the primary reason for having more probes for the SR variant is that the participants using the SR variant stay engaged in the study for a longer period (as seen in Figure 6a).

Rate of Skipped Responses across Time

We also compare the percentage of self-reports not answered in Figure 6c. We observe that for the No-SR group participants, the amount of *No Responses* remains almost identical over weeks. It is also observed that initially the amount of *No Responses* is high for the SR group participants, however, with every week it reduces. This may be attributed to a high number of probes that the SR group participants may have received in the initial period. However, from week 4, the amount of *No Response* for the SR group is even less than the same of the No-SR group. This indicates that participants of the SR group skip fewer probes with time and responds to more probes in the long-term. This may be attributed to the fact that once the participants find the feedback provided by the self-reflection interface useful, they record more emotion labels and provide less *No Response* labels.

All these findings indicate that the participation days and average probing rate of participants are significantly higher in the case of the SR group; which demonstrates the suitability of self-reflection to engage the participants in the long-term to obtain more responses.

Conclusion and Future Work

In this work, we show the promise of using self-reflection assisted emotion self-report collection in an ESM-based study. We design, and develop an Android application *EmoReflect*, which traces users' keyboard interactions, collects emotion self-reports and allows them to reflect on their typing behavior during the data collection. Our preliminary findings reveal that using self-reflection the participants feel more engaged with the study, which is manifested in terms of improvement in participation duration and the number of self-reports collected during the study.

At the same time, this early work provides important future research directions. First, although we collect a large number of self-reports, we want to validate the quality of such self-report in terms of emotion classification. Next, we want to investigate the influence of any confounding variables (e.g. application usage, age, and gender variability) on the outcome of the study. We also aim to improve the self-reflection interface for better visualization and to support other categories of questions to make the interface more engaging. Finally, as a long-term research plan, we aim to investigate the implications of self-reflection in mental health-related problems (e.g. stress, depression).

REFERENCES

[1] Niels Van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The Experience Sampling Method on Mobile Devices. *ACM Computing Surveys (CSUR)* 50, 6 (2017), 93.

[2] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. 2017. Understanding Self-Reflection: How People Reflect on Personal Data Through Visual Data Exploration. In *Proceedings of the PervasiveHealth*.

[3] Marilyn Wood Daudelin. 1996. Learning from experience through reflection. *Organizational dynamics* 24, 3 (1996), 36–48.

[4] Carlo C DiClemente, Angela S Marinilli, Manu Singh, and Lori E Bellino. 2001. The role of feedback in the process of health behavior change. *American journal of health behavior* 25, 3 (2001), 217–227.

[5] Mica R Endsley and others. 1997. The role of situation awareness in naturalistic decision making. *Naturalistic decision making* 269 (1997), 284.

[6] Rosta Farzan, Joan M DiMicco, David R Millen, Casey Dugan, Werner Geyer, and Elizabeth A Brownholtz. 2008. Results from deploying a participation incentive mechanism within the enterprise. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 563–572.

[7] Joel E Fischer, Chris Greenhalgh, and Steve Benford. 2011. Investigating episodes of mobile phone activity as indicators of opportune moments to deliver notifications. In *Proceedings of ACM MobileHCI*. 181–190.

[8] Zachary Fitz-Walter, Dian Tjondronegoro, and Peta Wyeth. 2011. Orientation passport: using gamification to engage university students. In *Proceedings of the 23rd Australian computer-human interaction conference*. ACM, 122–125.

[9] Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2019. Designing An Experience Sampling Method for Smartphone based Emotion Detection. *IEEE Transactions on Affective Computing* (2019).

[10] Joyce Ho and Stephen S Intille. 2005. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *Proceedings of ACM SIGCHI*. 909–918.

[11] Simo Hosio, Jorge Goncalves, Vili Lehdonvirta, Denzil Ferreira, and Vassilis Kostakos. 2014. Situated crowdsourcing using a market model. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, 55–64.

[12] Gary Hsieh, Ian Li, Anind Dey, Jodi Forlizzi, and Scott E Hudson. 2008. Using visualizations to increase compliance in experience sampling. In *Proceedings of the 10th international conference on Ubiquitous computing*. 164–167.

[13] Ravi Karkar, Jessica Schroeder, Daniel A Epstein, Laura R Pina, Jeffrey Scofield, James Fogarty, Julie A Kientz, Sean A Munson, Roger Vilardaga, and Jasmine Zia. 2017. TummyTrials: A Feasibility Study of Using Self-Experimentation to Detect Individualized Food Triggers. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 6850–6863.

[14] David A Kolb. 2014. *Experiential learning: Experience as the source of learning and development*. FT press.

[15] 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.

[16] Ian Li, Anind K Dey, and Jodi Forlizzi. 2011. Understanding my data, myself: supporting self-reflection with ubicomp technologies. In *Proceedings of the ACM UbiComp*.

[17] Peter Lynn. 2001. The impact of incentives on response rates to personal interview surveys: Role and perceptions of interviewers. *International Journal of Public Opinion Research* (2001).

[18] Abhinav Mehrotra, Jo Vermeulen, Veljko Pejovic, and Mirco Musolesi. 2015. Ask, but don't interrupt: the case for interruptibility-aware mobile experience sampling. In *Adjunct Proceedings of the ACM UbiComp/ISWC*. 723–732.

[19] Mohamed Musthag, Andrew Raij, Deepak Ganesan, Santosh Kumar, and Saul Shiffman. 2011. Exploring micro-incentive strategies for participant compensation in high-burden studies. In *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 435–444.

[20] Veljko Pejovic, Neal Lathia, Cecilia Mascolo, and Mirco Musolesi. 2016. Mobile-based experience sampling for behaviour research. In *Emotions and Personality in Personalized Services*. Springer, 141–161.

[21] Veljko Pejovic and Mirco Musolesi. 2014. InterruptMe: designing intelligent prompting mechanisms for pervasive applications. In *Proceedings of ACM UbiComp*. 897–908.

[22] Quantified Self 2019. (2019). `http://www.quantifiedself.com/`

[23] James A Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178.

[24] Neil J Salkind. 2010. *Encyclopedia of research design*. Vol. 1. Sage.

[25] Arthur A Stone, Ronald C Kessler, and Jennifer A Haythomthwatte. 1991. Measuring daily events and experiences: Decisions for the researcher. *Journal of personality* 59, 3 (1991), 575–607.

[26] Niels Van Berkel, Jorge Goncalves, Simo Hosio, and Vassilis Kostakos. 2017. Gamification of Mobile Experience Sampling Improves Data Quality and Quantity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 107.

[27] Niels van Berkel, Chu Luo, Denzil Ferreira, Jorge Goncalves, and Vassilis Kostakos. 2015. The curse of quantified-self: an endless quest for answers. In *Adjunct Proceedings of the ACM UbiComp / ISWC*.

[28] Dominik Weber, Alexandra Voit, Philipp Kratzer, and Niels Henze. 2016. In-situ investigation of notifications in multi-device environments. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1259–1264.

[29] Miriam Zisook, Sara Taylor, Akane Sano, and Rosalind Picard. 2016. SNAPSHOT Expose: Stage Based and Social Theory Based Applications to Reduce Stress and Improve Wellbeing. In *CHI 2016 Computing and Mental Health Workshop, San Jose, CA*.