A Retrospective and a Look Forward: Lessons Learned From Researching Emotions In-the-Wild

Benjamin Tag, Jorge Goncalves, Sarah Webber, Peter Koval, and Vassilis Kostakos, *University of Melbourne, Parkville, VIC, 3010, Australia*

Emotions have a significant impact on our decision-making, learning, awareness, social interactions, and mental and physical health. Even though large efforts have been put into quantifying human emotions, their subjectivity, context-dependence, and complexity render them as being almost unpredictable. However, while different streams of research in pervasive computing and psychology have made significant progress in the quantification of emotions, the most successful research results come out of controlled laboratory studies. In this article, we present a retrospective of a series of in-the-wild studies through the lens of human emotions. We are looking at the weaknesses and strengths of traditional research methods and present lessons learned. We furthermore call for a readjustment of research rigor and describe potential new research designs specific to out-of-the-lab studies.

The pervasive computing community has a long tradition of making methodological contributions to how we run human subjects studies outside the lab. Many of these contributions were originally driven by the early focus on context-aware computing, while more recently a growing interest in emotion and (mental) health has increasingly driven this work.[1]

In this article, we present a retrospective analysis of traditional methods that we and the community have used to study people's emotions and behaviors outside the lab. Over a number of years, we have conducted a range of studies, across different countries and populations, to investigate a variety of factors relevant to human emotion and technology use in the wild. Indicatively, we reflect on methods including: interviews, diary studies, the experience sampling method (ESM), and remote sensing.

Through our retrospective analysis, we reflect on the challenges and opportunities of conducting experiments outside the lab. In a post-COVID era, it is timely to put a spotlight on the methods we use as a community, identify methodological gaps and opportunities that have arisen, and propose ways in which our community can reconsider what is deemed "rigorous" beyond the lab. This article draws on our body of work on emotion as a lens for reflection, but we argue that many of the insights and lessons learned apply more broadly to human subjects studies out-of-the-lab.

HANDS-ON WITH IN-THE-WILD METHODS

Studying humans outside the lab has a rich and long tradition. The psychologist Kurt Lewin was one of the first to promote the scientific study of "the forces that structure daily thought and behavior,"[2] as early as 1935. At that time, the most rigorous research methods for investigating human behavior were diary studies, questionnaires, interviews, and observational studies. While diary studies and questionnaires offered a reliable solution for collecting behavioral and contextual data describing the phenomena of interest, interviews, and observational studies often led to disruptions of the behavior of interest, or required reliable recall of events and behavior.[3]

Intrapsychic experiences such as emotions are highly subjective, situational, and context-dependent,[4]

Digital Object Identifier 10.1109/MPRV.2021.3106272

which limits the extent to which the four methods listed above can capture a holistic account, as they entail one-time assessments, require recall, and are impacted by the observer effect. Some of these limitations are mitigated by newer methodologies made possible by remote communication and personal digital devices such as smartphones. These novel approaches include new modes of conducting *interviews* remotely, logging information for *diary studies* using digital photos, the *ESM*, and *remote sensing*.

A full survey of all methods used in emotion research in-the-wild is beyond the scope of this article. Such methods include the use of body movements; *in situ* annotations, e.g., with colors or photos; implicit methods, e.g., typing and touch interactions; analysis of social media behavior, mobility, technology use; and the combination of multimodal cues, such as visual, verbal, and biosignals. All these methods present challenges, e.g., *in situ* annotations interrupt activities, biophysical signals need to be mapped to emotional ground-truth data, and the analysis of mobility and social media behavior poses ethical questions. We focus our retrospective on interviews, diary studies, ESMs, and remote sensing, methods with which we have extensive experience, and which have been widely used and validated over the years. Moreover, particularly the former three methods are vital for collecting the emotional ground-truth needed to validate newer methods such as remote sensing.

Interviews

User interviews enable researchers to quickly gain rich qualitative data. Often they are conducted in-the-field or remotely, e.g., via email or teleconferencing. Well-designed interviews can yield deep insights into feelings, experiences, and motivations. They can be of long duration and result in large amounts of rich data, which entails considerable time and effort in conducting and analyzing interviews, especially if sample sizes are large.[5]

Interviews are common in technology-based studies to understand people's experience in different contexts. We have found, when using interviews to understand emotional interactions, it helps to provide users a context to reflect upon, such as a specific technology, sensation, or situation. These contexts create a scaffold for the interviewer to explore elements of user experiences that users may struggle to identify without prompting. Often, interviews are used in conjunction with other methods discussed, to answer questions which cannot be addressed directly through other methods or to complement the findings they yield. In our studies, we have conducted both in-person,[8] and remote interviews using voice or video communication such as Zoom and Skype.[9]

Diary Studies

In diary studies, participants are asked to log events, thoughts, experiences or other phenomena of interest to the researcher. Logs can consist of written notes, voice memos photographs, chat messages, or any other record that aids in recalling the behaviors, activities, or emotions of interest. Notes are often collected by the researchers for further analysis. Logging can span a few days or several months,[6] providing insights into experiences over time. Depending on the research question and the diary study design, typical study outcomes are chronological patterns of how experiences changes, and information on factors that influence experiences. A crucial dimension of diary study design is the logging schedule: e.g., every hour, at the end of the day, or when a specific event or action happens. We have extensive experience with conducting diary studies to investigate emotions, emotion regulation, and interaction with technologies. These studies have involved logging using paper, online forms, and smartphones (e.g.,[10]).

Experience Sampling Method

In ESM studies—while we are aware of the differences, we include ecological momentary assessments (EMA) in our discussion of ESMs, using them interchangeably, self-reports in the form of short surveys are proactively triggered at different points in time throughout the waking hours of the participants.[7] This allows for a systematic collection of contextual information without heavily relying on the participants' memories.

This is a method we have used extensively in combination with remote sensing to gain insights into the situational context of users. Sometimes we have used this method as an additional data point (e.g., user opinion[7]), while other times it has provided us with ground-truth labels (e.g., user's current context or emotional state[8]). On occasion, we have conducted studies where the experience sampling itself was the focus.[7] We have conducted a series of studies investigating how to improve accuracy in users' ESM responses, for example, by varying interruption frequency or considering contextual factors.

Remote Sensing

Remote sensing has largely been enabled by the widespread use of personal devices, especially mobile

devices, which provide new possibilities for automatically collecting contextual and behavioral data without requiring active input by participants. New approaches to continuous, unobtrusive data collection, e.g., through wearable sensors, including on-body sensing for heart-rate, electrodermal activity, and off-body systems such as cameras for facial expression or gait detection, open up new possibilities for out-of-the-lab research of emotions.[1,8]

A large number of our studies, increasingly looking at mental states and emotion, use remote sensing on smartphones, in smart-home environments, or from wearable technology for capturing data about users' contexts. Very often, we use this type of data to help us understand user behavior and mental states, or build predictive models of those. Examples include making predictions about what people will do on their phone,[8] whether they can be interrupted,[11] or their cognitive[12] or emotional state.[8] Effectively, sensor data are trained on some available ground-truth dataset of user emotion or behavior, such as self-reports or ESM labels. In combination, these allow for accurate predictions based on readily available sensor data.

LESSONS LEARNED

We report on our experiences with applying the above methods to study emotion outside the lab. Here we articulate our lessons learned in this context, and subsequently in the discussion we identify issues that are more broadly relevant to studies outside the lab.

Interviews

Conducting online or remote interviews can be challenging, but at the same time provides some unique advantages. Assuming that participants undertake remote interviews while at home, being in a familiar setting for a conversation allows them to easily draw links, in the moment, between the researchers' requests and their own lives.

For example, in one study we were concerned that interviewees may be reluctant to talk about emotional experiences. However, participants talked about difficult emotional experiences, challenging interpersonal circumstances, and intimate details of their inner lives. We posit that being in a familiar location (rather than a lab setting) puts participants at ease, helping them to feel comfortable reflecting on, and articulating, their emotional experience. The distance imposed by video call technology may help participants to feel "safe" to talk honestly with interviewers. It is, however, the responsibility of the interviewer to create and to maintain this safe feeling throughout the interview, e.g., by actively listening, but also being prepared to react to technical problems and resulting frustration, disruptions, and being aware of time.[13]

Technical difficulties may impede remote interviews. But as we anticipated, our interviewees had experience of using video call software for professional purposes. In addition, video calls are now widely used for social and personal purposes, and so are likely to be an increasingly familiar context for many participants to discuss aspects of their professional, personal, and emotional lives.

An additional benefit of video call software is that it allows for easy recording of interviews. However, we have also found in a small proportion of online interviews that it is difficult to read participants' emotional cues, which impairs our ability to engage in productive conversations about emotion, to a certain extent. This generally occurs when participants prefer to keep their cameras turned OFF or there is a poor network connection, which makes it more challenging to catch nonverbal clues, such as facial expressions or posture. We note also that in online interviews, as opposed to face-to-face interviews, emotional expression may be impaired by the limited field of view of the camera (which means some body language cues are not visible), and by speakers' needs to position themselves carefully in front of a camera. While the familiar setting may be beneficial for the interviewees' perception of safety, it also limits the control researchers have, as distractions in the room that are outside the field of view of the camera are hidden from the researcher. Together these factors may impede interviewers' ability to pick up on contextual clues and emotional tone, which are critical when talking about emotion. A thorough preparation, such as providing clear instructions, early scheduling, sufficient time for the interview, testing of all used technology, and a set of questions that have been piloted, are vital for ensuring that interviews are easy for interviewees and meaningful for the interviewer.

Having previously found that some participants are nervous about being video recorded, in our interviews, while being face-to-face with our interviewees, we selected to record only the audio. However, we posit that video recording online interviews is still far less confronting for participants than a full AV-recording setup in a lab. None of our participants refused to consent to neither audio nor video recording. We propose that video recordings of interviews and contextual information about participants' environments, which can both be easily captured through teleconferencing software, might aid in accurate interpretation

and analysis of accounts of emotional experiences. This is particularly valuable if transcripts are analyzed by researchers who did not observe or conduct the interviews. Moreover, emerging technologies enabling automated analysis of changes in the voice and facial expressions can enrich the traditionally rich qualitative insights with quantitative data.

Diary Studies

It is an inherent feature of diary studies that they are conducted remotely, outside the lab. A key difference imposed by COVID-19 was that the whole process has to be conducted remotely, including the crucial initiation session. As with interviews, an advantage of the participant being in their own home at initiation of a diary study is that they can easily identify and implement small prompts to remind themselves to complete the diary activities (e.g., placing a post-it note or notepad on their desk).

There are challenges to ensuring participant compliance with diary studies when no face-to-face meetings are planned. Smartphones, however, allow for lifting some of that burden of the participants' shoulders. It is possible to set automated reminders, or open up communication channels through messenger apps, that can be used to remind participants, or for participants to *in situ* communicate their notes. The multifunctionality of smartphones also enables the participant to use alternatives to note taking, such as photographs or voice memos. In a study on the use of video games for emotion regulation, we provided the participants with a diary template and decided to send daily reminders to our participants to ensure that they kept their diary.[9] In a second study on technology use for emotion regulation, we aimed to make diarizing a lightweight task. We asked participants to either take a note or a photograph (or screenshot) with their phone when they noticed they were using digital technologies as part of their everyday emotional lives. We offered to send reminders, but participants mostly declined.

In the latter study, we found that all participants complied with the request to record at least four events of technology use to regulate emotions over the course of a week. We note that there was some inconsistency across the group in terms of the number of events recorded. Also, there was no effort to balance sampling of events in terms of, e.g., time of day or emotional valence. If this is vital information for a research question though, we recommend to proactively send reminders to participants. We hypothesize that the sample of events is likely biased toward participants' cognitive availability—as they would be unlikely to record events when they were very busy or actively engaged in social interaction—and perhaps toward highly salient emotional experiences.

We found that the diary entries (photographs) are effective prompts for participants to jog their memory. In most cases, these are sufficient for participants to recall the context and their emotional experience. They further enable interviewees to talk about these experiences in depth. In our study, we elected to not gather participants' diary entries for further analysis. However, photographs and other recordings might be used to gather rich contextual information, which can aid in analyzing data such as participants' verbal reports and sensor data.

In a different study, we used a diary methodology to validate findings of an in-the-wild deployment. In this example, we used an interactive public display to capture the overall happiness of a particular community over a period of several weeks using a projective test.[10] We also used an independent diary reconstruction method study to validate the outcomes.

Experience Sampling Method

We usually initiate ESM studies by inviting participants to a short in-person briefing during which we can a) build a rapport with participants, and b) explain the ESM protocol and survey questions, allowing participants to seek clarification if needed. This initial session serves to motivate and engage participants in the research process, which is crucial for obtaining good compliance over the course of the ESM sampling period (typically 7 days or longer). This process has become more challenging during COVID-19 since all aspects of our ESM research had to move online. Rather than skipping this initial briefing altogether, we have used online screening surveys (containing attention checks and eligibility questions) to filter out potentially ineligible, careless, or unmotivated respondents. We have also requested participants to install our ESM software during the screening survey to check for compatibility with their devices. Finally, we have replaced in-person briefing sessions with short videos introducing the study, which participants are instructed to view before commencing the study. This is especially helpful when recruiting participants in different time zones and to enable participants to start the study on their own schedule. This also helps to lower the burden on the researcher, as scheduling these sessions can be cumbersome, e.g., when the sample is rather large. As these sessions can easily be skipped, they could be followed by a few key

comprehension questions to ensure participants have watched the intro videos attentively and understand all study requirements.

We have typically used signal-contingent pseudo-random ESM sampling schemes, which involve dividing participants' waking hours into several sampling windows and prompting participants to complete an ESM survey at a random moment during each sampling window. This aims to ensure that participants' experiences are sampled across a representative range of daily contexts and activities. However, even when using such sampling designs, there is a risk that participants may systematically miss ESM surveys when experiencing intense emotions (e.g., when highly anxious). This would be problematic because ESM is often assumed to offer the advantage (e.g., over lab studies) that it helps to capture the full range of emotions people experience in their everyday lives. Reassuringly, a recent study examining unobtrusive audio recordings coinciding with missing versus completed ESM surveys suggests that participants are not more likely to miss ESM surveys when experiencing more intense emotions.[14]

Another challenge to obtain representative self-report data, particularly on emotion using ESM, relates to people's ability to introspect, attend, and accurately identify their subjective feelings repeatedly over time. While this ability appears to differ between individuals,[7] our findings suggest that overall people's self-reported emotions are sufficiently meaningful and differentiated to be useful for studying emotions in daily life. A further major challenge for emotion research using ESM is that, at least in nonclinical populations, people rarely experience intense negative emotions in daily life.[15] As a result, it is common to observe a large number of participants rate near the lowest score—so called "floor effects" in distributions of ESM reports—of negative feelings. Perhaps even more problematically, this may limit within-person variability in negative emotions due to statistical confounds between the mean and variance.[15]

Finally, several challenges arise when analyzing the ESM data. Many phenomena of interest may be relatively brief, e.g., emotions typically last less than 1-2 hours.[16] Given the typical ESM sampling frequency is approximately 2 hours, it may be difficult to track the temporal unfolding of the studied phenomena. Instead, ESM may be more likely to capture less event-related events that fluctuate more slowly. In order to fill these gaps, and keep the burden as low as possible for participants, we have made good experiences with combining passive sensing techniques with ESM data collection. Also, it is important to consider whether collected data are likely to be skewed in any way. For example, daily life emotions appear to be characterized by nonlinear dynamics, which pose data analytic challenges and imply that standard (linear) models may not be appropriate.

Remote Sensing

With remote sensing techniques, participant recruitment depends significantly on the nature of the study, and the underlying technologies. For instance, when using smartphones, it is possible to remotely recruit participants using the appstore, but often an enrollment process is required to brief participants about the study objectives. This is particularly important when studying sensitive phenomena, such as mental health and emotions. Beyond smartphones, remote sensing often requires physical or in-person interaction, e.g., in the case of remote sensing in homes. For example, in a recent study, we had to send a device by physical mail to each participant to install in their home. We conducted an online enrollment session to support the setup remotely and ensure it works as expected. Using prototypes, apps, and devices can result in incomplete or inaccessible data and even participant frustration. For example, when sensing devices are not remotely accessible (e.g.,[12]), we have to rely on the participants to detect and report technical issues. Similarly, for studies that have involved cloud services or platforms, unexpected changes by the service provider may taint the collected data. Overall, we have found it useful to be available to support participants with any technical questions that might occur through email, or periodically check in with participants during a remote sensing study, e.g., by asking them to fill in a weekly questionnaire or conduct intermittent interviews.

A plethora of pervasive, mobile, and wearable sensors can be used for researching emotions, including but not limited to heart-rate variability (HRV), electrocardiography (ECG), and electrodermal activity (EDA).[17] Recently, one technology has been especially under scrutiny, namely facial expression analysis for emotion detection. While most of these systems are trained on ideal high quality image datasets, our benchmark analysis identified significant shortcomings on realistically distorted images.[18] However, we also provide a blueprint for validating the effectiveness of one of such systems to detect emotions outside the lab. Our study identified a bidirectional relationship between smartphone usage and emotional state.[8] We used ESM labels to validate the correctness of the facial recognition.

TABLE 1. Summary of pains and gains of methods used for researching emotions in-the-Wild.

Method	Gains	Pains	Recommendations	Examples
Interviews				[5], [8], [9]
	* rich qualitative data * can be done remotely and in the field * good for triangulation	* question design * time-consuming * large samples * large sample analysis * reliance on participants' memory	* provide users with context (e.g., situation, technology) * give clear instructions * schedule early and sufficient time * pre-analyze available data before interview	
Diary Studies				[10]
	* insights in chronological changes of experiences * digital devices can ease burden * not as intrusive as ESM * good for longitudinal studies	* require participant diligence (burden) * timing and frequency of diary * large sample * large sample analysis	* use digital technology to lower burden on participants (e.g., use photos or voice memos instead written notes)	
Experience Sampling Method				[7], [8]
	* systematic collection of context information * not relying on participant's memories * insights in slow chronological fluctuations * good to combine with other methods * variety of (digital) tools available	* can be intrusive and burdensome * risk of missing ESMs at vital times (e.g., anxiety episode) * sampling frequency is important * participants' ability to introspect * temporal unfolding of experiences might be missed * longitudinal recordings	* use for ground-truth data collection * conduct initiation session (face-to-face or remote) * ensure participant motivation * demo technology with participants * use personalized introduction videos for larger samples or remote initialisation * lower pressure on participants by providing full reimbursement for $<100\%$ compliance (e.g, $>80\%$) * provide additional incentives (e.g., individualized feedback)	
Remote Sensing				[1], [8], [11], [12]
	* alternative recruitment methods (e.g., through study app) * lower burden on user (automatic, passive sensing) * granularity (continuous sensing) * provides contextual, emotional, behavioral data	* no standardized recruitment * ground-truth data collection * technical requirements (computing power, signal processing, device requirements) * validation of new methods * high risk of noise in data * privacy and ethical issues	* use in combination with ESM or Diary to validate findings and build accurate prediction models * demo technology with participants * check-ins through short surveys or interviews to ensure technology works well * run longer studies to account for noise in data	

Although smartphones can capture a range of sensor data, it is challenging to capture the nuances of human behavior, social context, or emotion. A major advantage of remote sensing is that data can be collected continuously without requiring active user input, a key affordance in emotion related research. However, the inherently limited control over conditions outside the lab renders the collected data highly susceptible to noise. Not only do many on-body sensors require proper placement on the participants' bodies, but also it is usually not possible to check during a study if data are correctly collected without interrupting the participants' everyday lives. The high susceptibility to noise can be accounted for by running field studies over longer periods of time. But, often the data analysis happens post-hoc, meaning it only becomes clear if the data collection was successful after the participant has finished their study. This can render field studies costly and inefficient in case a repetition is necessary. Therefore, it is important to understand the limitations and assumptions present in the collected data as summarized in Table 1.

DISCUSSION

Mitigating Pains

The strengths and weaknesses of the methods we refer to are mostly well documented, and obviously depend on the exact phenomena being studied. Therefore, we want to comment on how these methods are often combined to mitigate shortcomings of one specific method, and reflect on combinations of methods that work well.

For instance, we have observed that ESM works very well with remote sensing. Typically, ESM studies now utilize participants' own smartphones for deploying and scheduling the questionnaires. This means that participants need to install some software on their phone (although not necessarily always). This provides an opportunity to enrich said software with sensing abilities, such that it can passively—without requiring active participant input—and continuously record data throughout the study. This method couples a continuous stream of sensor data with intermittent questionnaire data, offering richer insights. ESMs have been validated to robustly collect emotion information, and are especially interesting for collecting vital ground-truth data outside the lab.

A drawback of the reliance on ESMs is—as most literature suggests—that after three weeks, participants' response rates to ESMs will dramatically decline.[7] Therefore, for longitudinal studies it can be an option to employ a diary study approach, which is less intrusive and less demanding, and can additionally be conducted more easily on smartphones. Then, during follow-up interviews, those photos are used to drive the conversation and further unpack the participants' experiences. As the photos and other diary log techniques are used to jog the participants' memories about the situation or emotion of interest, it is important to collect this information as soon as possible after. Long delays can lead to a loss of detail and biases in the participants' recall.

Conducting interviews is typically encouraged when possible or practical. Sometimes it is not easy to conduct interviews due to a large sample size, or a geographically distributed sample. But most often, participants need to contact researchers before enrolling in a study, and at that point it is possible to arrange an interview. We have found it very helpful to prepare for interviews by carrying out some preliminary analysis or assessment of participant data. For instance, it is possible to summarize the collected sensor data for each participant, and use relevant graphs or charts to drive the interview. In a diary study, it helps to have looked at the diary entries before holding the interview.

Rethinking "rigour" Outside the Lab

Researchers in our field—ourselves included—often find themselves in a situation where they try to analyze data from field studies using techniques and approaches that have been prevalent in controlled lab experiments. This has been problematic for a number of reasons. Field studies are notoriously challenging to control, and to a large extent it remains questionable whether a field study should be controlled and shoehorned into a "lab study outside the lab." Furthermore, the way field studies are analyzed often reflects the expectations of the community and reviewers, who in turn are more likely to be favorable toward techniques they are familiar and comfortable with.

Typical complaints about field studies contain small sample sizes. In the field of human–computer interaction (HCI), the most prevalent sample size is 12 participants per study.[19] While the criticism that a sample is "too small" is often a consequence of the *threshold myth*, i.e., there is a threshold dictating when a sample is large enough, sample sizes can be justified. Qualitative studies often aim for *saturation*, the point where no new information can be elicited from new participants, quantitative studies have tools such as *power analysis* available for defining a proper sample size. However, none of these methods is without limitations and free of criticism, e.g., one factor

power analysis requires is the level of expected noise.[19] This is virtually impossible to predict when it comes to field studies.

While the overall aim of applied research is to discover, interpret, document, and develop methodologies and systems that increase human knowledge, lab, and field studies in HCI and pervasive computing aim at evaluating designs and systems. Perhaps an alternative approach to being statistically "rigorous" in the field is to aim to demonstrate that the richness of the phenomenon in question has been captured. Rather than aiming to get "enough data," one could explicitly aim to get data that is "rich enough" to describe a phenomenon. Rather than obtaining data in lab quality that reproduces the effect sizes of lab studies, field studies should focus on detecting the multifariousness of phenomena, e.g., the multitude of emotions felt at one point in time, rather than a significant level of one emotion triggered by a stimulus in a controlled environment. An alternative way to think about study design can be inspired from the notion of micro-randomized trials,[20] which refer to how (medical) interventions can be delivered dynamically to participants. Along these lines, perhaps we can consider how a field study with 12 participants can have 12 populations of n=1 rather than one population of n=12. This mindset shift could be a way to maximize richness of insights, especially when studying complex phenomena such as emotion. Of course, this very much depends on the aims of the study and the relative maturity of the literature on the topic. But nevertheless, a field study should be conducted to obtain data and provide answers that a lab study cannot fully generate.

Finally, rather than conducting lab studies outside the lab, we have to readjust our expectations of methodologies and rethink our approach to field studies. As we still need tools, such as ESMs and diaries to obtain valid ground truth labels, we are not calling for omitting traditional methods. However, we should take advantage of existing smart technologies to create true out-of-the-lab experiments. For one, we should focus on lowering the burden on study participants as only this guarantees the most natural behavior. For example, with increasing length of a study, the frequency of ESMs needed to collect ground-truth labels can be lowered while models based on passively collected sensor data can continuously become more accurate. After an initial period, the models can be revalidated by less and less frequently triggered ESMs. Depending on the individual, these periods may differ and require a less stringent but smarter study design. This will also require us to be better prepared for different devices, setups, lifestyles of our participants. These smart study designs will be adoptable, and adjust to individual participants. Diaries and interviews can be timed in response to extreme events or false predictions made. Consequently, we will be able to quantify real world phenomena in ways that live up to human behavior and emotions being complex, highly subjective, and context-dependent.

CONCLUSION

We have presented a retrospective analysis of traditional methods used to run studies outside the lab. By looking at studies using interviews, diaries, experience sampling, and remote sensing to study the complexity of human emotions outside the lab, we detailed challenges and opportunities. Synthesizing from our experiences, we sketch out potential changes to traditional study methodologies and call for a readjustment of the lab study rigour that we traditionally use to assess and validate field studies. When if not in a post-COVID era is the right time to adopt newly gained experiences and improve our approach to human subjects studies out-of-the-lab.

ACKNOWLEDGMENTS

This work was supported by the Australian Research Council under Grant DP190102627.

REFERENCES

1. O. Amft, J. Favela, S. Intille, M. Musolesi, and V. Kostakos, "Personalized pervasive health," *IEEE Pervasive Comput.*, vol. 19, no. 3, pp. 11–13, Jul.–Sep. 2020.
2. M. Csikszentmihalyi, *Flow and the Foundations of Positive Psychology*. Dordrecht, The Netherlands: Springer, 2014.
3. R. Larson and M. Csikszentmihalyi, "The experience sampling method," *Flow and the Foundations of Positive Psychology*. Dordrecht, The Netherlands: Springer, 2014.
4. L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychol. Sci. Public Int.*, vol. 20, no. 1, pp. 1–68, 2019.
5. A. Adams and A. L. Cox, "Questionnaires, in-depth interviews and focus groups," *Research Methods for Human-Computer Interaction*, Cambridge, U.K.: Cambridge Univ. Press, 2014.
6. N. Bolger, A. Davis, and E. Rafaeli, "Diary methods: Capturing life as it is lived," *Annu. Rev. Psychol.*, vol. 54, pp. 579–616, 2003.

7. N. van Berkel *et al.*, "Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports," *Int. J. Hum. Comput. Stud.*, vol. 125, pp. 118–128, 2019.
8. Z. Sarsenbayeva *et al.*, "Does smartphone use drive our emotions or vice versa? A causal analysis," in *Proc. ACM Int. Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–15.
9. Z. Sarsenbayeva, B. Tag, S. Yan, V. Kostakos, and J. Goncalves, "Using video games to regulate emotions," in *Proc. 32nd Australian Conf. Hum.-Comput. Interact.*, 2020, pp. 755–759.
10. J. Goncalves *et al.*, "Projective testing of diurnal collective emotion," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2014, pp. 487–497.
11. T. Dingler, B. Tag, S. Lehrer, and A. Schmidt, "Reading scheduler: Proactive recommendations to help users cope with their daily reading volume," in *Proc. 17th Int. Conf. Mobile Ubiquitous Multimedia*, 2018, pp. 239–244.
12. B. Tag, A. W. Vargo, A. Gupta, G. Chernyshov, K. Kunze, and T. Dingler, "Continuous alertness assessments: Using EOG glasses to unobtrusively monitor fatigue levels in-the-wild," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 464:1–464:12.
13. S. Weller, "Using internet video calls in qualitative longitudinal) interviews: Some implications for rapport," *Int. J. Soc. Res. Methodol.*, vol. 20, no. 6, pp. 613–625, Dec. 21, 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33342369/
14. J. Sun, M. Rhemtulla, and S. Vazire, "Eavesdropping on missing data: What are university students doing when they miss experience sampling reports?," *Pers. Soc. Psychol. Bull.*, 2020.
15. E. K. Kalokerinos *et al.*, "Neuroticism may not reflect emotional variability," in *Proc. Nat. Acad. Sci. U.S.A.*, vol. 117, no. 17, pp. 9270–9276, 2020.
16. P. Verduyn and S. Lavrijsen, "Which emotions last longest and why: The role of event importance and rumination," *Motivation Emotion*, vol. 39, no. 1, pp. 119–127, 2015.
17. A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors*, vol. 20, no. 3, 2020, Art. no. 592.
18. K. Yang *et al.*, "Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets," *Vis. Comput.*, vol. 37, pp. 1447–1466, 2021.
19. K. Caine, "Local standards for sample size at CHI," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2016, pp. 981–992.
20. P. Klasnja *et al.*, "Microrandomized trials: An experimental design for developing just-in-time adaptive interventions," *Health Psychol.*, vol. 34, no. S, pp. 1220–1228, 2015.

BENJAMIN TAG is a postdoctoral researcher with the School of Computing and Information Systems, University of Melbourne, Melbourne, Australia. His research interests include digital emotion regulation, human cognition, with a special focus on inferring mental state changes from biophysical signals in the wild. Contact him at benjamin.tag@unimelb.edu.au.

JORGE GONCALVES is a senior lecturer with the School of Computing and Information Systems, University of Melbourne, Melbourne, Australia. His research interests include ubiquitous computing, social computing, affective computing, and mobile sensing. Contact him at jorge.goncalves@unimelb.edu.au.

SARAH WEBBER is a research fellow with the School of Computing and Information Systems, University of Melbourne, Melbourne, Australia. Her research interests include the design of digital technologies for social connectedness and well being. Contact her at s.webber@unimelb.edu.au.

PETER KOVAL is a senior lecturer with the Melbourne School of Psychological Sciences, University of Melbourne, Melbourne, Australia. His research focuses on the dynamics of subjective emotional experience and the deliberate regulation of emotion in daily life. Contact him at p.koval@unimelb.edu.au.

VASSILIS KOSTAKOS is a professor of computer science with the University of Melbourne in Australia and the head of the Human-Computer Interaction Group. His research interests focus on ubiquitous computing, human-computer interaction, social computing, and Internet of Things. Contact him at vassilis.kostakos@unimelb.edu.au.