International Journal of Human-Computer Studies 137 (2020) 102396

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Human accuracy in mobile data collection

ARTICLE INFO

Keywords:
Experience sampling method
ESM
Ecological momentary assessment
EMA
Mobile sensing
Mobile crowdsourcing
Self-report

ABSTRACT

The collection of participant data 'in the wild' is widely employed by Human-Computer Interaction researchers. A variety of methods, including experience sampling, mobile crowdsourcing, and citizen science, rely on repeated participant contributions for data collection. Given this strong reliance on participant data, ensuring that the data is complete, reliable, timely, and accurate is key. Although previous work has made significant progress on ensuring that a sufficient amount of data is collected, the accuracy of human contributions has remained underexposed. In this article we argue for an emerging need for an increased focus on this aspect of human-labelled data. The articles published in this special issue demonstrate how a focus on the accuracy of the collected data has implications on all aspects of a study – ranging from study design to the analysis and reporting of results. We put forward a five-point research agenda in which we outline future opportunities in assessing and improving human accuracy in mobile data collection.

1. Introduction

Human-labelled data is at the core of data collection techniques employed in Human Computer Interaction (HCI). The widespread availability of smartphones and other mobile devices, combined with an increasing aspiration to study human behaviour 'in the wild' (*in situ*), has led to an increased application of mobile-based human data contributions. While the use of mobile devices allows for the collection of human-labelled data in authentic settings and surroundings (as collected via *e.g.*, experience sampling method (ESM) (van Berkel et al., 2017a; Csikszentmihalyi and Larson, 1987) or diary studies (Kahneman et al., 2004), self-tracking (Ptakauskaite et al., 2018), crowdsourcing (Hettiachchi et al., 2019), and citizen-science projects (Budde et al., 2017)), humans are notoriously inconsistent in the quality of their contributions. This can be caused by a variety of factors, including a lack of concentration on the task at hand, changes in motivation along the study duration, or simply as the result of study participants missing the required knowledge and skills. As researchers largely rely on human contributions, ensuring a sufficient level of accuracy in these contributions is essential to produce valid and replicable study results. Surprisingly, the accuracy of human labelled submissions through mobile data collection remains an underexplored area. Mobile devices, while occasionally scorned for their distractive effect on their users, offer a largely unexplored territory for researchers to monitor and improve human accuracy during *in situ* data collection.

This special issue focuses specifically on those types of data collection which lack ground truth *in situ*. Examples include, but are not limited to, reporting human emotion or experience, identifying or classifying events, and labelling or annotating observations. The lack of ground-truth forces researchers to consider novel and creative techniques to assess the quality of human submissions, as well as identify new approaches to increase and ensure the accuracy of these submissions. It is interesting to note that, while several of the aforementioned methodologies (*e.g.*, crowdsourcing) have developed method-specific approaches to increase data quality, few of these consider the possibilities and limitations introduced by mobile devices. For example, situated crowdsourcing has made use of public displays to increase human accuracy by tapping into local knowledge (Hosio et al., 2018). Similarly, citizen science has seen an increase towards Open Data, enabling citizens to verify existing data and to contribute to any gaps in the data. In self-report studies, many researchers have embraced mobile devices to present questions, but the use of sensors or novel display options to improve data quality remains underused (van Berkel et al., 2017a).

Given the longitudinal focus of these methodologies, spanning at least multiple days or weeks of data collection in the participant's 'real' environment (van Berkel et al., 2017a), there is a strong overlap in the challenges experienced by researchers as well as the potential for converging solutions used to overcome these challenges. In a categorisation of cross-methodological practises, Van Berkel et al. discuss promising solutions on mobile human contributions that can be applied prior to (*e.g.*, task design, participant training), during (*e.g.*, extrinsic motivators, feedback), and following data collection (*e.g.*, data filtering, response shift) (van Berkel et al., 2018). No previous effort has aimed to bring together the insights from these various methods into contributions that could benefit the wide range of HCI methodologies utilising human labelled data.

We first discuss the general background of *in situ* mobile data collection literature, pointing to the general lack of work focusing on human accuracy. We then introduce the papers included in this special issue, which highlight the diversity of efforts that are require to assess and improve human accuracy in mobile data collection. Finally, we put forward and discuss a research agenda for longitudinal *in situ* human data collection.

https://doi.org/10.1016/j.ijhcs.2020.102396

Available online 13 January 2020

2. Background

Mobile devices have established themselves as a popular research artefact over the past decade (Miller, 2012; Raento et al., 2009). However, despite being used increasingly often for human labelled data collection, the accuracy of human labelled submissions through mobile data collection remains an underexplored area. *In situ* studies, in which mobile devices are used to collect human-labelled data in regular life, are increasingly common (van Berkel et al., 2017a; Miller, 2012). Researchers interested in building automated detection algorithms through Machine Learning rely on mobile human input to establish ground truth on the phenomenon of interest, whereas researchers interested in human experiences typically rely on mobile self-reports of these experiences (see *e.g.* (Di Lascio et al., 2018; Doherty and Doherty, 2018)). This practice has become so wide-spread that *in situ* methods are, for momentary human states like emotional state, considered to be the gold standard to which other data collection methods are compared (Kahneman et al., 2004).

While it is well-known that human accuracy is subject to fluctuation over time and across contexts, common current practice is to consider all mobile human data submissions as both accurate and equal to one another. At the same time, HCI and related disciplines have a long history of studying and improving the accuracy of human data submissions. Crowdsourcing in particular has a rich history of aggregating and filtering submissions to obtain the correct outcomes (Galton, 1907; Hosio et al., 2018), citizen science has explored how scientific tools can be re-appropriated to be usable by non-experts (Budde et al., 2017), and those utilising the diary method have explored novel validations such as the then-test to measure the effects of research instruments on participant answers (Schwartz et al., 2004). These necessary techniques and methods for analysing and improving the accuracy of human data submissions are currently missing for mobile data collection.

3. The special issue

A total of four high-quality papers were accepted for publication in this special issue on Human Accuracy in Mobile Data Collection. The work covered in these articles spans the entire process of mobile data collection, ranging from study design and data collection to analysis of the results. While the majority of papers focus specifically on the smartphone as a device for data collection, one of the articles investigates the use of a wearable device. We summarise the four accepted articles below;

Ellis et al. (2019), in their article "Do smartphone usage scales predict behavior?", compare human accuracy across a range of smartphone usage scales (*e.g.*, which apps were used) against objective measures provided by the participant's personal smartphone. Their results show a poor prediction of objective usage behaviours for the majority of assessment scales. The article concludes by urging researchers to combine participant self-reports with objective measures of behaviour to more reliably study the impact of technology on individuals and society.

Turner et al. (2019), in their article "The influence of concurrent mobile notifications on individual responses", analyse a large dataset of smartphone notifications and identify the characteristics of notification management. In contrast to earlier work, Turnet et al. specifically focus on the coexistence of multiple notifications and its subsequent effect on user behaviour. The authors present considerations on delivering and presenting notifications to study participants in mobile data collection studies, in particular pointing to the fact that study notifications are likely to co-exist with notifications from other applications.

van Berkel et al. (2020), in their article "Overcoming Compliance Bias in Self-Report Studies: A Cross-Study Analysis", analyse a set of recent self-report studies (in the domain of smartphone usage) and discover substantial differences in the quantity of responses between participants. This phenomenon, dubbed 'compliance bias', can result in extensive distortions of study results when ignoring the uneven distribution of participant responses. The authors identify contextual, routine, and study-specific factors that affect participant response rates. Based on these insights, they propose a number of methods to mitigate compliance bias by taking into account the context of respondents.

Giannakos et al. (2020), in their article "Fitbit for Learning: Towards capturing the learning experience using wearable sensing", evaluate the use of wearable devices for assessment of learning processes during class activities. Specifically, the physiological datasets collected via wearable device are compared to self-reported learning outcomes. The authors have conducted *in situ* data collection with a set of participants in a classroom environment. The results are promising as they indicate the potential for a wearable device data streams to correspond to the level of the individuals' learning. A wearable is in this context improving the accuracy of human contributions.

4. Research agenda on human accuracy in mobile data collection

With this special issue we raise awareness of human accuracy in mobile sensing. Rather than considering the reliability of human provided data to be consistent and reliable at all times, we argue that researchers ought to consider fluctuations in human accuracy and the consequences for their subsequent data analysis and results. Methods such as the ESM and crowdsourcing were introduced specifically to increase the reliability and richness of human-labelled data by reducing reliance on an individual's ability to recall past events, to reduce group think, and enable the collecting of multiple data points throughout the day (van Berkel, 2019; Csikszentmihalyi and Larson, 1987).

Recognising the invaluable richness of data collection in authentic settings as enabled by mobile devices, we propose a research agenda in which researchers further embrace these devices to measure and improve the accuracy of human labelled data. Although, like any research agenda, our list is not conclusive – we hope to provide a useful starting point for researchers employing human-labelled data collection *in situ*.

1. *Integration of active and passive sensing.* Active sensing, *i.e.* human labelled data (via ESM), and passive sensing, *i.e.* sensor data, are predominantly collected side-by-side – only to be used collectively during data analysis. This is typically done when comparing the effect of context on participant answers. Integrating these data streams during data collection enables the use of dynamic questionnaire content and presentation. Furthermore, passive sensing can provide a continuous level of consistent data quality not feasible using human data collection alone (Ellis et al., 2019).
2. *Moving beyond time-based notification schedules.* The large majority of studies asking for participant input (e.g., via ESM) is following a randomised or interval-based time schedule (van Berkel et al., 2017a). Although these may initially appear to provide the most equal distribution of participant responses, response rates are often not equally distributed throughout the day – resulting in certain contexts being over-represented in the collected data (van Berkel et al., 2019; Lathia et al., 2013). Building on the aforementioned ambition towards integrated active and passive sensing, contextual

information can inform researchers when participants are likely to be available and able to provide reliable input (van Berkel et al., 2019; Intille et al., 2003). Similarly, keeping track of contexts from which participant input is under-represented can be used to obtain a more contextually diverse dataset.

3. *Explore wider device heterogeneity*. Although smartphones have established themselves as the go-to device for *in situ* data collection from participants, a wider range of devices should be explored. Already in 2003, Intille et al. explore how ubiquitous devices can provide a richer level of context (Intille et al., 2003). In a recent study, Paruthi et al. introduce a custom-made device for *situated* self-reports (Paruthi et al., 2018). Such a device could for example be used to collect human input from a range of people with a low barrier for entry. On a high spectrum of affordability lie studies deploying human data collection via smartwatches, which use is facilitated by the fact that they are mostly on its user's wrist and thus hard to ignore. However, recent work by Ponnada et al. highlights that simply moving questionnaires from the phone to the watch does not increase participant compliance, pointing to the need to consider *how* questionnaires are displayed to participants across device types (Ponnada et al., 2017).
4. *Cross- and peer-validation of contributions*. The online crowdsourcing literature has drawn on the concept of *vox populi* to obtain reliable insights by asking the same question to a group of people (Galton, 1907). In contrast, the use of momentary assessment has typically been focused on individual contributions with a few notable exceptions (see *e.g.*, (van Berkel et al., 2017b; Berrocal and Wac, 2018)). Scaling labelled data to a wider range of contributors enables richer insights and a more verifiable approach to data collection. Berrocal & Wac's 'peer-ceived' momentary assessment *PeerMA* specifically asks participants' peers to assess the state of the individual participants – providing insights in discrepancies between data contributions (Berrocal and Wac, 2018). Similarly, initial work has shown how participants can be used to evaluate the contributions of others *in situ* during data collection (van Berkel et al., 2017b).
5. *Standards for analysis and reporting*. Previous work has highlighted the wide range of inconsistencies and omissions in the reporting of self-report studies (van Berkel et al., 2017a). Consequently, comparing and replicating study outcomes is often unfeasible. In order to progress the methods used in mobile data collection studies, consistent reporting of study design choices is critical. This includes, but is not limited to, questionnaire design, its scheduling, data cleaning and filtering, and response rate calculation. As reported in (van Berkel et al., 2020), even a relatively straightforward outcome as a study's response rate may not report all details when observed on a study-wide level but requires further analysis to inspect for extensive discrepancies between participants.

5. Conclusive remarks

This special issue highlights both novel concerns and opportunities for researchers collecting human-labelled data *in situ*. Although previous work has primarily focused on increasing the number of data points collected per participant, the work presented in this special issue is indicative of the need to study the reliability of contributions obtained through human participants. Although it is often not feasible to replace human-labelled data collection with automated data collection using sensors, carefully considering how and when data can be collected from human participants is key in ensuring a reliable level of human accuracy. It is our expectation that future work will continue this line of work and proposes novel methods and techniques in which human accuracy in mobile data collection can be captured, analysed, and improved in lieu of available ground-truth data. Concretely, we put forward a five-point research agenda to form the foundation of future work into this area.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

van Berkel, N., 2019. Data Quality and Quantity in Mobile Experience Sampling.

van Berkel, N., Budde, M., Wijenayake, S., Goncalves, J., 2018. Improving accuracy in mobile human contributions: an overview. Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. ACM, pp. 594–599.

van Berkel, N., Ferreira, D., Kostakos, V., 2017. The experience sampling method on mobile devices. ACM Comput. Surv. 50 (6), 93:1–93:40. https://doi.org/10.1145/3123988.

van Berkel, N., Goncalves, J., Hosio, S., Kostakos, V., 2017. Gamification of mobile experience sampling improves data quality and quantity. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 1 (3), 107:1–107:21. https://doi.org/10.1145/3130972.

van Berkel, N., Goncalves, J., Hosio, S., Sarsenbayeva, Z., Velloso, E., Kostakos, V., 2020. Overcoming compliance bias in self-report studies: across-study analysis. Int. J. Human-Comput. Stud. 134, 1–12. https://doi.org/10.1016/j.ijhcs.2019.10.003.

van Berkel, N., Goncalves, J., Koval, P., Hosio, S., Dingler, T., Ferreira, D., Kostakos, V., 2019. Context-informed scheduling and analysis: Improving accuracy of mobile self-reports. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 51:1–51:12. https://doi.org/10.1145/3290605.3300281.

Berrocal, A., Wac, K., 2018. Peer-vasive computing: leveraging peers to enhance the accuracy of self-reports in mobile human studies. Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. ACM, New York, NY, USA, pp. 600–605. https://doi.org/10.1145/3267305.3267542.

Budde, M., Schankin, A., Hoffmann, J., Danz, M., Riedel, T., Beigl, M., 2017. Participatory sensing or participatory nonsense?: mitigating the effect of human error on data quality in citizen science. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 1 (3), 39:1–39:23. https://doi.org/10.1145/3131900.

Csikszentmihalyi, M., Larson, R., 1987. Validity and reliability of the Experience-Sampling Method. J. Nerv. Ment. Dis. 175 (9), 526–536.

Di Lascio, E., Gashi, S., Santini, S., 2018. Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2 (3), 103:1–103:21. https://doi.org/10.1145/3264913.

Doherty, K., Doherty, G., 2018. The construal of experience in HCI: understanding self-reports. Int. J. Human-Comput. Stud. 110, 63–74. https://doi.org/10.1016/j.ijhcs.2017.10.006.

Ellis, D.A., Davidson, B.I., Shaw, H., Geyer, K., 2019. Do smartphone usage scales predict behavior? Int. J. Human-Comput. Stud. 130, 86–92. https://doi.org/10.1016/j.ijhcs.2019.05.004.

Galton, F., 1907. Vox populi. Nature 75 (1949), 450–451. https://doi.org/10.1038/075450a0.

Giannakos, M.N., Sharma, K., Papavlasopoulou, S., Pappas, I., Kostakos, V., 2020. Fitbit for learning: towards capturing the learning experience using wearable sensing. Int. J. Human-Comput. Stud. 136, 1–14. https://doi.org/10.1016/j.ijhcs.2019.102384.

Hettiachchi, D., van Berkel, N., Dingler, T., Allison, F., Kostakos, V., Goncalves, J., 2019. Enabling creative crowd work through smart speakers. Proceedings of the CHI workshop on Designing Crowd-powered Creativity Support Systems. pp. 1–5.

Hosio, S., Goncalves, J., van Berkel, N., Klakegg, S., Konomi, S., Kostakos, V., 2018. Facilitating collocated crowdsourcing on situated displays. Human Comput. Interact. 33 (5-6), 335–371. https://doi.org/10.1080/07370024.2017.1344126.

Intille, S.S., Tapia, E.M., Rondoni, J., Beaudin, J., Kukla, C., Agarwal, S., Bao, L., Larson, K., 2003. Tools for studying behavior and technology in natural settings. In: Dey, A.K., Schmidt, A., McCarthy, J.F. (Eds.), UbiComp 2003: Ubiquitous Computing. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 157–174.

Kahneman, D., Krueger, A.B., Schkade, D.A., Schwarz, N., Stone, A.A., 2004. A survey method for characterizing daily life experience: The Day Reconstruction Method. Science 306 (5702), 1776–1780. https://doi.org/10.1126/science.1103572.

Lathia, N., Rachuri, K.K., Mascolo, C., Rentfrow, P.J., 2013. Contextual dissonance: design bias in sensor-based experience sampling methods. Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, New York, NY, USA, pp. 183–192. https://doi.org/10.1145/2493432.2493452.

Miller, G., 2012. The smartphone psychology manifesto. Perspect. Psychol. Sci. 7 (3), 221–237. https://doi.org/10.1177/1745691612441215.

Paruthi, G., Raj, S., Baek, S., Wang, C., Huang, C.-c., Chang, Y.-J., Newman, M.W., 2018. Heed: exploring the design of situated self-reporting devices. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2 (3), 132:1–132:21. https://doi.org/10.1145/3264942.

Ponnada, A., Haynes, C., Maniar, D., Manjourides, J., Intille, S., 2017. Microinteraction ecological momentary assessment response rates: effect of microinteractions or the smartwatch? Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 1 (3), 92:1–92:16. https://doi.org/10.1145/3130957.

Ptakauskaite, N., Cox, A.L., Musolesi, M., Mehrotra, A., Cheshire, J., Garattini, C., 2018. Personal informatics tools benefit from combining automatic and manual data

capture in the long-term. Proceedings of the CHI workshop on Next Steps Towards Long Term Self Tracking. pp. 1–6.
Raento, M., Oulasvirta, A., Eagle, N., 2009. Smartphones: an emerging tool for social scientists. Sociol. Methods Res. 37 (3), 426–454. https://doi.org/10.1177/0049124108330005.
Schwartz, C.E., Sprangers, M.A., Carey, A., Reed, G., 2004. Exploring response shift in longitudinal data. Psychol. Health 19 (1), 51–69. https://doi.org/10.1080/0887044031000118456.
Turner, L.D., Allen, S.M., Whitaker, R.M., 2019. The influence of concurrent mobile notifications on individual responses. Int. J. Human-Comput. Stud. 132, 70–80. https://doi.org/10.1016/j.ijhcs.2019.07.011.

Niels van Berkel[*,a], Jorge Goncalves[b], Katarzyna Wac[c,d], Simo Hosio[e], Anna L. Cox[f]
[a] *Aalborg University, Denmark*
[b] *The University of Melbourne, Australia*
[c] *University of Copenhagen, Denmark*
[d] *University of Geneva, Switzerland*
[e] *University of Oulu, Finland*
[f] *UCLIC, University College London, London, UK*
E-mail address: nielsvanberkel@cs.aau.dk (N. van Berkel).

[*] Corresponding author.