International Journal of Human-Computer Studies 134 (2020) 1–12

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

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

Overcoming compliance bias in self-report studies: A cross-study analysis

Niels van Berkel[a,*], Jorge Goncalves[b], Simo Hosio[c], Zhanna Sarsenbayeva[b], Eduardo Velloso[b], Vassilis Kostakos[b]

[a] University College London, United Kingdom
[b] The University of Melbourne, Australia
[c] University of Oulu, Finland

ARTICLE INFO

Keywords:
Experience Sampling Method
Ecological Momentary Assessment
ESM
Response rate
Completion rate
Compliance
Self-report
Bias

ABSTRACT

A popular methodology used for *in situ* observations is the Experience Sampling Method (ESM), in which participants intermittently answer short questionnaires. We analyse a set of recent ESM studies and find substantial differences in the number of collected responses across participants. These differences amount to 'compliance bias', as the experiences of responsive participants skew the results. Our work develops ways for researchers to ensure the collection of an adequate number of responses across participants. Through a cross-study analysis of ESM studies, we construct a model that describes the effect of contextual, routine, and study-specific factors on participants' response rate. In addition to previous work, which aims to maximise the number of total responses, this work also aims to achieve a more equal distribution of responses between participants. In order to achieve this goal, we analyse which contextual cues can be personalised to achieve a higher response rate. Our results highlight a number of factors that have a strong effect on participants' response rate and can guide the design of future experiments.

1. Introduction

Smartphones have become increasingly popular research tools in the HCI community, thanks to their widespread popularity and powerful sensing capabilities (Raento et al., 2009). Established methods are now being adapted into digital versions to be used with smartphones. One such method is the Experience Sampling Method (ESM), used to collect participant self-report data *in situ* on, for example, experiences or emotions (Larson and Csikszentmihalyi, 1983). In the ESM, participants are asked to answer multiple sets of questions throughout the day, resulting in a rich record of the participant's life regarding the phenomena of interest to the researcher. Therefore, response rate – the percentage of answered questions – is an important metric in ESM studies (van Berkel et al., 2017a). A high response rate ensures a wide spread of results across space and time, increasing the ecological validity of the study results (Hormuth, 1986), "*the occurrence and distribution of stimulus variables in the natural or customary habitat of an individual*" (Hormuth, 1986). Due to the importance of achieving high response rates, researchers have explored several ways to motivate participants. Recent examples include visualising response rates to participants (Hsieh et al., 2008), gamification (van Berkel et al., 2017b), psychological empowerment (Goncalves et al., 2014), and use of micro-incentives per individual questionnaires (Musthag et al., 2011).

In this paper, we explore the effect of contextual factors on participant response rates in ESM studies. First, we present an analysis of previous experiments that together highlight that the response rates of participants in studies that employ the ESM vary substantially. This is similar to what has been dubbed as *compliance bias* in the Health Sciences, and has been observed in studies on adherence to therapeutic protocols (Sackett, 1979). The challenge lies in that the differences in response rates cause more responsive participants to have a larger impact on the conclusions of a study. In essence, we show that researchers should expect their results to be biased by default, since response rates can vary substantially across participants. Second, we explore what causes participants to have reduced response rates and suggest three approaches to mitigate compliance bias in Experience Sampling studies.

To increase our understanding of (non-)responsive participants, we set out to establish the effect of smartphone usage on response rates. Relatively little work has considered how contextual and routine smartphone usage factors affect participant response rates. An important reason for this is that studies are typically carried out with the goal of investigating a well-defined phenomenon, and that the study design naturally stems from this research question. Participants are

* Corresponding author.
E-mail address: n.vanberkel@ucl.ac.uk (N. van Berkel).

https://doi.org/10.1016/j.ijhcs.2019.10.003
Received 19 November 2018; Received in revised form 8 September 2019; Accepted 7 October 2019
Available online 10 October 2019

Table 1
Overview of prior contributions on bias in ESM.

Bias	Effect
Design bias	Potential bias introduced by the study design. Lathia et al. (2013) identify differences in the probability of sampling across different contexts (time- and sensor-based) and find differences in frequency and skewness of data collection triggers.
Novelty bias	Change in behaviour of participant following the introduction of an unknown device van Berkel et al. (2017a). Given today's prevalence of smartphones, usage of participant devices has become more popular.
Observation effect (Hawthorne effect)	Participants may alter their behaviour as they realise they are being observed. Raento et al. (2009) mention how a participant may alter their phone conversations after being reminded that their conversation is recorded at the onset of their call.
Self-selection bias	Participants who sign up for ESM studies may differ from those who do not, although the effect appears limited (Hektner et al., 2007; Mulligan Casey et al., 2000).
Selective nonresponse bias	Participants may choose not to respond if they feel uncomfortable sharing or reflecting on their current experience (Csikszentmihalyi and Larson, 1987).

assumed to comply, when in reality the design of the instrument (*e.g.*, questionnaire items) itself affects compliance, as indicated by previous work in survey research (Lynn, 2001). Further, the literature describes several study-specific constructs which may influence response rate, including the level of rapport between researcher and participant, and the participants' intrinsic motivation (Larson and Csikszentmihalyi, 1983). This makes it challenging to determine whether a participant's decision to respond to an ESM questionnaire is due to the participant's context or due to study-specific constraints. Though recognising the context based on sensor data might help to distinguish these differences, inferring participant context is challenging due to development costs and lack of specialised skills and is therefore often avoided (Raento et al., 2009). Only through a synthesis of multiple studies it is possible to untangle the effect of contextual and routine factors on participant response rates. By adopting a modelling approach, we identify some of the key factors that affect participants' response rates in ESM studies and discuss how the factors manifest.

For these reasons, we have re-analysed data from recent and independent studies within the domain of HCI. Our cross-study analysis models the differences and similarities between these studies, and identifies which factors affect response rates. We found a number of contextual factors (*e.g.*, phone use, screen state) which have a large effect on response rate. Further, we found considerable differences in the effect of contextual factors across participants. Based on our findings, HCI researchers can optimize the scheduling of data collection based on individual participant context. This has the potential to collect data more evenly across participants and ultimately decrease compliance bias.

2. Related work

Since its introduction as a research method in the 1970's, the Experience Sampling Method (ESM) has continuously evolved alongside technological developments. Over time, participant-owned smartphones have become the *de facto* research instrument for ESM studies. Several researchers state that the use of the participant's personal device reduces the novelty effect (Raento et al., 2009), and allows for a more natural *in situ* observation (van Berkel et al., 2017a). Despite the increased complexity of the research instrument, not much literature exists on the effect of the device's context on study results. Wen et al. (2017) presented a meta-analysis on compliance rates among children and adolescents in ESM-like studies. Their analysis of 42 studies revealed mobile phone as the primary device for data collection (20 studies), with a weighted average compliance rate among participants of 78.3%. Interestingly, Wen et al. (2017) reported significant differences in response rates between different sampling frequencies, with average compliance increasing when sampled more often; "*the mean compliance rate was significantly lower in studies that prompted participants 2–3 times (73.5%) or 4–5 times (66.9%) compared with studies with a higher sampling frequency (6+ times: 89.3%)*" (Wen et al., 2017). Their results showed no effect of study duration on compliance rate. Jones et al. (2017) study compliance in ESM-like studies on substance usage. Their results showed an average response rate of 75.1%, with no effect of number of questionnaires or duration of the study. Furthermore, the used device type and the substance under investigation did not significantly affect response rates. In this work we focus specifically on the effect of smartphone context on (differences in) response rate.

2.1. Bias in experience sampling

The ESM and similar self-report methods (*e.g.*, Ecological Momentary Assessment (EMA)) were introduced to achieve two main objectives (Larson and Csikszentmihalyi, 1983). First, a shift from collecting participant data in the laboratory to data collection in the real world. The underlying motivation is to collect data in a more realistic setting, preserving the ecological validity of the study: "*imagery evoked in laboratory studies is not necessarily typical of experience encountered in real-life situations*" (Larson and Csikszentmihalyi, 1983). Second, researchers sought to reduce the reliance on the participants' ability to recall their experiences (van Berkel et al., 2017a). Systematic errors in reasoning, recall, and judgement – known as cognitive biases – negatively impact the validity of collected self-report data (Iida et al., 2012).

These conceptual improvements over previous methods aimed to increase data reliability. Since then, several studies and commentaries have identified biases in the use of the ESM, potentially affecting the reliability of study results if not adequately identified and handled during study design and data analysis. As with any study, biases can be introduced at any stage of the study procedure. Several works have attempted to identify the scale of biases specific to the use of the ESM and have sometimes proposed solutions to circumvent the introduced problem. Table 1 provides a summarized overview of these biases.

Despite the amount of work on improving study-wide response rates, no work has developed approaches to homogenise response rates across individual participants. As such, compliance bias is understudied in ESM – this work sets out to address this bias.

2.2. Improving participant response rates

In the methodologically related diary method, monetary incentives have been shown to increase response rate (Lynn, 2001). Musthag et al. (2011) explored the use of micro-incentives in ESM studies. They found that the amount of the offered micro-incentive did not affect response rate. However, their study did not compare response rates using micro-incentives against a typical 'bulk payment' provided at the end of the experiment.

Several researchers have made suggestions on how to interact with participants to ensure sufficient response rates, usually derived from first-hand experience rather than controlled experimentation. For example, Larson and Csikszentmihalyi (1983) encourage constructing a 'viable research alliance', where the participant is aware of the importance of the study and their role in data collection.

Stone et al. (1991) suggest researchers to contact participants who have missed multiple days of data contributions to encourage them to (re-) commence data contribution.

In a study capturing patient's pain levels, Stone et al. (2003) presented participants with a varying number of daily notifications in order to measure the effect of daily questionnaires on response rate. Participants carried an electronic diary attached to a belt strap, allowing for the completion of questionnaires through the device's touchscreen. In addition to a baseline condition, participants were assigned to receive either 3, 6, or 12 daily questionnaire prompts. Results showed no significant difference in response rates between condition, with all conditions showing high compliance (93.5 – 95.5%). Conner and Reid (2012) presented a study on happiness, in which participants ($N = 162$) received either 1, 3, or 6 daily text messages containing three questions for a duration of 13 days. Participants answered the questionnaires by responding three numbers through text messages. Results show a high average response rate of 96% (data from five participants was removed due to an individual response rate below 75%). Response rates did not differ significantly per condition ($p = .09$), but unfortunately individual response rates per condition were not reported. We note that the aforementioned high response rates are unusual in HCI studies (median response rate 69.9% ((van Berkel et al., 2017a)) and were not completed on smartphones.

Given the importance of the quantity of answers in ESM studies, HCI research has explored new techniques to increase study response rates. Hsieh et al. (2008) demonstrated how displaying participant responses rates directly to participants led to increased response rates in a desktop-based ESM study. Conner and Reid (2012) allowed participants to specify the time of day at which they were able to respond to questionnaires – reducing the intrusiveness of the method. This does however possibly introduce (unwanted) bias as we do not capture data at times which are inconvenient to participants but of interest to the study. Niels van Berkel et al. (2017b) introduced gamification elements in a smartphone-based study and found that participants in the gamified condition contributed significantly more self-reports. However, a potential downside of this approach is the fact that gamification is not suitable for all study subjects or participant demographics (Lefcheck and Freckleton, 2016). Zhang et al. (2016) explored answering questionnaires through 'unlock journaling' (answer a question while unlocking the phone). This method led to a higher frequency of reporting as well as a decreased intrusiveness. A limitation of this method is the single question offered per questionnaire, as well as the input constraints connected to gesture-based unlock input (*e.g.*, no typing). Hernandez et al. (2016) compared the presentation of questionnaires on smartwatches and head-worn devices (Google Glass) to smartphones. They found a 13% increase in response rate on both smartwatches and head-worn devices as compared to smartphones, although the difference was not statistically significant. Further, the study contained a different number of prompts between conditions (Hernandez et al., 2016), making direct comparison challenging. Intille et al. (2016) presented questionnaires through smartwatch-based '*microinteractions*', resulting in increased response rates – again, for both studies, the limited screen real estate hampers possibilities for questionnaire items. Finally, both Pejovic and Musolesi (2014), Stone et al. (1991) explored the use of an intelligent notification system which interrupts the user based on user context. The system resulted in reduced response times, and more favourably received notifications. Our work bears resemblance to previous work (Mehrotra et al., 2015, Pejovic and Musolesi, 2014). We are interested in identifying contextual variables that have an effect of response rate. Unlike previous work, however, we aim to identify these effects across *multiple* studies, not just any single study. As we describe next, previous work has linked a number of contextual variables to response rate in ESM studies.

2.3. Effect of context on response rate

Smartphones are devices that are mobile by design, available to be used anywhere and anytime. As such, the context in which these devices are used constantly changes. In fact, the context of use is so critical that it affects the way in which smartphones are used (Tossell et al., 2012). Pielot et al. (2017) showed how several features related to recent phone usage (*e.g.*, screen state, last application launch) can work as predictors for notification interaction. Contextual factors such as location, mood, and time have been shown to influence mobile browsing needs and behaviour (Lee et al., 2005). These contextual factors exist 'outside' of a user's smartphone, as they are either based on the user's personal or environmental context. Of particular interest for this study are contextual factors that manifest themselves *through* the user's smartphone. For example, incoming notifications, or a battery warning are all outside of the direct control of the user, yet are likely to influence smartphone usage. In the same way that phone usage is affected by these contextual factors, a participant's willingness to respond to ESM questionnaires may be affected by their smartphone context.

We analysed the literature for contextual and behavioural predictors shown to affect smartphone usage and identified whether they have been investigated in the context of the ESM. We summarise these results in Table 2. Our overview indicates that the effect of the majority of aspects shown to influence *phone usage* has not been tested in the context of the ESM. Furthermore, as these variables have been independently identified in mostly separate studies, the relative importance and impact of each one remains unclear. That is why in our work we analyse data from multiple independent studies and try to identify the relative importance of each contextual variable.

3. Data sets

To ensure consistency within our results, we selected three recent studies published in the HCI community which collected contextual data of both answered and unanswered questionnaires. All three studies use the AWARE framework (Ferreira et al., 2015) for collecting smartphone sensor data, ensuring consistency in data collection and formatting. Following data cleaning, our combined dataset consists of 8370 ESM notifications (4408, 1649, and 2313 notifications for Dataset 1, 2, and 3 respectively). In all studies, participants used their personal Android smartphones, providing realistic contextual data on their smartphone usage. Across all studies, participants were full-time students from a single University. Although it can be argued that this biases our results, it simultaneously ensures that we control for *e.g.* cultural differences and shifts in power dimensions between researchers and participants which would occur with a more varied sample of studies. We therefore consider this a reasonable trade-off between participant diversity and the control of external variables. We provide a summary of the three study designs and their respective ESM questionnaires below.

Dataset 1: Usage sessions (van Berkel et al., 2016). This study analysed the intentions of smartphone users upon unlocking their device. The goal of the study was to quantify the commonly held assumption that users who lock and unlock their phone shortly after are continuing the same "session" of interaction. To achieve this goal, the authors employed a combination of human sensing (detailed below) and active sensing (*i.e.*, the collection of smartphone sensor data – most notably phone usage). The sample consisted of 17 participants (13 male, 4 female), average age 26. In this seven-day field study, participants were presented with a pop-up ESM questionnaire each time they unlocked their device. The questionnaire consisted of a single question regarding their phone usage ("*Why did you start using your phone?*") with two predefined answer buttons ("*Continue previous objective*" and "*Start on a new objective*"). Upon completion of the study, participants received a fixed reward (movie voucher) as compensation.

Dataset 2: Battery value (Hosio et al., 2016). This study examined

Table 2
Literature overview of effect of several smartphone usage aspects on smartphone usage and ESM responses.

Aspects	Phone usage	Effect on ESM behaviour
Battery level	"*[users] limit device use for 'emergency situations'*" (Truong et al., 2010)	No information
Recent notification history	"[notifications] lead to a reduction in work productivity, including the resumption time from the interruption back to the primary task and the quality and amount of time available for decision making" (Okoshi, 2015)	No information
Recent app. usage	"a mobile user is likely to use different applications at different locations and access different websites at different times of the day" (Shepard et al., 2011)	No information
Last phone usage	"the majority of phone usage sessions that contain a gap (*i.e.*, phone went to standby mode) consist of only one additional continuous session" (van Berkel et al., 2016)	No information
Time of day (diurnal pattern)	"*Roughly 70% of the users in each dataset have a peak hour usage that is more than twice their mean usage.*" (Falaki et al., 2010)	
"*people are attentive to messages 12.1 h of the day, […], people are more attentive during the evening*" (Dingler and Pielot, 2015)	"*we observe that most participants respond across working hours, and late night / early morning hours are less active*" (van Berkel et al., 2016)	
Weekday / Weekend	"*attentiveness is higher during the week than on the weekend*" (Dingler and Pielot, 2015)	"*Sunday responses were fewer and had a greater variance in happiness*" (Csikszentmihalyi and Hunter, 2003)
Daily apps.	"*users exhibit extraordinarily diverse usage patterns*" (Shepard et al., 2011)	No information
Daily notif.	"[…] sentiment towards a notification varies with the type, completion level and complexity of an ongoing task and the recipient's relationship with the sender." (Mehrotra et al., 2016)	No information

the monetary value that smartphone users assign to their device's battery life, dependent on their context, needs, and current battery level. To quantify the monetary value held by participants, the study incorporated a reverse second-price sealed-bid auction protocol. Upon receiving an ESM question ("*How much money should we pay you for 10% units of your battery?*"), participants bid their desired amount of money for the exchange of 10% of their smartphone's current battery life. Using the aforementioned auction protocol, the participant with the lowest bid won the auction and received the money offered in the second-lowest bid after draining their battery by 10%. The study lasted for eight days and consisted of a total of 20 participants (17 male, 5 female) with an average age of 24. Participants received 13 notifications per day following an hourly schedule (interval contingent). ESM notifications expired after 10 min. Participants received a fixed reward (€50) as well as any money won in the auction.

Dataset 3: Scheduling effects (van Berkel et al., 2018). This study assessed the effect of different ESM schedules on participant response rate and recall accuracy. Participants answered questionnaires consisting of five questions, all focused on assessing their smartphone usage (*e.g.*, "*How many unique applications have you used since 12:00?*"). We consider only the initial question of each questionnaire as opposed to all individual questions (which would register as identical). Participants received up to six notifications a day, with a weekly-changing within-subjects design determining the timing of notifications; randomised (signal contingent), every other hour (interval contingent), or upon unlocking their phone (event contingent, but restricted to one per time-period). ESM notifications were set to expire after 15 min. Data was collected over a period of 3 weeks, with a total of 20 participants (15 male, 5 female), average age of 26. Following completion of the study, participants received a fixed reward (two movie vouchers).

The described datasets contain a mixture of interval-, signal-, and event-contingent notification schedules. These are the three widely-used categories for ESM questionnaire scheduling, all of which are used extensively in the field of Computer Science (van Berkel et al., 2017a). We, therefore, argue that, even though a vast number of scheduling configurations are in use, the presented datasets represent the three primary categories of ESM scheduling configurations.

4. Analysis I – compliance bias in ESM

In Analysis I we investigate the existence and magnitude of compliance bias in ESM. We define compliance bias as the introduction of systematic predilection into collected self-reports as the result of differences in response rate between participants. Because participant responses are typically analysed as a single entity, over- and under-representation of participants in the collected dataset can significantly bias the analysis of results. The literature shows that differences in the number of participant responses are not uncommon. Wiese et al. (2013) found that their most responsive participant completed 358 questionnaires, whereas their least responsive participant completed 29 questionnaires. Similar results are presented by Epp et al. (2011), reporting participant responses ranging from 2 to 219.

For our analysis of compliance bias, in addition to the three previously introduced studies, we also analyse the data from the *StudentLife* study (Wang et al., 2014), a publicly available dataset containing ESM responses. The StudentLife dataset was collected over a 10-week term from 48 University students, assessing their mental health, academic performance, and behaviour. The dataset contains various questionnaires, each triggered according to a different schedule. For consistency, we restrict our analysis to only one of these questionnaires. We select the questionnaire for which most responses were collected. Participants were asked to indicate their current state by selecting a photo amongst a grid of photos depicting various emotions (Photographic Affect Meter). Since this dataset contains only answered ESM questionnaires (no unanswered questionnaires), we are unable to calculate individual response rates per participant and instead consider the absolute number of participant responses.

4.1. Results and discussion

We calculate and visualise the number of responses as collected from individual participants for all four studies (Fig. 1). Substantial differences between individual participants exist in all four studies. This is also indicated by the large standard deviation in the collected number of responses between participants. All four studies are remarkably consistent in the fact that 30% of participants provide roughly 50% of the total collected responses. The 30% of participants with the lowest

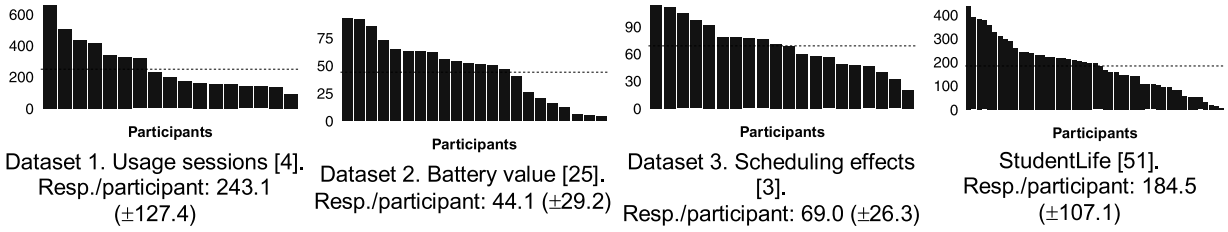

Fig. 1. Number of responses per individual participant across four different studies.

number of responses account for merely 6 – 17% of responses.

These results indicate that there is serious compliance bias in the four analysed studies, despite the fact that aggregated data is used to infer study results (van Berkel et al., 2016). Although these numbers are rarely reported in the literature, given these results and the stark differences reported by Wiese et al. (2013) and Epp et al. (2011), we have reasons to believe that compliance bias may be widespread in self-report studies. Therefore, the current approach of data analysis, in which all responses are considered equal, causes responsive participants to have a larger effect on study conclusions. Following the identified differences between participants, we now present a more detailed analysis of variables leading to participant non-response.

5. Analysis II – modelling compliance bias

Whereas we establish the existence of compliance bias in Analysis I, we now construct a more detailed insight of the effect of (smartphone-based) contextual, routine, and study-specific characteristics on response rate. Establishing the effect of these characteristics could assist researchers in designing more intelligent notification schemes or encourage future studies into the effect of these contextual factors on study participants.

5.1. Method

We combine the three aforementioned datasets into one dataset to investigate compliance bias more systematically than possible in a single study. Here, each row consists of a single ESM question (binary: answered or unanswered). We exclude the StudentLife dataset (included in Analysis I) in this analysis as the dataset is restricted to answered questionnaires only, missing the critical context of unanswered questionnaires.

In addition to participant responses, we collected contextual data in each of the three studies (detailed below). We removed elements in our dataset that were incorrectly marked as a notification (*e.g.*, applications which use notifications to display audio controls, resulting in a large number of incorrect 'notifications'). Data was recorded upon a state change for each respective data element, combined with both a unique random ID per participant and a timestamp. From the contextual usage data contained in our dataset, we calculated the following predictors:

Contextual predictors. Contextual predictors reflect the *in situ* participant state at the moment of ESM notification arrival. These predictors are constantly changing, and jointly affected by smartphone usage and the timing of ESM notifications.

- **Battery level**: Percentage of remaining battery life.
- **Charging status**: Smartphone's charging status, either 'charging' or 'not charging'.
- **Screen status**: State of the screen, either 'on' or 'off'.
- **Last phone usage**: Time in minutes since the phone was last used. Values above 60 min are clamped to 60 min.
- **Recent notifications**. Number of notifications received by the participant in the preceding 15 min.
- **Weekend/weekday**. We assigned either 'weekend' or 'weekday' based on the day the ESM notification was sent.
- **Time of day**. Questionnaire arrival in morning [05:00 – 11:59], afternoon [12:00 – 16:59], or evening [17:00 – 04:59].

Routine predictors. Routine predictors characterise the smartphone usage behaviour across the duration of the study.

- **Daily application usage**. Number of average unique applications a participant uses per day.
- **Daily notifications**. Number of average notifications a participant receives per day.

Study-specific characteristics. We specify one key ESM study configuration parameter, daily ESM notifications, which is not identical between the three studies. Other study-specific characteristics (*e.g.*, questionnaire input type, study duration) are not suitable to be included in the model. Though we expect these parameters to affect a participant's willingness to respond, including these parameters in the model would reduce its reliability. If included, these parameters would assess all potential differences between the studies rather than one single variable. For example, using 'input type' as a parameter, its value would be 'popup' for Dataset 1 and 'notification' for the two remaining datasets. As such, the effects ascribed to this coefficient would model any differences between the studies rather than actually measuring the effect of input type.

- **Daily ESM notifications**. Average daily ESM notifications per dataset are 41.2 (± 18.39), 11.0 (± 1.49), and 5.4 (± 0.28) for Dataset 1, 2, and 3 respectively.

Criteria were selected based on recommendations from the related work as well as their shared availability across datasets. Rows for which one of the above predictors were missing were removed from the dataset to ensure reliable model construction. Following this, 8185 rows of ESM questions remained (185 were removed).

6. Results

Given our binary predictor value of ESM responses – answered or unanswered – the distribution of participant responses is binomial. Rather than considering all of the collected data to originate from one coherent source, we have to account for the fact that both the number of observations per study as well as the number of observations per participant differ from one another. We therefore construct a generalised linear mixed-effects model (GLMM) with a binomial distribution describing the effect of the aforementioned predictors on compliance. We specify both 'participant ID' and 'study ID' as random effects in the model. As a result, our model treats participant- and study-specific variations as a nuisance term (*i.e.*, it is not of direct interest but should be controlled for). The use of generalised linear-mixed effects models for binary responses is an increasingly common paradigm applied to data collected in a longitudinal setting (Zhang et al., 2011). We create one general model based on the data of all participants. Then, we create separate models for the 30% of participants with the highest response

rate (per individual study, roughly half of the collected responses) and the 30% participants with the lowest response rate (again per individual study). We label these two separate groups as high-compliance and low-compliance.

Predictors that were not significant ($p < .05$) across any of the three models were discarded. This was true for two variables: charging status and weekend/weekday. The three models were recalculated excluding these two variables. We report the outcomes of each model (including odds ratio and confidence interval per predictor) in Table 4. The general model is statistically significant (χ^2 (12) = 354.5, $p < .001$) and explains 8% of variance in answering ESM notifications (*Marginal* $R^2 = 0.08$, *Conditional* $R^2 = .38$). The model for the most responsive participants is also statistically significant (χ^2 (12) = 97.1, $p < .001$) and explained 15% of variance in answering ESM notifications (*Marginal* $R^2 = 0.15$, *Conditional* $R^2 = 0.28$). Finally, the model for the least responsive participants is also statistically significant (χ^2 (12) = 141.8, $p < .001$) and explained 28% of variance in answering ESM notifications (*Marginal* $R^2 = 0.28$, *Conditional* $R^2 = 0.39$).

When assessing human behaviour we do not expect our model to include all relevant predictors (Cohen et al., 2002). R^2 describes the goodness-of-fit of the model and was calculated using the R package '*piecewiseSEM*' (Lefcheck and Freckleton, 2016), as based on the method introduced in (Nakagawa and Schielzeth, 2013) specifically aimed at GLMM. Marginal R^2 describes the variance explained solely by fixed factors, whereas the Conditional R^2 describes the variance explained by the combined fixed and random factors (in our case 'participant ID' and 'study ID') (Nakagawa and Schielzeth, 2013).

We present the confusion matrices for our classification models in Table 3. Further, Table 3 reports the accuracy of the three models as well as their respective sensitivity and specificity values. Sensitivity describes the percentage of cases in which the model predicts 'Answered' among actual 'Answered' cases (also described as 'recall' or 'true positive rate'). Specificity describes the percentage of cases in which the model predicts 'Unanswered' among the total number of actual 'Unanswered' cases.

To ensure that our chosen predictors are not in fact measuring the same phenomenon, we check for the existence of multicollinearity. Multicollinearity indicates whether one of the predictors in the model can be linearly predicted from a combination of any of the other predictors, reducing the validity of individual predictors. All our predictors report a variance inflation factor between 1.01 and 1.73, well below the often-used threshold of 5 or 10 to detect (severe) multicollinearity (Hair et al., 2010). We also check for linearity between predictors and dependent variable in the three constructed models. The analysis conducted with the '*caret*' package (Kuhn, 2017) reveals no linear combinations between any of the predictors and the dependent variable.

We plot the effect of each predictor in the model in Fig. 2. This shows the effect of, for example, the smartphone's screen status on ESM response rate, assuming that all other predictors remain unchanged. We note that the Y-axis is not consistent between graphs: its range is, in essence, an indicator of the effect size for each predictor. As shown in Fig. 2, we observe that:

- Participants are more likely to respond when they receive a notification while their screen is off; *ergo*, they are not using their phone (stronger for low-compliance participants).
- Participants are more likely to respond in the morning, with a small decline in the transition to afternoon and evening. The opposite holds for high-compliance participants.
- Participants are more likely to respond when they have recently used the phone. This effect applies to all participants but is reduced for high-compliance participants.
- High-compliance participants increase their likelihood of response when experiencing a high number of incoming notifications, whereas low-compliance participants are less likely to respond when receiving many notifications.
- A lower battery level results in a higher response rate. This effect is almost negligible for high-compliance participants.
- Number of daily notifications (averaged over study period) received by participants shows a discrepancy between participant clusters. Low-compliance participants with a high number of daily notifications are less responsive than their peers, whereas the opposite holds for high-compliance participants with a high number of daily notifications.
- More daily notifications lead to a higher response rate.
- In general, participants who use a high number of unique daily applications (averaged over the study period) are more likely to respond to ESM notifications.

Next, we take a more detailed look at a few of these predictors. This is because, in the case of continuous variables, the effect of a predictor is not always as straightforward as portrayed in Fig. 2. Therefore, we calculate the density plots for predictors which show an effect of context on response rate.

Effect of Last Phone Usage: We calculated the effect of last phone usage against ESM response rate, as displayed in Fig. 3 (excluding the values above 60 min.). A participant who has recently used their device is considerably more likely to answer an ESM notification, whereas a longer time since the last device usage results in a higher percentage of unanswered questionnaires. This effect is strongest for participants with a low response rate.

Effect of Recent Notifications: As shown in Fig. 4, the effect of recent smartphone notifications is non-linear. Participants with either zero or a large number of recent notifications are more likely to answer (given the density of answered over unanswered at these values). This effect is most limited for participants with a low response rate.

Effect of Battery Level: Fig. 5 shows the effect of different battery levels on ESM response rate. For many ranges of battery levels, the chance of an ESM being answered or unanswered is equal. Only as the battery level approaches 100% does the chance of an ESM going unanswered increase. This effect is strongest for high-compliance participants.

Effect of Daily Notifications: The effect of the daily number of notifications on response rate shows an inconsistent pattern (Fig. 6). We see that for both the general model and the model describing the low-compliance participants, receiving many daily notifications is likely to lead to a reduction in response rate. The opposite holds for high-compliance participants, where a low number of daily notifications actually results in a lower response rate than their (high-responsive) peers.

Effect of Daily Application Usage: Participants who use a low number of daily applications are less likely to respond than those who use a higher number of applications. As shown in Fig. 7, this effect holds across all three clusters, but is most prominent in the general model and the model assessing the most responsive participants. Participants with a low number of daily applications used are maybe more likely to use a secondary device (*e.g.*, laptop, tablet) for their browsing and communication needs, or are simply less likely to use their smartphone throughout their day – thereby failing to notice incoming ESM notifications.

Table 3
Confusion matrices for the three models.

	General model		High-compliance		Low-compliance	
	Predicted		Predicted		Predicted	
	Unansw.	Answ.	Unansw.	Answ.	Unansw.	Answ.
Unasw.	998	1190	14	230	705	398
Answ.	440	5557	8	2232	255	1070
Accuracy	80.1%		90.4%		73.1%	
Sensitivity	92.7%		99.6%		80.8%	
Specificity	45.6%		5.7%		63.9%	
Baseline	73.3%		90.2%		54.6%	

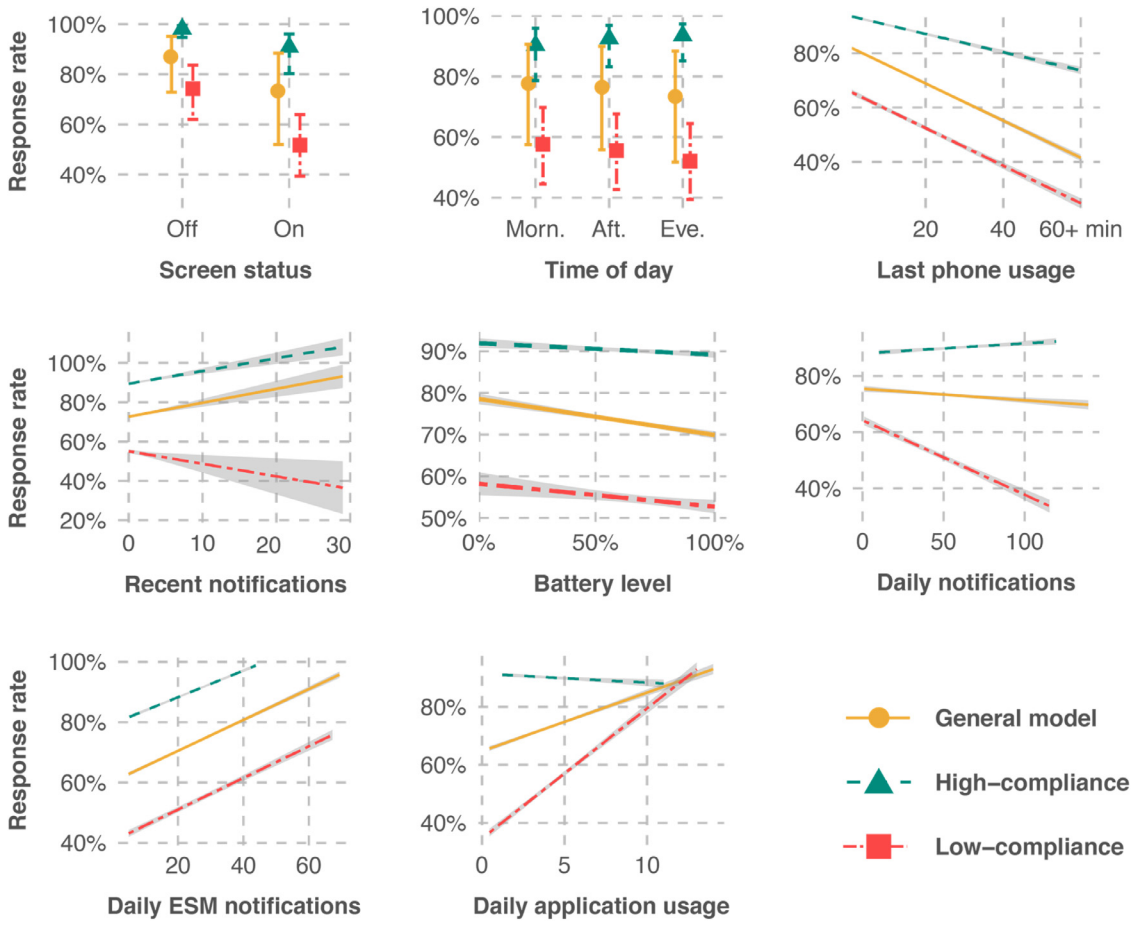

Fig. 2. Effect of individual predictors, with all other predictors remaining equal.

7. Discussion

Researchers employing the ESM rely on their participants' responses to make inferences about the topic being studied. Although a strength of the ESM is the repetitive collection of self-reports over time in the participants' natural environments, the same participant responses also become the *Achilles Heel* of the method. Several biases related to ESM studies have been previously identified, including design bias (Lathia et al., 2013), novelty bias (van Berkel et al., 2017a), and self-selection bias (Hektner et al., 2007, Mulligan Casey et al., 2000), among others (see Table 1). Further, a low number of participant responses has been discussed as detrimental to the reliability of ESM studies (Larson and Csikszentmihalyi, 1983). Recent discussion in Psychology and other disciplines have also critiqued the current bias towards '*WEIRD*' (well-educated and originating from industrialised, rich, and democratic countries) and university participant samples (Henrich et al., 2010). Such samples are also common among Experience Sampling studies, as seen in the studies included in our study, as well as most other recent ESM publications (*e.g.*, (Wang et al., 2014, Yang et al., 2016)). These participants are not a realistic representation of society, limiting study conclusions to the studied population.

Our results show that there is a considerable difference in the number of contributions between participants (Fig. 1). This phenomenon, which has not previously been systematically summarized and explored, appears to be widespread (*e.g.* (Epp et al., 2011, Wang et al., 2014, Yang et al., 2016)). We term this difference in responses between participants 'compliance bias'. To overcome these differences between participants, previous work recommends the removal of participants with a low response rate (Niels van Berkel et al., 2017a). For example, Epp et al. (2011) removed those with less than 50 responses, and Yang et al. (2016) removed participants who completed less than half of the assigned tasks. However, we argue this approach is problematic and at the very least not optimal. Not only is such a selection of acceptable

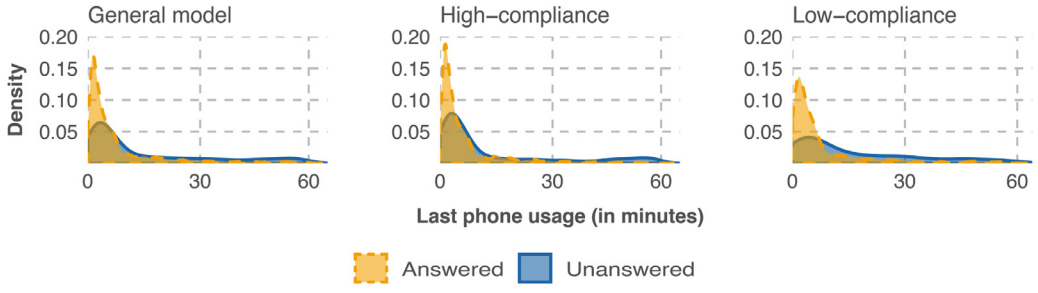

Fig. 3. Conditional density plot of last phone usage and ESM response rate.

Fig. 4. Conditional density plot of recent notifications and ESM response rate.

response numbers arbitrary, it also removes the experiences of a set of participants from the analysed data – biasing results to a smaller and more engaged participant sample. Furthermore, we note that the use of multilevel modelling, in which missing data is compensated through mean estimation, will lead to biases if collected data is not missing 'at random' (Enders, 2010). As we show in Fig. 1, collected self-report data is not equally distributed between participants and the missing data (*i.e.*, self-reports) are therefore considered 'Missing Not At Random'. We now interpret the presented results, offer suggestions on reducing compliance bias, and explain how compliance bias relates to previously identified biases in self-report studies.

7.1. Modelling results

From the total of 10 predictors analysed in this study, 8 were found to be significant in at least one of the models. From this, we conclude that different contextual factors significantly affect participant response rates. Other than time-related contextual variables, the selected predictors have not been previously explored in the context of ESM response rate (see Table 2). Questionnaire scheduling techniques in ESM studies often assume that a participant's ability to respond to a questionnaire is consistent both across sampling time and within the participant sample. Our results, on the contrary, show that participant response rates are not consistent across time and context, and that significant differences exist between participants. For example, we find that study participants are less likely to answer after leaving their phone unused for a long duration, or when their phone is currently being used. We also find that the response rate of high-compliance participants is even higher during the evening, while the opposite holds for low-compliance participants. Similarly, high-compliance participants are more responsive when having just received a higher number of notifications, whereas low-compliance respondents become even less responsive under the same condition.

Contextual Predictors: The context of a participant who receives an ESM questionnaire affects the participant's ability and willingness to reply. This is why, for example, many ESM studies avoid sending notifications during the night (Hosio et al., 2016). Pejovic and Musolesi (2014) identify a variety of contextual factors related to ESM responsiveness, but do not consider smartphone usage context. Our results show that various smartphone usage factors significantly affect response rate.

Participants' *screen status*, either 'on' or 'off', has a significant effect on response rate. Interestingly, we find a higher response rate when the screen of the device is turned off. This indicates that when participants are currently already using their device for a specific purpose, their focus on that task tends to take precedence over answering an ESM notification. In addition, we found that *last phone usage* is also a significant indicator of ESM response rate. We observe that an ESM questionnaire is more likely to be answered if the participant has recently used their phone. A large time gap may indicate that the participant is busy with other activities or simply not in the vicinity of the device. This is in line with previous work on smartphone usage (van Berkel et al., 2016, Dey et al., 2011), but has not been considered in light of ESM studies. We infer that a participant is more likely to respond to an ESM notification when the phone is on the periphery of attention but not in active use. These results hold for both high- and low-compliance participants.

Time of day, categorised as morning, afternoon, and evening, also had a significant effect on ESM response rate. For most participants, notifications sent in the morning or afternoon hours are more likely to be answered than those in the evening hours. As a result, if a researcher is to send notifications equally distributed over the course of day, the collected response data will be skewed towards the morning and afternoon. The opposite holds for high-responsive participants, who are actually responding slightly more frequently in the evening hours. In their analysis of smartphone user attentiveness to notifications, Dingler & Pielot found that levels of attentiveness are higher during evenings (Dingler and Pielot, 2015). Our results show that, in contrast, study participants are less responsive to ESM questionnaires during evening hours. This highlights a difference between notifications originating from user-installed applications and questionnaires.

Previous work has identified that notifications are experienced as interrupting (Cutrell et al., 2001), unfavourable (Pejovic and Musolesi, 2014), and annoying following a high frequency of notifications (Church and de Oliveira, 2013). As such, it can be argued that interrupting participants with a questionnaire following a high number of notifications may lead to frustrated participants and dismissed ESM notifications. Conversely, incoming notifications may indicate active usage of the device and therefore indicate a suitable moment for the interruption. Our results show that the effect of *recent notification history*

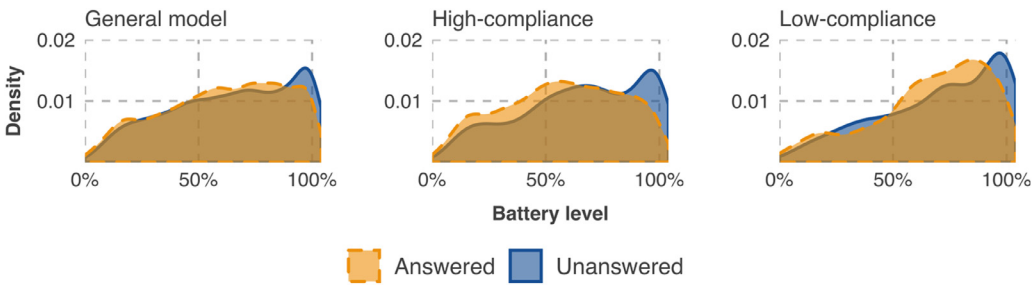

Fig. 5. Conditional density plot of battery level and ESM response rate.

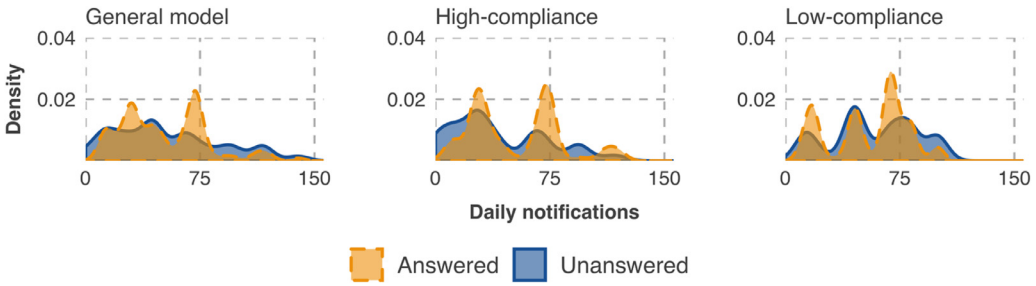

Fig. 6. Conditional density plot of daily notifications and ESM response rate.

differs between high-compliance participants and low-compliance participants. Among low-compliance participants we see a decrease in their responsiveness when receiving a large number of notifications.

Finally, a participant's *smartphone's battery life* was found to have a significant effect on participant response rate. According to the generated model, a higher battery life results in a lower response rate, although the size of the effect is limited. As shown in Fig. 5, the effect of battery life is not linear but instead primarily prominent at the 'edge-value' of battery life. The chance of ESM notifications going unanswered is highest in the case of full or almost-full battery life. A battery level of 100% is an indicator that the phone is likely to be connected to a power source and charging. Despite the significant effect of battery level, the effect size is limited (Table 4) – explaining the exclusion of charging status as a predictor.

Earlier work on the ESM recommends participant sampling across both weekdays and weekends, to capture a diverse set of experiences (Hektner et al., 2007). Csikszentmihalyi and Hunter (2003) report fewer participant responses on Sundays. Our results imply that, for our data, no systematic bias was introduced by sampling across both weekdays and weekends. In the three investigated studies, participants used their personal devices, whereas the studies completed by Csikszentmihalyi and Hunter (2003) required participants to carry both a beeper and pen-and-paper – making it more likely for participants to forget their recording device or experience it as burdensome.

Routine Predictors: Smartphone usage differs drastically between people. Thus, smartphone users can be categorised by their usage habits (Xu et al., 2011). Here, we investigated two routine smartphone usage parameters: application usage and notifications received.

We find that participants that use a larger number of *unique daily applications* are more responsive to ESM notifications. This effect holds true specifically among low-compliance participants. Further, we find differences between high- and low-compliance participants in the effect of *daily notifications*. Among high-compliance participants, response rate is even higher for those with a high number of daily notifications. For low-compliance participants, a high number of daily notifications is an indicator of even lower response rates. Whereas high-compliance participants seem to be able and willing to manage ESM requests among incoming smartphone notifications, the opposite holds for low-compliance participants.

Study-specific Predictors: The effect of methodological configurations in Experience Sampling on response rate has typically been discussed from the perspective of participant strain. For example, Consolvo and Walker (2003) suggest designing questionnaires that minimise participants' burden by avoiding open-ended questions or reducing the number of notifications. Mehrotra et al. (2015) note that "*users may fail to respond honestly, or may even ignore the questionnaire prompts if they perceive the study as too burdensome.*"

The sampled studies contain differences in the amount of *daily ESM notifications* presented to participants. The number of daily questionnaires is often a balance between the researchers' desire for a rich data set *versus* participant burden (Zirkel et al., 2015). Our results show that an increase in the number of ESM notifications leads to an increase in response rate. This effect is highly consistent between all three models and is in line with a previous meta-analysis by Wen et al. (2017). However, we do note that the sampled studies contained differences in questionnaire structure. In the 'Usage Sessions' study, participants received a *popup message* with a binary question upon device unlock. While this may have been experienced as interrupting, the required effort was lower when compared to the two other studies (unlocking the phone and opening the notification). We control for this difference by specifying study (and participant) as random factors in our model. Although there may be a variety of other study-specific predictors causing differences in response rate (*e.g.*, study topic), our data does not allow for analysing the effect of such factors and we therefore do not pose any claims regarding other study-specific factors. Nevertheless, our analysis shows a multitude of factors that do significantly impact participant response rates – explaining up to 28% of variance among low-compliance participants, the most critical participant group when considering compliance bias. Given the nature of our cross-study analysis we can expect these factors to hold for future ESM studies.

7.2. Mitigating compliance bias in ESM

Our analysis of multiple recent ESM studies shows that compliance bias exists, and that it has a substantial effect on the interpretation of study results. From the analysed studies, half of contributions were made by 30% of participations, whereas the 30% with the lowest number of contributions were responsible for only 6–17% of responses (see Fig. 1). A good first step towards addressing compliance bias in

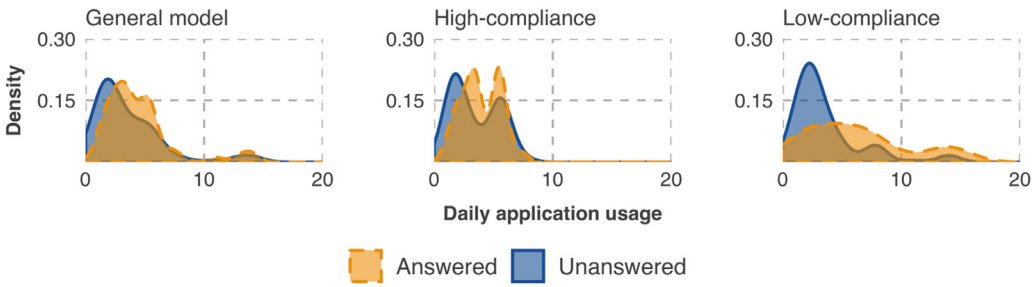

Fig. 7. Conditional density plot of daily unique application usage and ESM response rate.

Table 4
Effect of model factors on response rate (odds ratio, confidence interval, and p-value).

	General model			High-compliance			Low-compliance		
	OR	CI	p	OR	CI	p	OR	CI	p
(Intercept)	7.52	2.0–28.0	<0.003**	15.26	3.2–72.1	<0.001***	0.6	0.2–1.6	0.274
Screen status – On	0.40	0.3–0.5	<0.001***	0.20	0.1–0.3	<0.001***	0.4	0.3–0.5	<0.001***
Time of day – Evening	0.79	0.7–0.9	0.004**	1.54	1.0–2.3	0.033*	0.8	0.6–1.0	0.079
Time of day – Afternoon	0.93	0.8–1.1	0.362	1.33	0.9–1.9	0.128	0.9	0.7–1.2	0.510
Last phone usage	0.98	1.0–1.0	<0.001***	0.98	1.0–1.0	<0.001***	1.0	1.0–1.0	<0.001***
Recent notifications	0.99	1.0–1.0	0.048*	1.03	1.0–1.1	0.435	1.0	1.0–1.0	0.501
Battery level	1.00	1.0–1.0	0.012*	1.00	1.0–1.0	0.353	1.0	1.0–1.0	0.836
Daily notifications	1.00	1.0–1.0	0.799	1.00	1.0–1.0	0.970	1.0	1.0–1.0	0.049*
Daily ESM notifications	1.01	0.9–1.1	0.392	1.04	1.0–1.1	0.092	1.0	1.0–1.1	0.042*
Daily application usage	1.06	1.0–1.2	0.240	1.08	0.9–1.2	0.271	1.2	1.1–1.4	0.003**

ESM is therefore to recognize that this bias exists, and to urge researchers to analyse and report the difference in collected participant responses. We note that for studies which achieve a high overall response rate (*e.g.*, >95%), the likelihood of compliance bias is reduced as individual responses rates are more closely aligned. However, we do note that this is rare, as indicated not only by our presented analysis but also by extensive literature reviews in the area of Addiction (avg. response rate of 75.1% across 126 studies (Jones et al., 2017)), children and adolescents (avg. response rate of 78.3% across 42 studies (Wen et al., 2017)), and Computer Science (avg. response rate of 69.6% across 110 studies (van Berkel et al., 2017a)).

Compliance bias can be reduced by either oversampling those participants that are likely to be underrepresented in the collected dataset, or by undersampling highly responsive participants. While discarding responses following data collection is the most straightforward approach to undersampling, it is also detrimental to many of the core aspects of the ESM. Typically, the more responses are collected, the richer the analysis that can be performed by the researcher (Csikszentmihalyi and Larson, 1987). Therefore, researchers are likely to want to focus on *increasing the amount* of collected responses of low-respondents rather than decreasing responses in an attempt to homogenise response rates. One approach to increase the number of collected responses among low-respondents is to extend the study duration for low-compliance participants. This will result in a higher number of absolute responses, but is often impractical due to research agreements, will delay the analysis, and conflicts with other practical study arrangements. We therefore introduce and discuss three different strategies to homogenise participant responses.

Contextual optimisation. Low-responsive participants are not only more likely to ignore the questionnaire when their device is in active use, but also when it is requesting their attention (recent notifications (Table 4). This shows that high-responsive participants are more willing to overcome these 'barriers' to answer a question or retrieving their smartphone after not having used it for a long time. To overcome this difference, one strategy to balance response rates is to personalise questionnaire scheduling to those contexts in which a participant is likely to respond. In a way, this follows prior work by Church et al. (2014) who allowed participants to specify the times of day at which they would receive ESM questionnaires. We argue that future work should explore the personalisation of ESM sampling beyond time-restrictions, including contextual factors. These personalised schedules adapt their schedule to interrupt participants at moments in which a response is likely. This builds on earlier recommendations (*e.g.*, (Mehrotra et al., 2015, Pejovic and Musolesi, 2014)). Personalised scheduling based on participant context will result in a more equal response rate between participants, reducing the effect of compliance bias. The results we have presented here provide a starting point for which contextual variables to consider in this approach.

Increase answer opportunities. Participants with a low response rate respond to ESM questionnaires less frequently. To collect an equal number of responses between participants, participants with a lower response rate in comparison to their peers can be sampled more frequently. Naturally, this will mean that the study orchestrators will have to find a way to adjust the rate on the go, while the study is progressing. With the use of mobile phones and modern ESM tools, this is luckily quite feasible already. While this will result in a further decrease of their individual response rate (as well as global study response rate), it allows the researcher to homogenise the absolute number of collected responses between participants. This approach does not consider participant context and could lead to increased annoyance among participants – potentially having a detrimental effect on the quality of collected responses. One way to reduce this annoyance is to oversample only in those contexts where the participants miss the questionnaire (expired questionnaires) rather than actively ignore it (dismisses questionnaires).

Context-based oversampling and undersampling. A different approach to addressing compliance bias is to focus on the breath of contexts covered in collected responses rather than the total number of collected responses. For example, when a participant is biased towards answering questionnaires when they are messaging friends, future questionnaires should undersample during this context, and extra questionnaires should be scheduled during contexts which lack data. While this is likely to result in a lower number of overall responses, the difference in the number of collected responses between participants is likely to be reduced and the ecological validity of the results is increased.

Naturally, these suggestions are not always applicable. For instance, Church et al.'s study design (Church et al., 2014) prevented the collection of participant responses at moments that participants deemed inconvenient. Similarly, the sampling of experiences at moments at which a participant is more likely to respond will reduce the variety of contexts captured during the study. The effect of this on study results is highly dependent on the study question and the larger goal of the study. When focusing on, for example, a participant's stress levels, it is critical to sample responses across a variety of (smartphone usage) contexts, including when a participant is experiencing a large number of notifications or an alarmingly low battery level. This is why we urge researchers to strongly consider which parameters to include in personalised ESM models based on established literature.

Finally, when addressing compliance bias in self-report studies, it is important to consider the potential interplay with other biases. Attempts at reducing compliance bias can result in the introduction or reinforcement of other biases. First, sampling participants during contexts in which they are likely to be more responsive will limit the number of contexts in which participants answer questionnaires (Lathia et al., 2013). Simultaneously, smartphone usage behaviour may differ between participants, which could result in participants receiving a significantly different number of ESM questionnaires if following a single notification scheme for all.

Second, the recruitment of university students will likely lead to different smartphone usage patterns than displayed by other samples of the population. Therefore, the use of sampling techniques tested among

university student participants may not necessarily be successful among other parts of the population.

Third, novelty bias can affect compliance bias over time. It is well known that participant engagement drops over time, resulting in a reducing number of responses (Stone et al., 1991). As a result, study results are likely to be biased towards the initial sampling period where participants still experience the study as interesting.

Our results show that, in addition to study-specific predictors, a combination of contextual and routine smartphone usage predictors as measured from participant smartphones significantly affect ESM response rate. Further, we show that these predictors affect high- and low-compliance participants differently. These findings can be useful for future ESM studies and their question scheduling, as a high response rate is crucial to securing ecological validity of the study (Hormuth, 1986). We position our work as one of the contributions of HCI to the continuous development of the ESM – with contributions ranging from novel input methods to more intelligent interruption mechanisms.

7.3. Limitations and future work

Our analysis of ESM study data contains several limitations. Given the nature of the cross-study analysis reported in Analysis II, we are limited to the use of datasets which contain a substantial overlap in the types of sensor data collected. Even though the publicly available ESM dataset StudentLife (Wang et al., 2014) contains a wide range of contextual data, the dataset is limited to answered questionnaires and contextual features do not fully overlap. In the presented analysis, the context of both answered and unanswered ESM questionnaires is essential. Consequently, Analysis II is limited to three unique studies given the high additional costs of running additional studies. An ideal number of studies for a cross-study analysis does not exist. Future work should identify in more detail the effect of individual contextual and routine parameters on response rate, and how this newfound knowledge can drive notification schedules while retaining ecological validity. Furthermore, we note that the contextual factors considered in this study are dictated by the information provided by (battery-conservative) smartphone sensors. While this facilitates the future application of our findings, it limits the contextual richness of our analysis. For example, factors such as the company of the participant (*e.g.*, alone, with friends) or their current activity (*e.g.*, work, sport) are not considered in this study. Although we believe these contextual factors would be useful, the analysed datasets did not contain the information necessary to derive these factors.

The population in the three datasets (Analysis II) is limited to the student population of one University. While the similar background of participants ensures that the identified differences are not due to differences in student-researcher power-relationships, motivation, or other cultural aspects, the study sample is limited in variety. We note that our results do not necessarily replicate among other target groups. Furthermore, our results may not generalise to other study designs, including, but not limited to, different disciplines, alternative input techniques, or compensation structures (*e.g.*, micro-incentives). However, we note that the studies included in this analysis are in line with the HCI literature with regards to study duration, sample size, and response rate (Niels van Berkel et al., 2017a). Furthermore, the studied population sample (young adults, well-educated) differs from the general population, including in their use of smartphones. However, our results are directly applicable to the many researchers relying on student populations. We hope that future work can replicate the method demonstrated in this work with a more diverse sample of participants and study designs.

8. Conclusion

Through an analysis of four recent ESM studies, including the publicly available StudentLife dataset (Wang et al., 2014), we identify substantial differences in the number of collected responses between study participants. This introduces a considerable bias in the analysis of study results, as the experiences of high-compliance participants have a disproportionate effect during study analysis. We term this compliance bias. Following the establishment of this phenomenon in ESM studies, we perform a cross-study analysis of three recent studies investigating the effect of contextual, routine, and study-specific factors on participants' response rate. We demonstrate that a number of contextual factors are highly predictive of participant response rate. Prominently, sampling when the phone has recently been used and when the screen is off (indicating that it is no longer being used) is likely to lead to higher response rates. Further, our results show that there are considerable differences in the effect of these factors on responsive and irresponsive participants. We propose different scheduling techniques that could help mitigate compliance bias by homogenising the number of participant responses. Finally, we discuss potential concerns with regards to other biases.

Declaration of Competing Interest

None.

References

van Berkel, N., Ferreira, D., Kostakos, V., 2017a. The experience sampling methods on mobile devices. ACM Comput. Surv. 50 (6). https://doi.org/10.1145/3123988. 93:91-93:40.

van Berkel, N., Goncalves, J., Hosio, S., Kostakos, V., 2017b. Gamification of mobile experience sampling improves data quality and quantity. Proceed. ACM Interact. Mob. Wearable Ubiquitous Technol. 1 (3). https://doi.org/10.1145/3130972. 107:101-107:121.

van Berkel, N., Goncalves, J., Lovén, L., Ferreira, D., Hosio, S., Kostakos, V., 2018. Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports. Int. J. Hum. Comput. Stud. https://doi.org/10.1016/j.ijhcs.2018.12.002. press.

van Berkel, N., Luo, C., Anagnostopoulos, T., Ferreira, D., Goncalves, J., Hosio, S., Kostakos, V., 2016. A systematic assessment of smartphone usage gaps. Proceed. ACM Confer. Human Factor. Comput. Syst. 4711–4721. https://doi.org/10.1145/2858036.2858348.

Church, K., Cherubini, M., Oliver, N., 2014. A large-scale study of daily information needs captured in situ. ACM Trans. Comput.-Human Interact. 21 (2), 1–46. https://doi.org/10.1145/2552193.

Church, K., de Oliveira, R., 2013. What's up with whatsapp?: Comparing mobile instant messaging behaviors with traditional SMS. In: *Proceedings of the ACM Conference on Human-Computer Interaction with Mobile Devices and Services*, ACM, pp. 352–361. https://doi.org/10.1145/2493190.2493225.

Cohen, J., Cohen, P., G. West, S., S. Aiken, L., 2002. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Routledge.

Conner, T.S., Reid, K.A., 2012. Effects of intensive mobile happiness reporting in daily life. Soc. Psychol. Personal. Sci. 3 (3), 315–323. https://doi.org/10.1177/1948550611419677.

Consolvo, S., Walker, M., 2003. Using the experience sampling method to evaluate ubicomp applications. IEEE Pervas. Comput. 2 (2), 24–31. https://doi.org/10.1109/MPRV.2003.1203750.

Csikszentmihalyi, M., Hunter, J., 2003. Happiness in everyday life: the uses of experience sampling. J. Happiness Stud. 4 (2), 185–199. https://doi.org/10.1023/a:1024409732742.

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

Cutrell, E., Czerwinski, M., Horvitz, E., 2001. Notification, disruption, and memory: effects of messaging interruptions on memory and performance. Human-Comput. Interact. – INTERACT 263–269.

Dey, A.K., Wac, K., Ferreira, D., Tassini, K., Hong, J.H., Ramos, J., 2011. Getting closer: an empirical investigation of the proximity of user to their smart phones. In: *Proceedings of the international Conference on Ubiquitous Computing*, ACM, pp. 163–172. https://doi.org/10.1145/2030112.2030135.

Dingler, T., Pielot, M., 2015. I'll be there for you: quantifying attentiveness towards mobile messaging. In: *Proceedings of the ACM Conference on Human-Computer Interaction with Mobile Devices and Services*, ACM, pp. 1–5. https://doi.org/10.1145/2785830.2785840.

Enders, C.K, 2010. Applied Missing Data Analysis. Guilford Press.

Epp, C., Lippold, M., L. Mandryk, R., 2011. Identifying emotional states using keystroke dynamics. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, pp. 715–724. https://doi.org/10.1145/1978942.1979046.

Falaki, H., Mahajan, R., Kandula, S., Lymberopoulos, D., Govindan, R., Estrin, D., 2010. Diversity in smartphone usage. In: *Proceedings of the ACM Conference on Human-Computer Interaction with Mobile Devices and Services*, ACM, pp. 179–194. https://doi.

org/10.1145/1814433.1814453.
Ferreira, D., Kostakos, V., K. Dey, A., 2015. AWARE: mobile context instrumentation framework. Front. ICT 2 (6), 1–9. https://doi.org/10.3389/fict.2015.00006.
Goncalves, J., Kostakos, V., Karapanos, E., Barreto, M., Camacho, T., Tomasic, A., Zimmerman, J., 2014. Citizen motivation on the go: the role of psychological empowerment. Interact. Comput. 26 (3), 196–207. https://doi.org/10.1093/iwc/iwt035.
Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., 2010. Multivariate Data Analysis. Pearson.
Hektner, J.M., Schmidt, J.A., Csikszentmihalyi, M., 2007. Experience Sampling Method: Measuring the Quality of Everyday Life. Sage.
Henrich, J., Heine, S.J, Norenzayan, A., 2010. Most people are not weird. Nature 466 (7302), 29.
Hernandez, J., McDuff, D., Infante, C., Maes, P., Quigley, K., Picard, R., 2016. Wearable ESM: differences in the experience sampling method across wearable devices. In: *Proceedings of the ACM Conference on Human-Computer Interaction with Mobile Devices and Services*, ACM, pp. 195–205. https://doi.org/10.1145/2935334.2935340.
Hormuth, S.E., 1986. The sampling of experiences *in situ*. J. Pers. https://doi.org/10.1111/j.1467-6494.1986.tb00395.x.
Hosio, S., Ferreira, D., Goncalves, J., van Berkel, N., Luo, C., Ahmed, M., Flores, H., Kostakos, V., 2016. Monetary assessment of battery life on smartphones. Proceedings of the ACM Conference on Human Factors in Computing Systems 1869–1880. https://doi.org/10.1145/2858036.2858285.
Hsieh, G., Li, I., Dey, A., Forlizzi, J., E. Hudson, S., 2008. Using visualizations to increase compliance in experience sampling. In: *Proceedings of the ACM International Conference on Ubiquitous Computing*, ACM, pp. 164–167. https://doi.org/10.1145/1409635.1409657.
Iida, M., Shrout, P.E., Laurenceau, J.-P.P., Bolger, N., 2012. Using diary methods in psychological research. In: Cooper, H. (Ed.), APA Handbook of Research Methods in Psychology. American Psychological Association, pp. 277–305.
Intille, S., Haynes, C., Maniar, D., Ponnada, A., Manjourides, J., 2016. μEMA: microinteraction-based ecological momentary assessment (EMA) using a smartwatch. In: *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM,, pp. 1124–1128. https://doi.org/10.1145/2971648.2971717.
Jones, A., Remmerswaal, D., Verveer, I., Robinson, E., Ingmar, H, Franken, .A.., K. Fred Wen, C., Field, M., 2017. Compliance with ecological momentary assessment protocols in substance users: a meta-analysis. Addiction. https://doi.org/10.1111/add.14503. Online first.
M. Kuhn. 2017. caret: Classification and Regression Training (R package version 6.0-78). In: Astrophysics Source Code Library.
Larson, R., Csikszentmihalyi, M., 1983. The experience sampling method. In: Csikszentmihalyi, M. (Ed.), Flow and the Foundations of Positive Psychology. Wiley Jossey-Bass, pp. 41–56.
Lathia, N., Rachuri, K.K., Mascolo, C., Rentfrow, P.J., 2013. Contextual dissonance: design bias in sensor-based experience sampling methods. In: *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM, pp. 183–192. https://doi.org/10.1145/2493432.2493452.
Lee, I., Kim, J., Kim, J., 2005. Use contexts for the mobile internet: a longitudinal study monitoring actual use of mobile internet services. Int. J. Hum.–Comput. Interact. 18 (3), 269–292. https://doi.org/10.1207/s15327590ijhc1803_2.
Lefcheck, J.S., Freckleton, R., 2016. piecewiseSEM: piecewise structural equation modelling in R for ecology, evolution, and systematics. Method. Ecol. Evolut. 7 (5), 573–579. https://doi.org/10.1111/2041-210X.12512.
Lynn, P., 2001. The impact of incentives on response rates to personal interview surveys: role and perceptions of interviewers. Int. J. Public Opin. Res. 13 (3), 326–336. https://doi.org/10.1093/ijpor/13.3.326.
Mehrotra, A., Pejovic, V., Vermeulen, J., Hendley, R., Musolesi, M., 2016. My phone and me: understanding people's receptivity to mobile notifications. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, pp. 1021–1032. https://doi.org/10.1145/2858036.2858566.
Mehrotra, A., Vermeulen, J., Pejovic, V., Musolesi, M., 2015. Ask, but don't interrupt: the case for interruptibility-aware mobile experience sampling. In: *EA Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, ACM, pp. 723–732. https://doi.org/10.1145/2800835.2804397.
Mulligan Casey, B., Schneider, B., Wolfe, R., 2000. NBER Working Paper No. t0265.
Musthag, M., Raij, A., Ganesan, D., Kumar, S., Shiffman, S., 2011. Exploring micro-incentive strategies for participant compensation in high-burden studies. In: *Proceedings of the ACM International Conference on Ubiquitous Computing*, ACM, pp. 435–444. https://doi.org/10.1145/2030112.2030170.
Nakagawa, S., Schielzeth, H., 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. Method. Ecol. Evolut. 4 (2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x.
Okoshi, T., 2015. Attelia: reducing user's cognitive load due to interruptive notifications on smart phones. In *pervasive computing and communications (PerCom)*. In: *2015 IEEE International Conference on*, IEEE, https://doi.org/10.1109/PERCOM.2015.7146515.
Pejovic, V., Musolesi, M., 2014. InterruptMe: designing intelligent prompting mechanisms for pervasive applications. In: *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM, pp. 897–908. https://doi.org/10.1145/2632048.2632062.
Pielot, M., Cardoso, B., Katevas, K., Serrà, J., Matic, A., Oliver, N., 2017. Beyond interruptibility: predicting opportune moments to engage mobile phone users. Proceed. ACM Interact. Mob. Wearable Ubiquitous Technol. 1 (3) 91:25-91:91.
Raento, M., Oulasvirta, A., Eagle, N., 2009. Smartphones: an emerging tool for social scientists. Sociol. Method. Res. 37 (3), 426–454. https://doi.org/10.1177/0049124108330005.
Sackett, D.L., 1979. Bias in analytic research. J. Chronic. Dis. 32 (1), 51–63. https://doi.org/10.1016/0021-9681(79)90012-2.
Shepard, C., Rahmati, A., Tossell, C., Zhong, L., Kortum, P., 2011. LiveLab: measuring wireless networks and smartphone users in the field. ACM SIGMETRICS Perform. Evaluat. Rev. 38 (3), 15. https://doi.org/10.1145/1925019.1925023.
Stone, A.A., Broderick, J.E., Schwartz, J.E., Shiffman, S., Litcher-Kelly, L., Calvanese, P., 2003. Intensive momentary reporting of pain with an electronic diary: reactivity, compliance, and patient satisfaction. Pain 104 (1), 343–351. https://doi.org/10.1016/S0304-3959(03)00040-X.
Stone, A., Kessler, R., Haythomthwatte, J., 1991. Measuring daily events and experiences: decisions for the researcher. J. Pers. 59 (3), 575–607. https://doi.org/10.1111/j.1467-6494.1991.tb00260.x.
Tossell, C., Kortum, P., Rahmati, A., Shepard, C., Zhong, L., 2012. Characterizing web use on smartphones. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, pp. 2769–2778. https://doi.org/10.1145/2207676.2208676.
Truong, K.N., Kientz, J.A., Sohn, T., Rosenzweig, A., Fonville, A., Smith, T., 2010. The design and evaluation of a task-centered battery interface. In: *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, ACM, pp. 341–350. https://doi.org/10.1145/1864349.1864400.
Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., T. Campbell, A., 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM, pp. 3–14. https://doi.org/10.1145/2632048.2632054.
Wen, C.K.F., Schneider, S., Stone, A.A., Spruijt-Metz, D., 2017. Compliance with mobile ecological momentary assessment protocols in children and adolescents: a systematic review and meta-analysis. J. Med. Internet Res. 19 (4), e132. https://doi.org/10.2196/jmir.6641.
Wiese, J., Saponas, T.S., Bernheim Brush, A.J., 2013. Phoneprioception: enabling mobile phones to infer where they are kept. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, pp. 2157–2166. https://doi.org/10.1145/2470654.2481296.
Xu, Q., Erman, J., Gerber, A., Mao, Z., Pang, J., Venkataraman, S., 2011. Identifying diverse usage behaviors of smartphone apps. In: *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, ACM, pp. 329–344. https://doi.org/10.1145/2068816.2068847.
Yang, Y., Clark, G.D., Lindqvist, J., Oulasvirta, A., 2016. Free-Form gesture authentication in the wild. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, pp. 3722–3735. https://doi.org/10.1145/2858036.2858270.
Zhang, H., Lu, N., Feng, C., W. Thurston, S., Xia, Y., M. Tu., X., 2011. On fitting generalized linear mixed-effects models for binary responses using different statistical packages. Stat. Med. 30 (20), 2562–2572. https://doi.org/10.1002/sim.4265.
Zhang, X., Pina, L.R., Fogarty, J., 2016. Examining unlock journaling with diaries and reminders for in situ self-report in health and wellness. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, pp. 5658–5664. https://doi.org/10.1145/2858036.2858360.
Zirkel, S., Garcia, J.A., Murphy, M.C., 2015. Experience-Sampling research methods and their potential for education research. Educ. Research. 44 (1), 7–16. https://doi.org/10.3102/0013189X14566879.