

Exploring the State-of-Receptivity for mHealth Interventions

FLORIAN KÜNZLER*, ETH Zürich
VARUN MISHRA*†, Dartmouth College
JAN-NIKLAS KRAMER, University of St. Gallen
DAVID KOTZ, Dartmouth College
ELGAR FLEISCH, ETH Zürich and University of St. Gallen
TOBIAS KOWATSCH, ETH Zürich and University of St. Gallen

Recent advancements in sensing techniques for mHealth applications have led to successful development and deployments of several mHealth intervention designs, including Just-In-Time Adaptive Interventions (JITAI). JITAIs show great potential because they aim to provide the right type and amount of support, at the right time. Timing the delivery of a JITAI such as the user is receptive and available to engage with the intervention is crucial for a JITAI to succeed. Although previous research has extensively explored the role of context in users' responsiveness towards generic phone notifications, it has not been thoroughly explored for actual mHealth interventions. In this work, we explore the factors affecting users' *receptivity* towards JITAIs. To this end, we conducted a study with 189 participants, over a period of 6 weeks, where participants received interventions to improve their physical activity levels. The interventions were delivered by a chatbot-based digital coach – Ally – which was available on Android and iOS platforms.

We define several metrics to gauge receptivity towards the interventions, and found that (1) several participant-specific characteristics (age, personality, and device type) show significant associations with the overall participant receptivity over the course of the study, and that (2) several contextual factors (day/time, phone battery, phone interaction, physical activity, and location), show significant associations with the participant receptivity, in-the-moment. Further, we explore the relationship between the effectiveness of the intervention and receptivity towards those interventions; based on our analyses, we speculate that being receptive to interventions helped participants achieve physical activity goals, which in turn motivated participants to be more receptive to future interventions. Finally, we build machine-learning models to detect receptivity, with up to a 77% increase in F1 score over a biased random classifier.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**; • **Applied computing** → *Health care information systems*.

Additional Key Words and Phrases: Receptivity, Intervention, Interruption, Mobile Health, Engagement

ACM Reference Format:

Florian Künzler, Varun Mishra, Jan-Niklas Kramer, David Kotz, Elgar Fleisch, and Tobias Kowatsch. 2019. Exploring the State-of-Receptivity for mHealth Interventions. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 140 (December 2019), 27 pages. <https://doi.org/10.1145/3369805>

*Both authors contributed equally to this research.

†Corresponding author

Authors' addresses: Florian Künzler, ETH Zürich; Varun Mishra, Dartmouth College, varun@cs.dartmouth.edu; Jan-Niklas Kramer, University of St. Gallen; David Kotz, Dartmouth College; Elgar Fleisch, ETH Zürich, University of St. Gallen; Tobias Kowatsch, ETH Zürich, University of St. Gallen.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2019 Copyright held by the owner/author(s).

2474-9567/2019/12-ART140

<https://doi.org/10.1145/3369805>

1 INTRODUCTION

Advances in mobile, wearable and embedded sensing technology have created new opportunities for research into physical, mental, social, emotional, and behavioral health. Researchers have used this ‘mHealth’ (mobile health) technology to study stress, depression, mood, personality change, schizophrenia, physical activity and addictive behavior, among other things. From a deeper understanding of health-related behavior, their ultimate goal is to develop effective interventions that can lead people toward healthier behavior.

One novel intervention design is the just-in-time adaptive intervention (JITAI), which aims to provide the right type and amount of support, at the right time, adapted as-needed for the individuals’ internal and contextual state [28, 29]. JITAI-like interventions have successfully been delivered to patients with alcohol disorder [13], schizophrenia [6] and other behavioral conditions [5, 7, 16, 37]. For many such conditions, interventions are most effective if delivered at the right time. But what is the ‘right’ time? First, they should be delivered at the onset of the relevant behavior [2], or a psychological or contextual state that might lead to that behavior. Second, they should be delivered at a moment when the person will actually see, absorb, and engage with the intervention. Consider an example: Alice has anxiety disorder and will soon need to give a public presentation. Her smartwatch includes technology that can detect and anticipate stress, and an interventional app that helps her manage stressful situations. The app should deliver its calming recommendations prior to the presentation – and perhaps afterward – but certainly not on-screen in the middle of her presentation!

Most prior mHealth research has focused on the first factor: inferring psychological, behavioral, or contextual state [15, 24, 25]. In this paper we focus on the second factor, i.e., determining the times or conditions when the person would be *receptive* to an intervention, i.e., in a *state of receptivity*. As defined by Nahum-Shani et al., *receptivity* is the person’s ability to receive, process, and use the support (intervention) provided [28]. To this end, we built a JITAI system to promote physical activity, available on Android and iOS devices, that delivered an actual behavior-change intervention aimed at increasing the participant’s daily step count. We enrolled a representative population of 189 participants, over a period of 6 weeks in free-living conditions. Participants received occasional notifications that encouraged them to engage in dialog with the digital coach Ally, a chatbot motivating individuals to increase their physical activity.

The ubiquitous-computing research community has explored the related concept of *interruptibility*, and *engagement*. *Interruptibility* is defined as a person’s ability to be interrupted by a notification by immediately taking action to open and view the notification content [21], whereas *engagement* refers to the involvement into something that attracts and holds a person’s attention [30, 34]. In the domain of smartphone notifications, engagement can be thought of what happens after a person is interrupted. For example, a person viewing a recently delivered SMS notification can be at a state of interruptibility, however only when the person replies to the text message will s/he be engaged. Hence, a person can be interruptible but not engaged. Similarly, a person can be engaged and not interruptible, e.g., self initiated engagement tasks, like playing games on smartphone. This has led to a variety of works in the field of ‘interruptibility management’, ‘context-aware notifications’, and ‘smart notifications’. This body of work aims to identify times or conditions where users are most likely to accept or view phone-based notifications [12], so that the system can deliver notifications with lower user burden [11] and increase likelihood that users view the notification content [27]. To model user interruptibility and engagement, researchers have explored context like location [27, 34], activity [31, 34], boredom [35], personality [22], phone interaction [35] and more. Although such research may be effective at delivering routine phone notifications, none of those methods are sufficient for timing notifications intended as behavioral interventions. For most phone-based notifications, e.g., messages, email, and social media, the recipient is interruptible if they derive an immediate and clear benefit – either a transfer of important information or the opportunity for an amusing diversion. Behavioral interventions are different. Even participants who are motivated for behavioral change may prefer to ignore or postpone specific intervention activities... particularly if they require challenging behaviors.

Although some of these prior studies have looked at contextual factors related to users' interruptibility, virtually no studies have investigated states of receptivity for mHealth based interventions. Receptivity to interventions, as defined by Nahum-Shani et al, is the person's ability to receive, process, and use the support (intervention) provided [29]. To compare with interruptibility and engagement – in the context of smartphone notifications – receptivity may be (loosely) conceptualized to encompass the combination of interruptibility (*willingness to receive* an intervention), engagement (*receive* the intervention), and the person's subjective perception of the intervention provided (*process* and *use* the intervention). Sarker et al. [38] made one such attempt; instead of delivering actual interventions, however, they used surveys and Ecological Momentary Assessment (EMA) questions that they claimed would have similar engagement levels as interventions.¹ The authors also provided monetary incentives to encourage the participants to respond to EMA prompts, which might have biased the participants' receptivity. Further, in addition to contextual features, the authors used the content of the participants' response to the EMA, over the course of the study, in their method for gauging receptivity.

The results of this paper **contribute** an understanding of the following research questions:

- How do intrinsic factors (i.e., participant characteristics like, age or personality) relate to receptivity?
- How do contextual factors (e.g., location, time of day) relate to receptivity?
- Does participant receptivity differ on iOS and Android platforms?
- How does the outcome of the intervention relate to receptivity? That is, do participants achieve higher positive outcomes by being more receptive to the intervention... and when they achieve positive outcomes, do they become more receptive?
- Can we build machine-learning models to infer when participants are receptive towards interventions? Even with fairly simple classifiers (e.g., Random Forests), we observed significant improvements in F1-score of up to 77% over a naive classifier baseline.

We begin with a review of related work on interruptibility and receptivity. We then describe the details of our study and research methods. We then conduct an extensive analysis of the results from the study and address the research questions listed above. We conclude with a discussion of outstanding issues and provide an outlook on future work.

2 RELATED WORK

Mobile application developers commonly rely on *push notifications* to capture the user's attention. While some push notifications are necessary and welcomed by the user, others are seen as interruptive. Many researchers have investigated the effect of various features on the users' *interruptibility*. In this section we discuss techniques and factors explored by researchers that affect user's interruptibility. Further, we also discuss prior work that has attempted to capture users' *receptivity* in the domain of mHealth.

Most research in the field of interruptibility management of mobile push notifications is focused on application-independent systems, where they treat all types of notifications to be similar, and observe the factors affecting response to those notifications. Several researchers have investigated the correlation between the user's context and interruptibility such as the time of the day [4, 20, 33, 35]. While most of these studies found that time is a useful feature to determine a user's state of interruptibility [33, 35], some research found that time is not a helpful input [20, 40]. Other research has considered the user's location [21, 34, 38]; again, the results are inconsistent. Some research found location is an important predictor of interruptibility [33, 35, 38], while others found the opposite [21]. Researchers have investigated other context, such as the use of Bluetooth information to determine if the user is in a social context [33], communication (SMS and Call logs) [11, 35], phone battery information [34] and Wi-Fi [33, 35]. These contexts have shown to indicate interruptible moments to deliver notifications.

¹Although EMA is common in behavioral studies, including intervention studies, EMA is means to *collect* self-reported information and not specifically aimed at *changing* behavioral or mental condition.

In addition to the user's context, researchers have investigated the user's physical activity and have found significant correlation between activity and interruptibility, e.g. [14, 21, 32, 38]. Okoshi et al. found that breakpoints between two different activities, e.g. switching from walking to standing, are opportune moments to trigger push notifications [32], while Mehrotra et al. found certain activities (e.g., biking) to be less opportune than others (e.g., walking) [21].

In addition to context and physical activities, researchers have investigated the correlation between interruptibility and an individual's personality traits and psychological state. Sarker et al. found that features derived from participants' stress levels (calculated from physiological sensors) are significant in predicting interruptibility [38]. They also found that happy or energetic participants are more available, compared to stressed ones. Personality traits were also found to be significant predictors of how fast someone replies to a prompt [22]. Mehrotra et al. identified neuroticism and extroversion as discriminating features of how fast people respond to push notifications [22]. The results from all the above mentioned works, however, have not been generalizable: different studies report different and often conflicting results [19]. This could be because of small sample sizes, and homogeneity amongst the participants in a study, who usually belong to a similar demographic group, e.g., students or office workers.

A few researchers have also looked beyond interruptibility to user *engagement* with content. Pielot et al. conducted a study with 337 participants, where they observed whether and how participants engaged with a set of eight different types of content after clicking on a notification [34]. They build machine-learning models to infer whether a participant would engage with the notification content, which performed 66.6% better than the baseline classifier. However, they did provide monetary incentives to participants to use the app for a specific period of time and to respond to a minimum number of notifications, which the authors agree could create a bias on how participants interacted with the notification prompts.

Some research looked at interruptions in the domain of mHealth. Sarker et al. explored discriminative features and further built machine-learning models to identify receptivity to just-in-time interventions (JITI) with 30 participants [38]. The authors, however, did not use an actual intervention during the study, instead they sent out EMA prompts, claiming interaction with EMA prompts would be similar to interaction with interventions. The authors used passive smartphone data, a wearable sensor suite with ECG and respiration data, and self-reported EMA responses to determine moments of receptivity. The authors also provided monetary incentives to the participants for completing the EMA prompts, which could have biased how the participants responded to those prompts. Mishra et al. also explored the role of contextual breakpoints on responsiveness to EMA [23].

In other work, Morrison et al. conducted a study with a mobile stress-management intervention [27]. They randomly assigned 77 participants to one of three groups, each receiving interventions via push notifications at different timings. One group received notifications occasionally (not every day), one group daily, and one group intelligently using sensors (using insights gained from previous research). They found that the daily and the intelligent group responded to significantly more prompts than the occasional group. There was, however, no significant difference in responsiveness between the intelligent prompts and the frequent prompts groups. Further, they call for further research into health behavior-change intervention using different sensor combinations.

Finally, all the works mentioned above have looked at *interruptions* or *receptivity* using only one mobile platform – Android. No work has studied the substantial population of iOS users, although it is possible that UI differences may affect interruptibility or receptivity.

Our current work attempts to address several open questions and limitations in prior work. Instead of a single source of participants, e.g., university students, we conducted a large study ($n = 189$) representative of the population in the German-speaking part of Switzerland. We focus on some initial survey data collected before the study started and passively collected data from smartphone sensors, across the two major smartphone operating systems – Android and iOS. Finally, we evaluate user *receptivity* – response and engagement with intervention – by delivering participants an actual intervention aimed at a positive outcome (increased physical activity),

instead of proxies like EMA [38] or survey tasks [34]. Finally, in addition to user-specific traits and contextual features, we also explore how the effectiveness of the intervention (i.e., achieving daily step goal) relates to users' receptivity of future interventions and vice versa, something that has not been explored before because of the use of EMA or other proxies.

3 METHOD

In this section, we discuss our study goals and design, the smartphone app used in the study (Ally) and the different types of data collected, followed by the study description and some statistics about the data collected over the course of the study. We also define the different metrics used for gauging receptivity, and finally discuss our analysis plan.

3.1 Study Goals and Intervention Design

One of the major goals for the study was to explore participants' state-of-receptivity when participants were receiving real interventions, instead of asking EMA questions, or engaging participants in other proxy tasks. To this end, we collaborated with psychologists and developed **Ally**, a chatbot for delivering interventions designed to promote physical activity by targeting to increase the average step count. We followed a study protocol similar to the one by Kramer et al. [18].

The app, available on both iOS and Android, is based on the MobileCoach platform, developed by Filler et al. [10, 17]. The app consists of a chat interface where the users get different intervention messages and can respond from a set of choices. The app also consists of a dashboard with an overview of the previous week's performance, along with the achieved number of steps and the step goal for that day. Next to the step data, users could view the distance walked and calories burnt. The chat interface and the dashboard interface of the app are shown in Fig 1.

After enrolling in the study and downloading the app, the study started with a 10-day baseline assessment phase, during which the app collected step data from the user and used this information to tailor personalized step goals.

After the baseline phase, a six-week intervention period followed. The app calculated a personalized goal using historical step data, using the 60th percentile of the last 9 days as a new goal for the day, and updated daily. The interventions components in our study were chat-based conversational messages from Ally, the coach, and pre-defined answer options from the participant. The starting message of these conversations were delivered to every user at random times within certain time frames. We had four different conversations in our study: (1) goal setting, which was delivered between 8-10 a.m. and set the step goals for the day, (2) self-monitoring prompt, used to inform the participants about the day's progress and encourage them to complete the goal, was delivered randomly to 50% of the participants between 10 a.m. and 6 p.m. everyday, (3) goal achievement prompt, to inform participants if they completed their goal and encourage them to complete future goals, was delivered at 8 p.m. everyday, and finally (4) the weekly planning intervention, to help the participants overcome any barriers to physical activity, was randomly delivered to 50% of the participants once a week.

In the study, the participants were randomly assigned to one of three financial incentive groups, for the entire duration of the study. One group received a cash bonus of 1 CHF (approximately 1 USD) for achieving a daily step

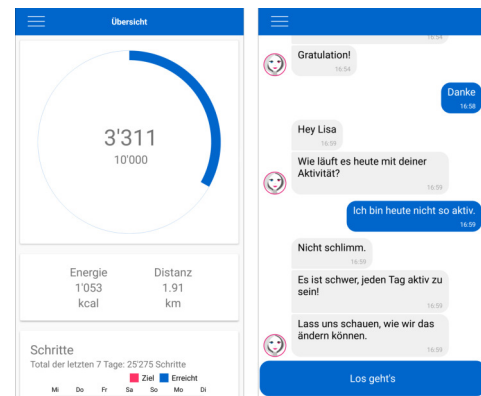


Fig. 1. The Ally app: The screenshot on the left shows the dashboard with a step count, calories burnt, and distance walked. The second screenshot shows the chat interface.

goal,² a second group donated the reward to charity, and a third group acted as control group, without receiving any financial reward. The participants could optionally fill out two web-based questionnaires (at enrollment and post-study), for which they were compensated with 10 CHF.

Fit within the JITAI framework: As proposed by Nahum-Shani et al. [29], JITAI has six key elements: a distal outcome, proximal outcomes, decision points, intervention options, tailoring variables, and decision rules.³

For our study, the *distal outcome* is increase in physical activity (average step count). The weekly and daily proportions of participant days that step goals are achieved during the intervention period are the primary proximal outcomes, along with other secondary proximal outcomes. The *decision point* is the pre-randomized time in which the message is sent to the participants (between 8-10 am, 10 am-6 pm, or 8pm), the *tailoring variable* is the cumulative step count till that time, the *intervention options* are possible support options, in either setting a step goal, or checking up on the current step count mid-day, or checking the step count at the end of the day, or providing an encouragement message. The *decision rules* link intervention options to the tailoring variable. Depending on the accumulated step count, relative to the daily step goal, a different intervention message was sent to either congratulate on the progress or to encourage the participant for further improvement.

How receptivity informs efficacious JITAI: In the JITAI framework defined by Nahum-Shani et al., *receptivity* is considered as a *tailoring variable*. At a given *decision point*, the *decision rules* can look up the tailoring variable of receptivity (along with other tailoring variables) to decide which *intervention option* to use. Depending on the design, the intervention options could include “no intervention” as an option. Hence, receptivity will help determine if, what and how an intervention will be delivered at a given decision point.

Assumptions about receptivity: In our intervention design, all conversations⁴ start with a greeting, which is a generic message like "Hello [name of person]", or "Good Evening [name of person]"; only if the participant responds to the greeting does the coach start sending the intervention messages based on the decision rules and the tailoring variable. To be consistent, whenever we discuss receptivity to a message, we are referring to the *initiating message*, i.e., the start of conversation message from Ally. Since participants respond to the initiating message without looking at the actual intervention, we assume that response to and subsequent engagement with the initial message is independent of the actual intervention type (e.g. length or content).

According to Nahum-Shani et al., receptivity is a function of both internal (e.g., mood) and contextual (e.g., location) factors [28]. Our work, however, is limited to only include the contextual part of receptivity; while time-varying personal factors pertaining to the person’s psychological, physiological or social state are also important factors, they are not evaluated in this work.

Hence, in our work, when we refer to ‘receptivity’, we mean only the ‘contextual’ part. Even with our assumptions and limitations about receptivity, we believe our work is a major step forward from previous works in being the first to study receptivity to interventions and to better define decision rules and efficiently use the decision points in JITAIs.

²Unlike previous works, the cash bonus of 1 CHF was awarded for completion of the *step goal only*, irrespective of how many times they respond or engage with interventions. However, to be thorough we conducted a one-way ANOVA to observe whether there was any significant effect of financial incentive type and the receptivity metrics over the course of the study. We observed that financial incentive had no significant effect on any of the receptivity metrics defined in Section 3.3.2. Hence, we believe that the interaction with the messages were naturalistic and free of financial bias.

³According to Nahum-Shani et al., *distal outcome* is the ultimate goal the intervention is intended to achieve, *proximal outcomes* are short-term goals the intervention is intended to achieve, *decision points* are points in time at which an intervention decision must be made, *tailoring variables* are information concerning the individual that is used for individualization, *intervention options* are an array of possible treatments/actions that might be employed at any given decision point, and *decision rules* are a way to operationalize the adaptation by specifying which intervention option to offer, for whom, and when.

⁴We provide examples of sample conversations in the supplementary material.

3.2 Choice of Factors and Sensor Data

In our work we explore how different factors relate to receptivity. We divide the factors into two groups, (a) intrinsic factors (which remain constant for a participant throughout the study, e.g., age, device type, gender, and personality) collected from pre-study surveys and (b) contextual factors (which may vary for each intervention message, e.g., time of day, location, physical activity) collected passively from smartphones. Our choice of intrinsic and contextual factors is based on a combination of review of prior works on context-aware notifications and the feasibility of obtaining different data from the phone.

Choice of intrinsic factors: Prior work on context-aware notifications have largely considered participant receptivity on Android devices. Given the significantly different OS architecture and notification delivery style between iOS and Android devices, we evaluate device type as factor towards receptivity. Further, prior works have shown associations between age and gender with smartphone usage [1] and availability for notification content [34]. In fact, Pielot et al. found age to be a top predictor for detecting engagement with notification [34]. Further, Mehrotra et al. found that different personality traits have an effect on the notification response time [22]. Hence, we consider device type, age, gender, and personality as the intrinsic factors for evaluating receptivity. While device type may not be necessarily “intrinsic” to a participant, it stayed constant over the course of the study, hence for the context of our work, we group it under intrinsic factors. Further, we acknowledge there could be other intrinsic factors (like employment status or marital status), and we discuss this later in the paper.

Choice of contextual factors: As discussed in the related work section, a variety of factors have earlier been shown to be associated with availability for notification prompts: physical activity, location, audio, phone battery state, conversation, communication (call/sms) logs, device usage, app usage, engagement with apps, etc. However, we faced certain limitations when it came to collecting certain data types; e.g., Apple does not let developers collect Audio (passively, in the background) without it being a VoIP app, and Google recently banned the use of Accessibility Services. Because we intended the app serve primarily as an intervention delivery application, we had to make some sacrifices with the rate and type of sensor data that was collected in the background. To keep things similar between Android and iOS, we collected almost all of the same sensors from the two devices. Eventually, we ended up collecting GPS location, Physical Activity, Date/Time, Battery status, Lock/Unlock events (iOS only), Screen on/off events (Android only), Wi-Fi connection state (iOS only) and Proximity sensor (Android only). This choice of sensors along with moderate duty-cycling resulted in only a 4-6% battery drain over 24 hours.

Hence, we considered the following contextual factors in our work: physical activity, location, battery, device usage, and time, all of which have shown some association with phone usage and availability to app notifications in the past.

3.3 Defining Receptivity for Interventions

Next, we define the metrics for *receptivity*, as used in the context of this study. Previous work in context-aware notification management systems has used several such metrics – seen time, decision time, response rate, satisfaction rate, notification acceptance and so on [11, 27, 31]. While these metrics are acceptable for generic notifications, for the case of intervention delivery we need specific metrics that can address the *state-of-receptivity* of an individual. We define two categories of metrics, (1) metrics that capture the *receptivity* of a person *in-the-moment*, and (2) metrics that capture the receptivity of a person over a period of time.

3.3.1 Metrics for in-the-moment receptivity. In this section, we define three metrics which we use for gauging a person’s receptivity *in-the-moment*.

- *Just-in-time response:* Based on the JITAI framework, it is important to trigger the intervention when the person is most receptive towards it. Mehrotra et al. [21] showed that users accept (i.e., view the content of)

over 60% of their notifications within 10 minutes of delivery, after which the notifications are left unhandled for a long time. They concluded that the maximum time a user should take to handle a notification that arrived when they were in an *interruptible moment* is 10 minutes. We thus decided to use 10 minutes after the intervention delivery as the period for an in-the-moment response. If the user views and replies to an initiating message, within 10 minutes, we conclude the user was in a receptive state at the time of the intervention message delivery and set the *just-in-time response* for that initiating message as ‘true’. We acknowledge that different intervention goals might have different thresholds for what is considered a receptive moment, and could range from a few minutes (in case of smoking cessation interventions) to a couple of days (for long term behavioral change interventions). Hence, a window of 10 minute might seem arbitrary. However, we use 10 minutes as a starting reference, as there is some quantitative evidence for a 10 minute window in interruptibility research. Consider an intervention whose initiating message was delivered at time t . If the participant responds (replies) within $t + 10$ minutes then *just-in-time response* is true.

- *Response delay*: This is the time (in minutes) taken between receiving an intervention based message (the time when the initiating message was delivered as a notification) and replying to it. If the participant replied to the initiating message (delivered at time t) at time t' , then the response delay is $t' - t$.
- *Conversation engagement*: An alert for an intervention can sometimes require minimal user input, e.g., just one click, but often it requires some level of engagement with the participant, where they might have to do a quick two-minute survey or take 20 deep breaths. Engagement with relevant content is something that has rarely been explored by other researchers. Pielot et al. [34] did a study where they were trying to detect participant engagement by sending them notifications with games, puzzles, news articles, etc. In our case, however, we attempt to look at participant engagement with the chatbot, which is providing the interventions. For our purposes, we define a participant was *engaged* in a conversation with the bot if the participant replied to more than one message within the 10 minute window from time of the first message delivery, i.e., if the participant replied to two or more messages within 10 minutes (i.e., $t + 10$ minutes), we mark that instance as *true* for conversation engagement.

3.3.2 Metrics for receptivity over an extended period of time. We also define four metrics for gauging a person’s receptivity over an extended period of time. We compute each metric for each participant individually over a given period: (1) *Just-in-time response rate*: The fraction of initiating messages for which the participant had a *just-in-time response*, over a given period, (2) *Overall response rate*: The fraction of initiating messages responded to by the participant (*just-in-time* or not), over a given period, (3) *Conversation rate*: The fraction of initiating messages for which the participant engaged in a conversation, over a given period, and (4) *Average response delay*: The mean response delay, over a given period.

3.4 Enrollment and Data Collected

We teamed up with a large Swiss health insurance company to conduct this study. The study was approved by the ethics-review boards (IRB) of both a Swiss and a US university. According to the agreement and the research plan, the insurance company would invite their customers to be part of the study. All participants had to be over 18 years old. Further, given the content of the intervention messages, they had to be German speaking and not working night shifts. After screening for eligibility, 274 participants signed the consent form and downloaded the Android or iOS version of the app.

After the participants installed the Ally app from the respective app stores, there was a 10-day baseline period during which the participants just received and responded to some welcome messages, and their step counts were recorded to determine their later step goals. After 10 days, the actual intervention study started and continued for 6 weeks. Out of the 274 users, we had step count data for 227 during the initial 10 days. The remaining 47 were

automatically discarded because without their step count data they could not be part of the intervention. Further, as the baseline period was ending, we had 214 participants reach the intervention phase. As the intervention phase started, after the first day, we had step count data for 203 participants. These were the 203 users who we could include in our study of receptivity towards interventions. However, several of them responded to very few messages throughout the intervention period, which can be considered as outliers, and if included in the final dataset would skew the results. We calculated the number of responses from the lowest 5 percentile of the participants and set that number (6 messages) as a threshold for inclusion in the analysis. Thus we included participants who had the top 95 percentile response count, which resulted in 189 participants (141 iOS and 48 Android). Of the 189 participants, 70 were male and 119 were female. The median age was 40 years \pm 13.7 years. Additional statistics about the data collected are shown in Table 1. Further, the iOS and Android sub-groups had very similar demographics, age and gender distribution, as compared to each other and the overall study population.

Table 1. Study Stats: Initiating messages are the start-of-conversation messages delivered to participants, initial responses are the responses to the initiating messages, just-in-time response % is the same as *just-in-time response rate* from Section 3.3.2, conversations engaged % is the same as *conversation rate* from Section 3.3.2., averaged across all participants.

	iOS		Android	
Initiating messages	10366		3576	
Initial responses	7466	72.02%	2912	81.45%
Just-in-time responses	2202	21.24%	896	25.06%
Conversations engaged	1659	16.00%	712	19.91%

Handling Missing Data: For iOS users (a majority of our participants), there is no way to explicitly detect when the phone is switched off/on, if the app has been uninstalled, or if the user does not provide our apps permission to collect any data and does not reply to messages. It is very hard to distinguish between such cases. We use the availability of sensor data as an indication that the device was turned on and the app was installed. So for each initiating message, if there was no sensor data a few hours before the delivery of the message, we assume that the data collection was interrupted and we do not include that initiating message in our analysis (even if the participant responded to that message). We follow the same approach for Android users, to maintain uniformity in data processing.

While this is an useful approach for handling missing data, there could be edge cases where the participant switched off their phone for a few minutes during which an initiating message was delivered. Since the phone was switched off, the push notification would not be visible once the phone is turned on. The initiating message would be included in the analysis as the phone was switched off for a few minutes and there was sensing data available before that. In such cases, even if the participant was in a state-of-receptivity, due to lack of response to an initiating message, they would be categorized as not receptive. However, since we sent only 2-3 initiating messages a day, we argue the odds of this particular edge case was quite low.

3.5 Analysis Plan

In our work we explore how different factors (intrinsic or contextual) relate to the receptivity metrics for the interventions.

Further, we explore the relationship between intervention effectiveness and the receptivity metrics. Finally, we build machine-learning models that can account for different factors (both intrinsic and contextual) to infer *in-the-moment* receptivity, based on the three metrics discussed in Section 3.3.1.

Depending on the type of the dependent and independent variables and the type of analyses required, we use several statistical methods to complete our analyses – Welch’s t-test, one-way ANOVA, two-way ANOVA, linear

models, generalized linear models, logistic regression models, Chi-Square tests, repeated measures ANOVA and mixed-effects generalized linear models.

Important: in this paper we use several statistical tests to explore the relationship between different factors and metrics. In our evaluations, when we discuss the *effect* of a variable on a metric, we refer to statistical effects only. We do not and cannot imply that the variables have a *causal effect* on a particular receptivity metric.

4 EFFECT OF INTRINSIC FACTORS

In this section, we evaluate how different intrinsic factors relate to the participants' reaction to intervention messages. We evaluate the effects of the intrinsic factors on the aggregated receptivity metrics from Section 3.3.2, over the course of the study, i.e., the just-in-time response rate, overall response rate, average response delay, and the conversation rate.

Effect of Device Type: Participants were free to use either iOS or Android version of Ally. While the app user interface was identical on both the platforms, there are some fundamental differences in the user experience between both the operating systems and how they handle notifications. Hence, we evaluated whether the choice of device plays a role in the participants' responsiveness towards the intervention messages.

We conduct multiple Welch's two-sample *t*-tests to observe the differences in message interactions between the two groups – iOS and Android. We observed that Android users had a significantly higher total response rate, $t(92.38) = 2.422, p = 0.017$, along with significantly lower response delay, $t(74.42) = -2.099, p = 0.039$. This observation suggests that through-out the study, on average, Android users responded to more interventions, and were faster than iOS users while responding. We did not observe any significant differences for just-in-time response rates and conversation rates. This implies that while iOS and Android users responded similarly *in-the-moment* of intervention delivery; Android users, however, came back to complete the intervention at a later time, more often. We believe the outcome might be related to how the two OSs bundle notifications on the lock screen. iOS 11 (the latest OS during the time of the study) used to bundle and push notifications to an 'Earlier Today' section, giving more visibility to newer notifications; in contrast, Android usually bundles notifications by apps, regardless of when they arrived. In Android, if the user does not explicitly 'click' or 'swipe-away' the notification, it will be visible every time a user pulls down the notification pane; however, in iOS 11 an extra gesture was required to bring up the 'Earlier Today' section.

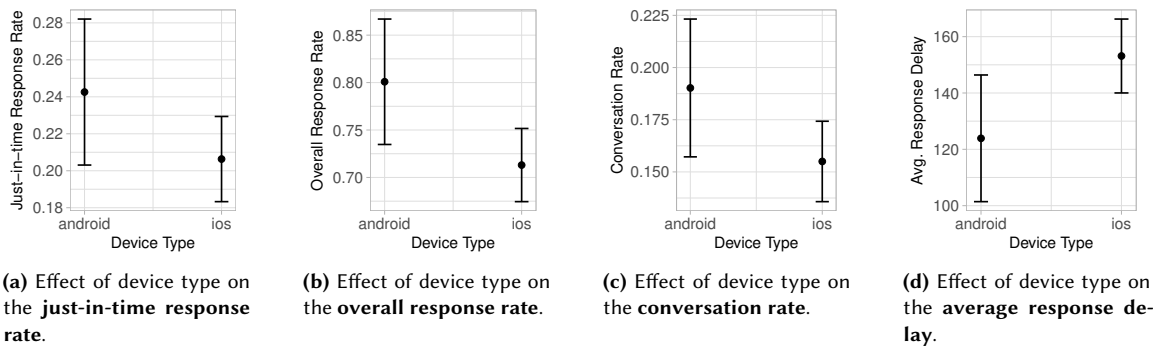


Fig. 2. Effect of the type of device on different *state-of-receptivity* metrics. The points represent the mean and the bars represent the 95% Confidence Interval (CI).

Effect of Age: We observed that the age of the person had a significant effect $F(1, 177) = 4.798, p = 0.029$ on the just-in-time response rate. We observe from Fig. 3 that the just-in-time response rate increased with age, suggesting that older people were, on average, more responsive to the intervention within the 10-minute threshold. Further, we observed a significant effect for conversation engagement rate, $F(1, 177) = 7.150, p = 0.008$, with older people also engaging in higher number of conversations.

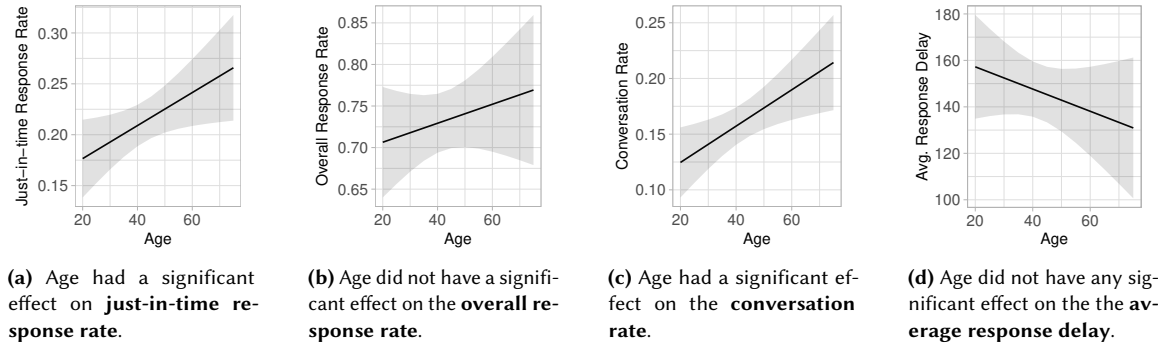


Fig. 3. Effect of age of the participants on the different *state-of-receptivity* metrics. The slope represents the mean and the shaded region represents the 95% CI.

Effect of Gender: We found that the gender of a person did not have a significant effect on any of the receptivity metrics.

Effect of Personality: At the start of our study, the participants filled out a personality survey, the BFI-10 [36], from which we computed the score (on a scale of 1–10) for the five personality traits – Extroversion, Agreeableness, Conscientiousness, Neuroticism and Openness. Of the 189 included participants, we had complete survey information for 154 participants, so we look at the effect of personality traits across those 154 participants.

We fit multiple linear models for each of the metrics (just-in-time response rate, total response rate, average response delay, and conversation rate) as the dependent variable and the five personality traits as independent variables. We noticed that some personality traits had a significant effect on the just-in-time response rate, conversation rate and total response rate.

Table 2. Effect of personality on **just-in-time response rate**, **overall response rate** and **conversation rate**.

	Just-in-time response rate				Overall response rate				Conversation rate			
	Estimated Coefficient	Std. Error	<i>t</i> value	<i>p</i> value	Estimated Coefficient	Std. Error	<i>t</i> value	<i>p</i> value	Estimated Coefficient	Std. Error	<i>t</i> value	<i>p</i> value
Openness	0.0029893	0.0056388	0.530	0.5968	0.008960	0.010211	0.877	0.38166	0.003005	0.004620	0.650	0.5165
Agreeableness	0.0140628	0.0075328	1.867	0.0639	0.030003	0.013641	2.199	0.02940 *	0.013482	0.006172	2.184	0.0305 *
Extroversion	-0.0009457	0.0062873	-0.150	0.8806	0.005709	0.011386	0.501	0.61680	-0.002767	0.005152	-0.537	0.5920
Conscientiousness	0.0048500	0.0068795	0.705	0.4819	0.032599	0.012458	2.617	0.00980 **	0.004493	0.005637	0.797	0.4267
Neuroticism	0.0112682	0.0056814	1.983	0.0492 *	0.030189	0.010289	2.934	0.00388 **	0.005266	0.004655	1.131	0.2598
.	<i>p</i> < 0.1	N=154			N=154				N=154			
*	<i>p</i> < 0.05	Adjusted $R^2 = 0.685$			Adjusted $R^2 = 0.886$				Adjusted $R^2 = 0.648$			
**	<i>p</i> < 0.01	F(5,148)=67.71, <i>p</i> <0.001			F(5,148)=240.7, <i>p</i> <0.001				F(5,148)=57.54, <i>p</i> <0.001			

Table 2 shows the parameters of the fitted linear model. We observe that the neuroticism trait significantly affects the just-in-time response rate. Participants with high level of neuroticism showed higher just-in-time

response rate, which is an interesting observation. Ehrenberg et al. found that neurotic individuals spent more time text messaging and reported stronger mobile-phone addictive tendencies [9], which could be a possible explanation for the significant effect on just-in-time response rate. Further, we observe that agreeableness, conscientiousness and neuroticism traits showed significant effect with the total response rate, with higher scores in those traits leading to higher response rates. Finally, agreeableness also showed significant effect with conversation rate, such that participants with higher agreeableness score were likely to have higher conversation rates. A person with higher agreeableness score is *sympathetic, good-natured, and cooperative* [39], and that might be the reason they comply and engage in more conversations once they have replied to the initial message. These, however, are speculations and we reiterate that we report statistical correlations that do not necessarily imply causal effects.

5 EFFECT OF CONTEXTUAL FACTORS

In this section we evaluate how contextual factors (which might change often) relate to the interaction with intervention messages. Given our earlier analysis highlighting the significant differences between Android and iOS users along with the fact that there were some differences in the types of sensor data collected from the two platforms, we perform the contextual analyses and report results separately for both groups, i.e., 141 iOS users and 48 Android users. Further, we evaluate the contextual effects on the three metrics of intervention receptivity, as defined in Section 3.3.1 – just-in-time response, conversation engagement, and the response delay.

5.1 Effect of Time of Delivery

Based on our intervention delivery strategy from Section 3.1, we group the hours of the day into three categories: (1) before 10 a.m., (2) 10 a.m. – 6 p.m., and (3) after 6 p.m. We further categorize the day of the week into weekday or weekend. For evaluation, we fit binomial Generalized Linear Models (GLM) to understand the effect of the time of day and type of day on just-in-time response and conversation likelihood, and perform two-way ANOVA to understand effect on response delay.

iOS users: We observed that time of day had a significant effect on just-in-time response, $\chi^2(2) = 47.237, p < 0.001$. Post-hoc analysis using Tukey contrasts revealed that iOS users were most likely to answer prompts delivered during the day, i.e., between 10 a.m. and 6 p.m., and they were least likely to answer to prompts during the evening, i.e., after 6 p.m. Type of day was also significant, with $\chi^2(1) = 7.853, p < 0.005$. We found that iOS users were more likely to respond just-in-time on weekdays, as compared to weekends. In fact, from Fig. 4a, we observe that iOS users were least likely to respond on weekend mornings, i.e., before 10 a.m., and most likely to respond on weekdays between 10 a.m. and 6 p.m.

iOS users showed a significant interaction between time of day and type of day on conversation engagement, $\chi^2(2) = 10.179, p < 0.001$, and on response delay, $F(2, 7352) = 34.933, p < 0.001$. Post-hoc analysis shows that iOS users were least likely to engage in conversation on weekend mornings; in fact on the weekends, the users were most likely to engage in conversations during the night, as shown in Fig. 4b. Further, iOS users replied the fastest during the day (10 a.m. – 6 p.m.) on weekdays and the slowest in the same time period of weekends; during the other times there was not a lot of difference in response delay, as evident in Fig 4c.

Android users: We observed that only the time of the day had a significant effect on just-in-time response, $\chi^2(2) = 3.0, p = 0.223$, regardless of the type of day, which was not significant. Post-hoc analysis for time of day revealed trends similar to iOS users, i.e., participants were most likely to respond between 10 a.m. and 6 p.m., and least likely to respond in the morning before 10 a.m. From Fig. 5a. Android users seemed to be more responsive on weekends during the day (10 a.m. – 6 p.m.), as compared to weekdays during the day; although insignificant, it is opposite to what was observed from iOS users.

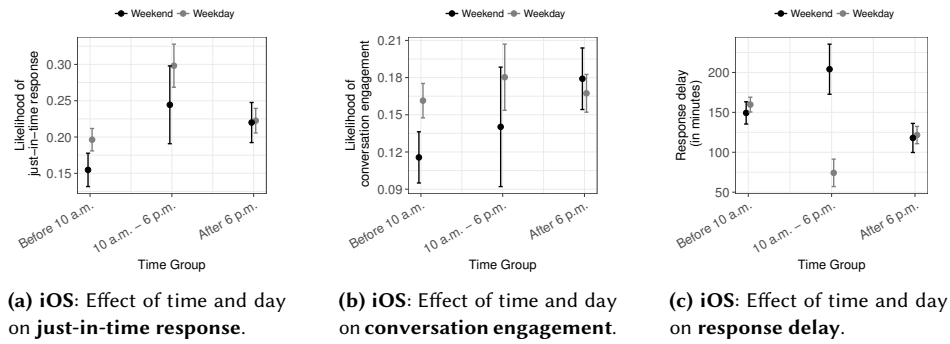


Fig. 4. Effect of interaction between time of day and type of day on the different receptivity metrics for iOS users.

Further, for Android users, we found a significant effect of time of day on conversation engagement, $\chi^2(2) = 11.861$, $p = 0.002$, and on response delay, $F(2, 2856) = 31.101$, $p < 0.001$. Post-hoc analysis revealed that conversation engagement was significantly lower during the mornings (before 10 a.m.), as shown in Fig. 5b. Further, Android users replied the fastest during the day (10 a.m. – 6 p.m.), regardless of the type of day, and slowest during the morning. Unlike iOS users, the trends on weekdays and weekends were consistent for Android users, as can be seen in Fig. 5c.

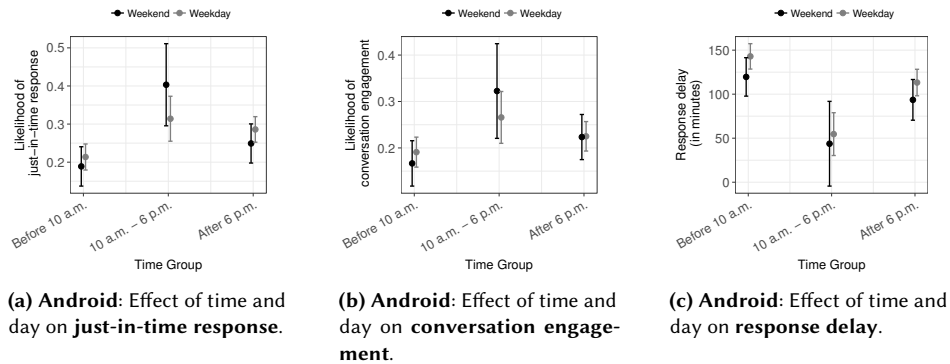


Fig. 5. Effect of interaction between time of day and type of day on the different receptivity and Android users.

5.2 Effect of Phone Battery

Next, we look at the relationship between the device battery and the receptivity metrics. We consider two features about the device's battery – the battery level of the device and the battery status of the device, i.e., charging, unplugged, full charge. We removed all instances where battery status of the device was “unknown”, as it does not provide any information. We start with the effect on the just-in-time response rate.

The battery status showed a significant effect on just-in-time response for both device types – iOS: $\chi^2(2) = 9.825$, $p < 0.001$, and Android: $\chi^2(2) = 11.5235$, $p = 0.003$. On post-hoc analysis, we found that in both cases the

likelihood of a just-in-time response was lower when the battery status was “Fully Charged”, as evident in Fig. 6. This makes sense intuitively because if the phone battery status is “Fully Charged” then the phone has been charging for quite some time and the participant does not have the phone handy and hence will not be able to provide just-in-time responses. iOS users also had a significant effect of the battery status on the likelihood of conversation engagement, $\chi^2(2) = 7.389, p < 0.001$.

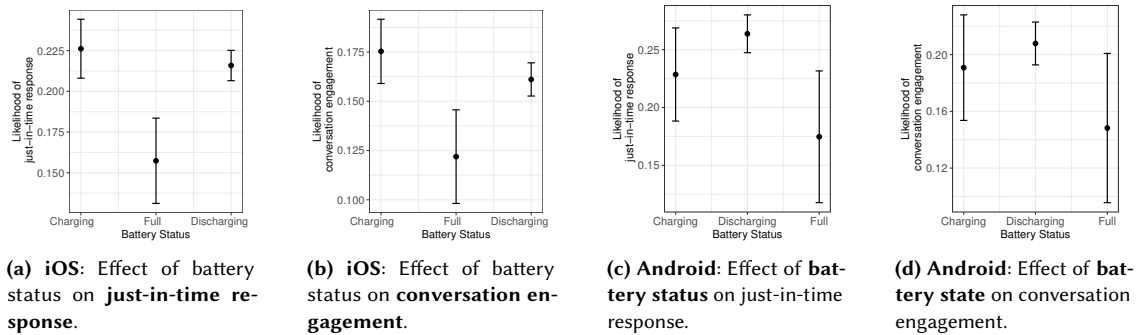


Fig. 6. Effect of **battery status** on the likelihood of a **just-in-time response** and **conversation engagement**

For both iOS and Android users, we found an interesting effect of battery level on the response delay – higher battery levels led to higher response times (iOS: $F(1, 7436) = 9.171, p = 0.002$, and Android: $F(1, 2781) = 5.590, p = 0.018$). We observed that for Android users, the battery status had a significant effect on the response delay, $F(2, 2781) = 22.984, p < 0.001$. Further, we found that the “Charging” state resulted in significantly higher response times, as evident in Fig. 7, which is different from the trends observed from iOS users. One possible reason could be the charging time difference between Android and iOS devices. iOS devices usually have smaller batteries and could charge faster than Android users, hence the device spends less time in “Charging” state and more time in “Full” state.

Based on this significant effect of charging state on response delay among Android users along with the (previously mentioned) effect of time on the receptivity metrics, we can speculate the cause for this – usually people charge their phones during the night (after returning from work, or while sleeping) or during the day (before leaving for work), which means that when the phone is charging, the participants might not have their phone close, which leads to a significantly longer response time. The observation about significantly shorter response delay during the day (10 a.m. – 6 p.m.) seems to corroborate this speculation.

We believe it is important to state that here battery information, by itself, might not mean much. It can, however, act as a proxy to some causal effect, e.g., participant is indoors, or away from phone.

5.3 Effect of Device Interaction

Further, we investigate how different interactions with the phone, e.g., lock/unlock, screen on/off, Wi-Fi connection, etc., relates to intervention receptivity. Given the difference in OS and developer APIs for Android and iOS, the variables for device interaction are slightly different for each platform. We explain the variables in Table 3.

iOS users: We found that there was a significant effect of *lock state* on just-in-time response, $\chi^2(1) = 19.115, p < 0.001$, and on conversation engagement, $\chi^2(1) = 15.570, p < 0.001$, suggesting a higher likelihood of response and conversation engagement if the phone is unlocked. Further, *lock change time* had a significant effect on all three receptivity metrics (just-in-time response: $\chi^2(1) = 15.519, p < 0.001$, conversation engagement: $\chi^2(1) = 7.771,$

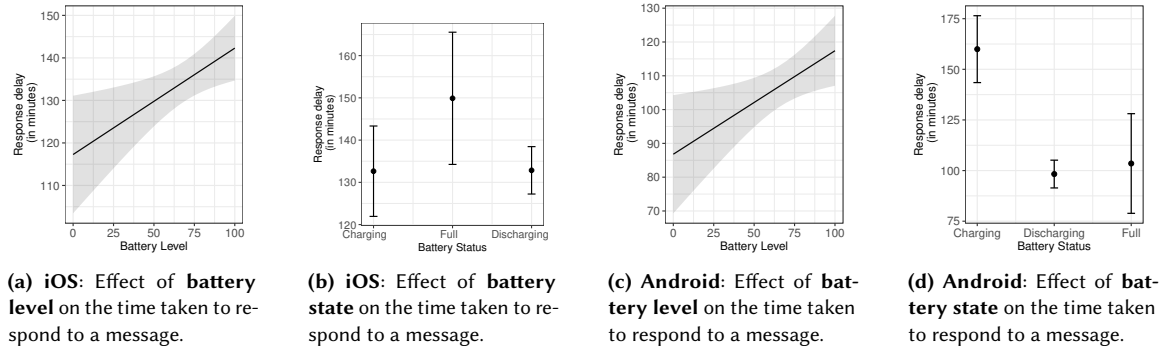
Fig. 7. Effect of battery on the **response delay**.

Table 3. Variables for device interaction for iOS and Android.

	iOS		Android
Lock State	Locked or unlocked (boolean)	Screen State	Screen on or off (boolean)
Lock change time	Seconds since last phone lock	Screen on/off count	Number of times the screen has been turned on/off in the last 5 minutes
Phone unlock count	Number of times the phone has been unlocked in the last 5 minutes	Screen change time	Seconds since the last screen off
Wi-Fi Connection	Connected to Wi-Fi (boolean)	Proximity	Value from proximity sensor (boolean)

$p = 0.005$ and response delay: $F(1, 5393) = 5.079, p = 0.024$). We found as time since last lock event increased, likelihood of response, conversation engagement decreased and the response delay increased, suggesting that if messages were not delivered at the right time, they could go unopened and ignored for a long time. For iOS users, we also observed that as more lock/unlock events happened in the time preceding the intervention, e.g., fiddling with the phone out of boredom, the more was the likelihood of just-in-time response ($\chi^2(1) = 9.747, p = 0.001$). Similar trends have been observed by Pielot et al. [35]. Finally, we observed if the participant's devices were connected to Wi-Fi, there was higher likelihood of a just-in-time response ($\chi^2(1) = 30.405, p < 0.001$) and conversation engagement ($\chi^2(1) = 6.517, p = 0.010$). Here, Wi-Fi connection could be a proxy to some causal factors, e.g., the participant is indoors, or at a known location, which led to an increased likelihood of response. We show the effect of different device interaction variables on just-in-time response in Fig. 8.

Android users: For Android, we collected the screen on/off data instead of phone lock/unlock data. While both are similar, there could be subtle differences between screen on/off and phone lock/unlock, e.g., participants checking time on their phone without unlocking it. We found that, similar to trends observed for iOS users, *screen state* and *screen change time* showed significant effects on just-in-time response ($\chi^2(1) = 15.545, p < 0.001$, and $\chi^2(1) = 46.732, p < 0.001$, respectively) and conversation engagement ($\chi^2(1) = 16.448, p < 0.001$ and $\chi^2(1) = 40.655, p < 0.001$, respectively). Further, the longer it had been since the last screen change, the longer it took to respond to the message ($F(1, 2435) = 26.801, p < 0.001$). We also observed that as more *screen on/off* events occurred, the likelihood of just-in-time response and conversation engagement increased ($\chi^2(1) = 17.308, p < 0.001$ and $\chi^2(1) = 14.758, p < 0.001$, respectively). Further, we observed when *proximity* was 0, i.e., something was blocking the proximity sensor (e.g., the phone was in the pocket, or in the bag, or while talking on the phone), the likelihood for a response and conversation engagement was significantly lower than if the proximity value was 1, $\chi^2(1) = 31.563, p < 0.001$ and $\chi^2(1) = 29.481, p < 0.001$, respectively. Further, if the proximity was 0, it took users

significantly longer to respond $F(1, 2435) = 16.536, p < 0.001$. We show the effect of different device interaction variables on just-in-time response in Fig. 9.

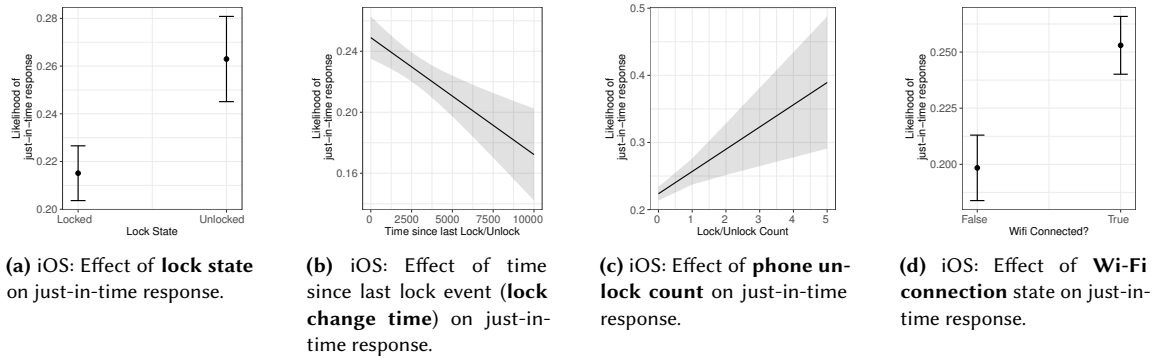


Fig. 8. Effect of different **iOS** phone interaction measures on **just-in-time response**.

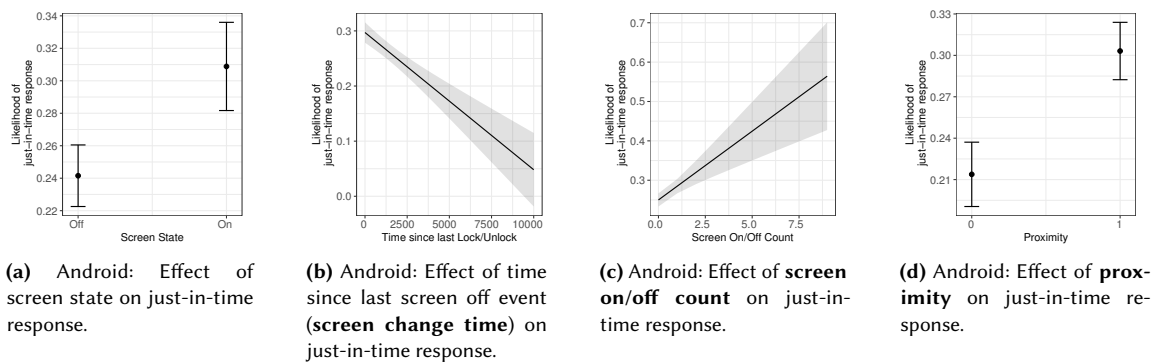


Fig. 9. Effect of different **Android** phone interaction measures on **just-in-time response**.

5.4 Effect of Physical Activity

In this section, we evaluate the effect of the participant’s physical activity on the receptivity metrics. We obtain the physical activity by subscribing to CoreMotion Activity Manager on iOS [3] and Google Activity Recognition Api on Android [8]. On iOS, we were able to receive the following activities – in vehicle, on bicycle, on foot, running, still and shaking. According to iOS CoreMotion Activity Manager updates, *on foot* means *walking*; *running* is a different activity. Due to the low and sporadic occurrences of shaking activity, we removed the instances where the user activity was shaking, leaving us with 5 detected activity types. Similarly, on Android we received – in vehicle, on bicycle, running, still, walking, on foot, and tilting. Due to the low occurrences of tilting and the fact that in Android *on foot* is a complete superset of walking and running, we remove instances with tilting and on foot activities, leaving us with 5 detected activity types.

For iOS users, we observed a significant effect of the physical activity type on the just-in-time response, $\chi^2(4) = 22.299, p < 0.001$, and conversation engagement, $\chi^2(4) = 19.439, p < 0.001$. Post-hoc analysis showed that likelihood of response and conversation engagement while the participants were *on foot*, i.e., walking, was significantly higher than when the participants were still. The result is in line with observations by Pielot et al. [34] and makes sense intuitively, as when the phone is still, it is probably lying on a table or not in physical possession of the participant, whereas when the phone detects a walking activity, then it *is* on the participant's person, and hence there would be a higher chance of them responding to the intervention message. Similarly, while the participants were in a vehicle (e.g., driving), the likelihood for response was lower than if the participants were on foot, which again seems reasonable as people should not be interacting with the phone while driving. Fig. 10 shows the effects of activity on the different receptivity metrics for iOS users.

In the case of Android users, we did not observe any significant effect of activity on the different receptivity metrics, but the trends are similar to that of iOS users Fig. 11.

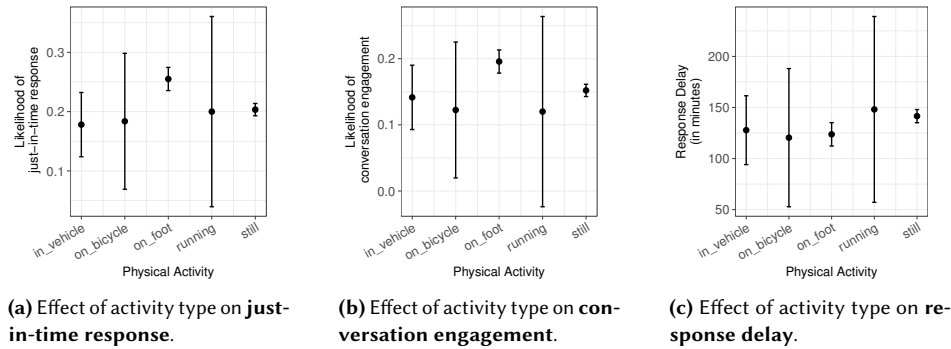


Fig. 10. iOS: Effect of activity type on the receptivity metrics.

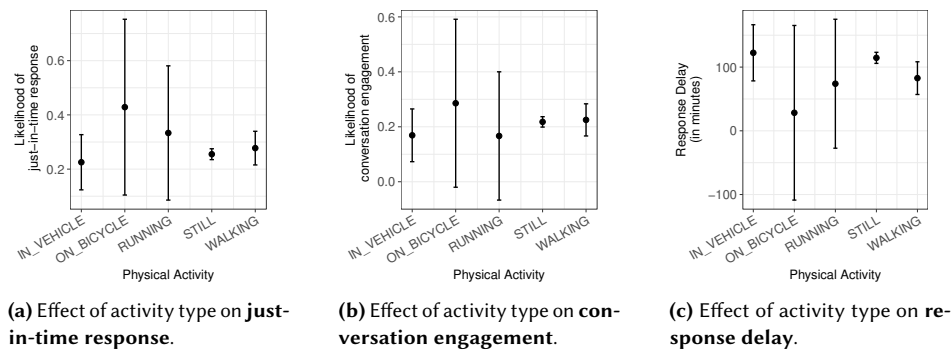


Fig. 11. Android: Effect of activity type on receptivity metrics.

5.5 Effect of Location

During the course of the study, Ally sampled the participants' fine-grain location every 20 minutes. We used the DBSCAN algorithm to generate location clusters from the raw location [26] and categorize them as *home*, *work*, *transit* and *other*. First, we filtered and removed raw GPS instances where the accuracy was over 100 meters, and then we generated clusters with a radius of 100 meters and a minimum of 20 samples per cluster. The rationale for a minimum of 20 samples was to capture places where a participant might visit and stay for some time, or frequently visit for short periods of time. We wanted to avoid clusters being generated while the participant was traveling, or was a one-time visit for a short period of time. Having generated these clusters, we labeled the most frequent location cluster between 10 p.m. and 6 a.m. every night, as *home*. We further labeled the most frequent location cluster between 10 a.m and 4 p.m. every weekday, that is not home, as *work*; all the remaining clusters were labeled as *other*, and all un-clustered locations were grouped as *transit*. We are making some assumptions here while labeling clusters to simplify the problem.

Only Android users had a significant effect of location on just-in-time response ($\chi^2(3) = 22.579, p < 0.001$) and on conversation engagement ($\chi^2(3) = 12.614, p = 0.005$). Post-hoc analysis reveals that Android users seemed to be least likely to respond or to engage in conversation at *other* locations, i.e., locations they visit regularly or frequently enough, but it is not their home or work. *Other* locations could be grocery stores, friend's house, a favorite restaurant, etc. While the difference was not significant amongst iOS users, it is reassuring to see a similar trend, as evident from Fig. 12. Further, we observed a significant effect of the type of location on the response delay for both iOS ($F(3, 7357) = 3.351, p = 0.018$) and Android ($F(3, 2674) = 2.581, p = 0.05$) users. On post-hoc analyses, it seems that iOS users took significantly longer to respond to messages at *work*, as compared to at *home*, or when in *transit*. On the other hand, Android users took significantly longer at *other* locations, as shown in Fig. 13. This is a peculiar difference observed amongst Android and iOS users.⁵

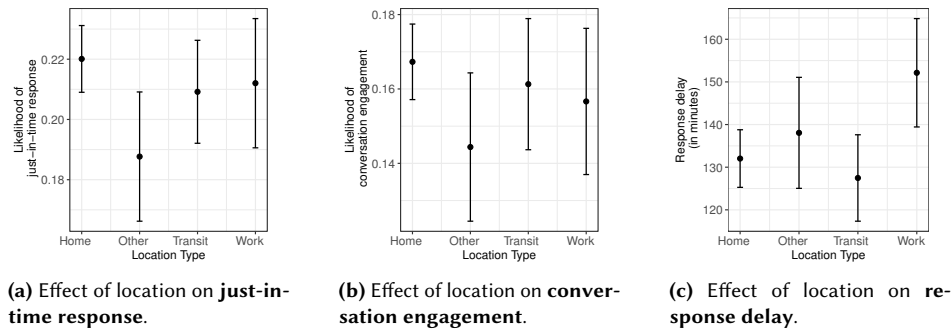


Fig. 12. iOS: Effect of location on receptivity metrics

6 RELATION BETWEEN RECEPTIVITY AND EFFECTIVENESS OF INTERVENTION

In the previous sections, we discuss the effect individual factors or contextual factors have on the users' receptivity. In this section, we explore the relation between the intervention's outcome (step count) and receptivity. While contextual factors have shown to affect receptivity towards notifications [21, 22], EMAs [23, 38], or interventions (in our case), the relationship between the effectiveness of the intervention and receptivity has not been explored

⁵While location is an objective measure across the two devices, and our clustering algorithm was identical for both device types, it seems there is an unknown confounding factor in play here.

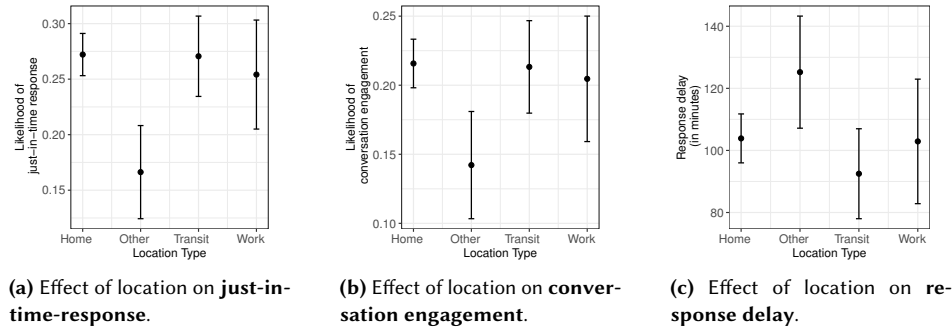


Fig. 13. **Android**: Effect of **location** on receptivity metrics

before. Our hypothesis is that if users do find the interventions useful or effective, it can be a motivating factor towards being more receptive towards them. Unlike several mental and behavioral health conditions, where it is difficult to gauge the effectiveness of an intervention remotely, without additional input from the user, we have a more objective and quantifiable measure, i.e., step count. If the user completed more steps than the goal set for that day, we say the user *completed the goal* for that day. *Goal completion* is a binary measure for each day, and we use it as a proxy for the effectiveness of the intervention. In this section, we explore the relationship of *goal completion* with the different receptivity metrics. We start by looking at the effect of *goal completion rate* (the fraction of days a user completed the goals) on the overall response rate, just-in-time response rate, conversation rate, and average response delay, over the course of the study. These are the same metrics we use in Section 3.3.2.

6.1 Goal Completion and Receptivity over the Course of the Study

We start with a correlation analysis between the goal completion rate and the different receptivity metrics. We observed correlations with overall response rate ($r = 0.53, p < 0.001$), just-in-time response rate ($r = 0.42, p < 0.001$), conversation rate ($r = 0.38, p < 0.001$) and average response delay ($r = -0.27, p < 0.001$). Based on the result, there seems to be a significant association between receptivity and step-goal completion. In fact, participants close to high goal completion rates had an overall response rate averaging at approximately 95%, i.e., they responded to 95% of all messages they received.

The above results being correlations, we do not know whether achieving more goals led to users being more receptive to interventions, or whether a participant's receptivity to interventions led to higher goal completion, over the course of the study. We explore this in the next section, where we observe the statistical effect of daily goal completion on the receptivity metrics for the next day and vice versa.

6.2 Analyzing the Relationship between Daily Goal Completion and Receptivity

In this section we analyze the temporal relationship between goal completion and receptivity. For each participant, we considered the binary goal completion outcome for a day d . Next, we calculated the receptivity metrics for day d and day $d + 1$. The metrics are daily response rate, daily just-in-time response rate, daily conversation rate and average response delay for that day. These are the same metrics from Section 3.3.2, but are now calculated over the course of a day, instead of the entire study duration.

We start by exploring the relationship between receptivity on day d with goal completion for the same day d . We used a mixed-effects generalized linear model with Type II LR Chi-square testing for this analysis with the *participant id* as the random variable, which controls for variability between participants. We fit different models

with each of receptivity metrics (daily response rate, daily just-in-time response rate, daily conversation rate, and average response delay) as independent variables and the goal completion (binary) as the dependent variable. This test determines whether receptivity had an effect on goal completion. We then explore the relationship between goal completion on day d with receptivity on day $d + 1$. To this end, we use repeated measures ANOVA, again with *participant id* as the random variable, with each of the receptivity metrics as a dependent variable and the goal completion (binary) as the independent variable. This test determines whether goal completion on a particular day had a significant effect on receptivity the next day.

From the first analysis, we observed significant effects of the three receptivity metrics on their daily goal completion. From the trends observed in Fig. 14, participants were up to 62% more likely to complete their daily goal if they had higher response rate to interventions that day ($\chi^2(1) = 134.780, p < 0.001$), and participants were up to 30% more likely to complete their goals if they had higher *just-in-time* response ($\chi^2(1) = 55.320, p < 0.001$), with participants with high *just-in-time* responses showing close to 68% likelihood of goal completion. Further, participants were up to 31% more likely to complete their goals the next day if they engaged in a conversation ($\chi^2(1) = 52.732, p < 0.001$), with participants with high conversation engagement rates showing close to 73% likelihood of goal completion. Finally, the likelihood of goal completion reduced by approximately 36% as the average response delay increased ($\chi^2(1) = 36.518, p < 0.001$).

This analysis increases our confidence in the choice of receptivity metrics; it is encouraging to see that being receptive to interventions was strongly related to goal completions for that day. This observation further strengthens our motivation for exploring receptivity for interventions.

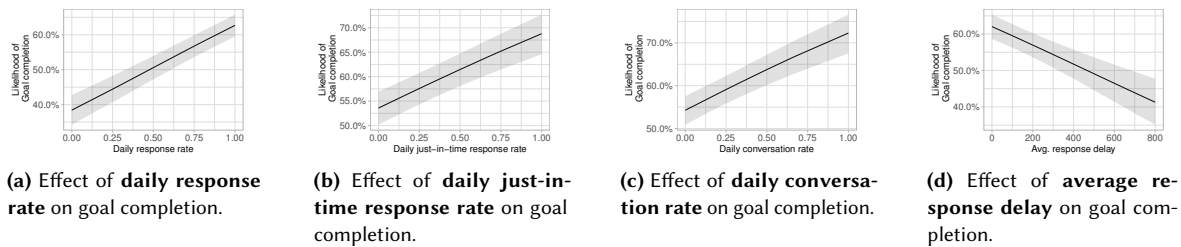


Fig. 14. Effect of different receptivity metrics on **goal completion**.

Our second analysis revealed a significant effect of goal completion (on day d) on the different receptivity metrics (on day $d + 1$): daily response rate ($F(1, 7030.7) = 42.892, p < 0.001$), daily just-in-time response rate ($F(1, 7120.6) = 28.606, p < 0.001$), daily conversation rate ($F(1, 7123.9) = 18.163, p < 0.001$), and average response delay ($F(1, 6161.9) = 20.364, p < 0.001$); suggesting that participants who completed their step goal for a day had higher receptivity the next day. We show the effects in Fig. 15. We speculate that completion of goals the previous day motivated participants to be more receptive to future intervention messages.

While we conduct a temporal analyses of receptivity and goal completion in this section, we cannot imply a cause-effect relationship between these variables, as we could not and did not control for other confounding variables, nor was it possible to do so under the experimental design of our longitudinal study. However, we speculate that receptivity to interventions motivates participants to achieve their step goals, and achieving of step goals further motivates the person to engage and be more receptive to future intervention messages, in a virtuous cycle. We believe, this result is extremely valuable, and has not been explored before.

Given the cyclic relation between receptivity and goal completion, we hypothesize that receptivity on day d has an effect on receptivity on day $d + 1$: Given the results above, there could be a transitive relation. Using a mixed-effects generalized linear model, we found that the data supported our hypothesis. We found significant

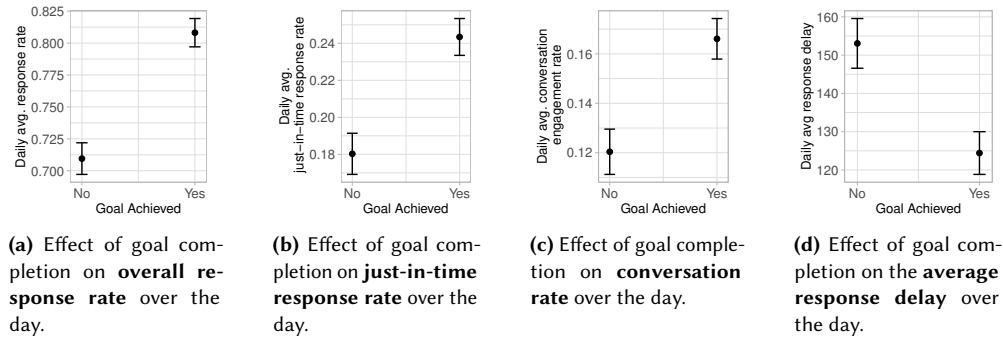


Fig. 15. Effect of **goal completion** on the receptivity metrics for the next day.

effects of daily response rate on day d with the the daily response rate ($F(1, 7082.7) = 345.81, p < 0.001$), the just-in-time response rate ($F(1, 6196.2) = 96.02, p < 0.001$), conversation rate ($F(1, 6228) = 59.18, p < 0.001$) and avg. response delay ($F(1, 5643.8) = 95.12, p < 0.001$) on day $d + 1$. Based on our results, receptivity on the day before could be an additional predictor for receptivity.

7 MACHINE-LEARNING MODELS FOR PREDICTING RECEPTIVITY METRICS

In this section we explore the extent to which the receptivity metrics (just-in-time response, conversation engagement, and response delay from Section 3.3.1) can be estimated using the contextual and intrinsic factors discussed in previous sections. To this end we train machine-learning models and evaluate them by cross-validation. Given the differences in data collected from iOS and Android devices, along with the fact that sometimes they showed significantly different effects for a similar factor, we decided to build and evaluate models separately for iOS and Android.⁶ For each of the receptivity metrics, we built two models, (1) with just contextual features, like location, activity, device interaction, etc., and (2) with the combination of contextual features and intrinsic factors like personality, age, and gender. The purpose of building two separate models is to observe whether incorporating intrinsic factors affect classifier performance. We did not build a model with just the intrinsic factors as they remain constant throughout the study and would not be useful for predicting *in-the-moment* receptivity metrics, just by themselves. While prediction of just-in-time response and conversation engagement is a binary classification problem, the estimation of the third metric, i.e., response delay, is a regression-based estimation. We use Random Forests for both the classification and regression tasks. Random Forests are simple, yet robust classifiers as they perform reasonably well without over-fitting. We report 10-Fold cross-validation results, along with the Precision, Recall and F1 scores.

Machine learning models for iOS: We start with the prediction of just-in-time responses, which is a binary variable. The results for classification with and without intrinsic factors are shown in Fig. 16a. To put these results into perspective, we also implemented a baseline classifier that randomly classified 0 or 1 based on the probability distribution of the training set. We observe that even without intrinsic factors, we can detect just-in-time responses with an F1 score of 0.31, which is approximately 50% higher than the baseline classifier. Further, introducing the intrinsic factors into the model leads to a sharp increase in Precision to 0.38, which is a 80% increase over the baseline classifier, and brings the F1 score to 0.36, which is a 71% increase over the baseline.

⁶We do not incorporate the day goal completion information in these preliminary models. We intend to explore this in the future.

Next, we move on to detection of conversation engagements, which also is a binary variable. The results for classification with and without intrinsic factors are shown in Fig. 16b. We also show the performance from a random baseline classifier for comparison. Using just the contextual features, there is not a lot of improvement in Precision (over the baseline), but the Recall is higher, and brings the F1 score to 0.21, which is a 31% improvement over the baseline. Incorporating the intrinsic factors as well leads to an increase in performance and brings the F1 score to 0.29, which is a 81% increase over the baseline classifier.

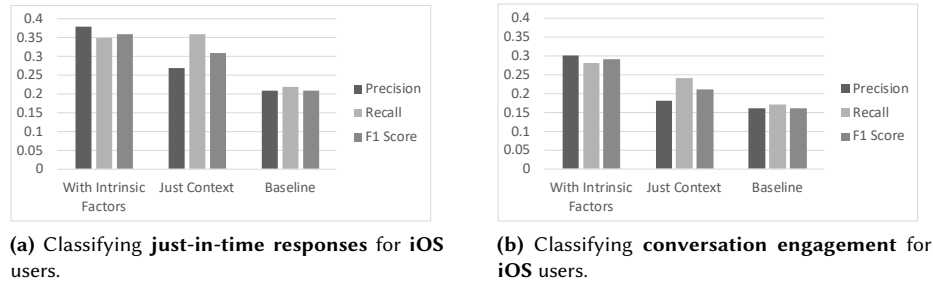


Fig. 16. iOS classification performance

Finally, for estimating the response delay, based on a regression based classifier, we observe a correlation coefficient $r = 0.29$, $p < 0.001$ using Random Forests, as compared to a baseline of $r = -0.032$, $p = 0.01$ using a ZeroR classifier (which always predicts the mean value of the training set).

Machine learning models for Android: For detecting just-in-time responses, we observed that using just contextual features led to a 51% improvement in F1 score over the baseline of a biased random guess. The addition of intrinsic factors led to an improvement of approximately 77% over the baseline, as can be seen in Fig. 17a. The improvements in the task of just-in-time response detection for both iOS and Android are quite similar to each other.

For detecting conversation engagement, a classifier using just contextual features led to an increase of approximately 60% in F1 score over the biased random guess baseline. The addition of intrinsic factors did not change the Precision much, but was a major boost to the recall, which led to an F1 score improvement of 100% over the baseline, i.e., double the F1 score of a biased random guess, as evident in Fig. 17b.

For estimating the response delay for Android users, we observed a correlation coefficient $r = 0.33$, $p < 0.001$, as compared to a baseline of $r = -0.073$, $p = 0.04$ using a ZeroR classifier.

The results shown here (i.e., the % gain in performance) as a feasibility test, perform better than similar results from previous works, e.g., by Sarker et al. [38], where they carry out 10-fold cross-validation to detect availability to their surveys, and achieve almost a 50% increase over the base classifier, or by Pielot et al. [34] where the authors detect engagement with notification content with an increase of 66.6% over the baseline. Neither of these previous works have tried to estimate response delay, i.e., the time taken to respond. The preliminary results give us confidence in the potential for real-time prediction of *just-in-time responses* along with *conversation engagement* and *response delay* using sophisticated ML techniques in the future, and hopefully further improve on the preliminary results discussed here.

Predictive features: Further, we look into the predictive power of the different factors in the Random Forest model. Based on an accuracy based search, we found, for iOS, the best feature subset was: *age, gender, agreeableness, neuroticism, openness, type of day (weekday or weekend), battery level, battery status, lock state, phone unlock count,*

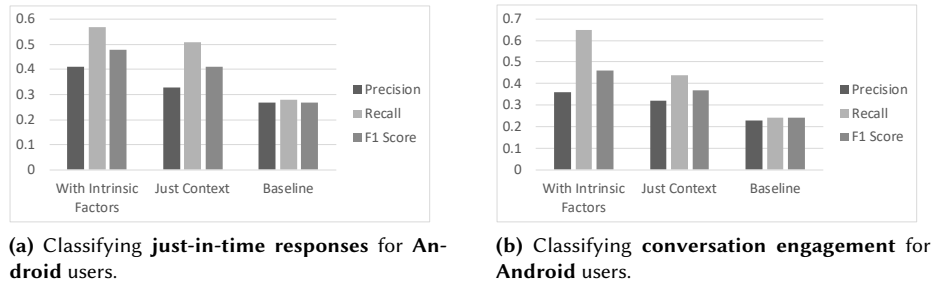


Fig. 17. Android classification performance

Wi-Fi connection, and *location type*. It is interesting to note that the feature subset contains several features that did not have a significant effect when analyzed individually (gender, agreeableness, openness and battery level), and leaves out some features that did have significant effects (activity type, time since last lock state change). This is where the correlations and interactions between different variables come into effect.

For Android, we found the best feature subset was *age*, *agreeableness*, *conscientiousness*, *neuroticism*, *openness*, *day of week*, *time of day*, *battery level*, *battery status*, *screen state*, and *screen state change time*. Similar to iOS users, the subset includes several features that did not have a significant effect when analyzed individually (agreeableness, conscientiousness, openness and battery level), and leaves out several features which had a significant effect (location type, proximity and screen on/off count). However, we do see an overlap between iOS and Android, which suggests that it might be possible to build a common model for both devices, something we plan to explore in future work. It is important to note that the above features were selected when using a Random Forest classifier. Different machine learning models might have different predictive features.

8 DISCUSSION

In this section, we discuss the implications of our results along with several open questions that need to be explored.

- *Analysis of intrinsic factors*: In our work we explored how different intrinsic factors relate to receptivity with interventions. We observed several significant relationships between age, device type and some personality traits with different receptivity metrics. Further, we observed that incorporating intrinsic factors (along with contextual factors) in ML models, to detect receptivity, helps boost the detection performance. Based on this result, we believe that these “intrinsic” factors can be used to build group-based contextual ML models that can perform better than a generalized ML model. We envision the group-specific ML models to use the dynamic contextual factors to detect/predict receptivity as the intervention progresses, but the model parameters would be tuned to optimize performance to that specific group; for example, models built for neurotic individuals, who show stronger mobile-phone addictive tendencies, might have different weights for the device interaction as compared to individuals with low neuroticism scores. These group-based models have potential serve as a good middle-ground or transition between a generalized model and individualized model.

Further, while more investigation is required, the findings discussed in our work could potentially influence creation of ‘intervention options’ for specific groups of people. For example, in our analyses, we observed that participants with higher neuroticism scores responded to more initiating messages, but did not have higher conversation engagement. Based on this result, intervention options for a study aiming to study

neurotic participants could be designed to include shorter (single-tap) messages spread throughout the day, instead of requiring them to engage in a conversation. Hence, in a JITAI, decision rules could use the tailoring variable of receptivity (from ML models) to determine which intervention option to use at a given decision time.

There could be other participant-specific characteristics that show associations with availability for interventions. We considered other characteristics which we hypothesized could affect receptivity – employment status, income levels, financial incentive type, and weight of the participant at the start of the study. However, none of these had any significant on any of the different receptivity metrics. Since the results were not significant, in our work we only highlighted characteristics that have been explored in the past, to be inline with previous works. There could be other intrinsic factors (e.g., marital status or number of children), that we did not account for.

It is possible that for intrinsic factors, there exist some correlations between factors. Such correlations (if any) could be grossly misleading about the inference we make. For example, if gender and device type is strongly correlated and we make an inference that women are more receptive, it could be because of a particular device type. In this case, device type could have *caused* the increase receptivity, and not because women are more receptive. We perform correlation analyses between the factors and discuss in supplementary material. We did not observe several strong correlations except between gender and age. Women negatively correlated with age, implying in our study, women were younger as compared to men. However, this did not affect receptivity, as we observed that age had an effect on receptivity metrics, but gender did not.

- *Analyses of contextual factors*: We observed significant associations between several intrinsic and contextual variables and the receptivity metrics. In our work, we only observed the main effects of the different variables, without accounting for interaction between different contextual terms. It is completely possible that interactions exist between different contexts, e.g., we observed that iOS users had a significantly higher just-in-time response if they were walking at (or near) home, which was 50% higher than being still at home, and significantly lower response rate if they were walking at *other* locations. The purpose was to find individual variables that can help capture receptivity (over a population) even when other contextual variables were missing or unavailable.

Further, while we believe these variables can be used directly in machine-learning models, it is possible that combining different contextual variables to determine a descriptive context, e.g., studying, working, partying, might lead to better predictive and more interpretable models. Given the nature of our study, we could not collect certain types of sensor data, e.g., audio, or SMS and call logs, as we did not want participants to experience severe battery drain or be concerned about their privacy. We intend to collect these data streams in the future and build models using a descriptive context, and explore whether it leads to better predictive performance. We believe, models built using contextual data can help accurately detect receptivity, strengthening the tailoring variable to improve selection of appropriate decision rules at the decision points, hence enabling efficacious JITAI.

- *Relationship between goal completion and receptivity*: From the results in Section 6.2, we speculate that receptivity to interventions led to higher goal completion that day, and completing goals on a day motivated the participants to be more receptive the next day. It seems that receptivity and goal completion have a virtuous effect on each other. Potentially, researchers could get a sense of the effectiveness of an intervention based on how participants engage with that intervention, and hence could possibly improve, modify or change the intervention being administered. While the results look promising, please take note of the point discussing the generalizability of our results with other JITAI, below.
- *Differences between Android and iOS*: Our results explore how participants respond to interventions delivered on different mobile operating systems. As the demographics of the participants (age and gender) using

Android and iOS was similar, we believe the notification presentation, alert and grouping methodologies amongst the two OSs played a role in how participants respond. Further, both the OSs allow apps to access different types of sensor data. While Android is more flexible in allowing apps access to a variety of data, iOS has always been very restrictive. The easy availability of sensor data is one of the reasons the ubiquitous computing community focuses on Android devices for interruptibility research. Based on our results, we call for more research across different platforms to determine the similarities and differences in how the participants respond to notifications and interventions between the two major platforms.

- *Generalizability of our results to other JITAI*: Since we consider receptivity to the initial greeting message, we believe our results could generalize to other JITAI use cases with similar level of intervention engagement, or to JITAI that use a chatbot for digital coaching. It is, however, possible that if the interventions require the users to do some substantial activities, e.g., taking 10 deep breaths, meditate for 10 seconds, go for a walk, or take a particular medications our results might not apply. We currently have other intervention based studies underway and we plan to explore the generalizability of these results to other studies in future work. Another aspect to consider is the burden of engaging with interventions. In our study, we sent a consistent 2-3 initiating messages every day; further research needs to be done to understand how receptivity changes if the number of messages increase/decrease drastically across days.

9 CONCLUSION

In this work, we explored several factors affecting users' receptivity in JITAIs. We conducted a study with 189 participants, over a period of 6 weeks, in which participants received actual interventions to improve their physical activity levels. To deliver the interventions, we developed a chatbot based digital coach – Ally – and made it available on Android and iOS platforms. We defined several metrics to gauge receptivity towards the interventions; and from our exploratory analyses, we observed (1) several participant specific traits (age, personality and device type) showed significant associations with the overall participant receptivity over the course of the study, and (2) several contextual factors (day/time, phone battery, phone interaction, physical activity and location) showed significant associations with the participant receptivity, *in-the-moment*. Further, we explored the relationship between effectiveness of the intervention and receptivity towards interventions, and based on our analyses, we speculate that being receptive to interventions helped achieve physical activity goals, which in turn motivated participants to be more receptive to future interventions. Finally, we built machine-learning models and observed significant improvements over a baseline classifier in detecting and estimating the different receptivity metrics.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers and editors for their helpful feedback. This research results from a joint research effort of the Center for Digital Health Interventions at ETH Zürich and the Center for Technology and Behavioral Health at Dartmouth College. It was supported by the NIH National Institute of Drug Abuse under award number NIH/NIDA P30DA029926, and by CSS Health Insurance, Switzerland. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.

REFERENCES

- [1] Ionut Andone, Konrad Blaszkiewicz, Mark Eibes, Boris Trendafilov, Christian Montag, and Alexander Markowetz. 2016. How Age and Gender Affect Smartphone Usage. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16)*. ACM, New York, NY, USA, 9–12. <https://doi.org/10.1145/2968219.2971451>
- [2] Adrienne Andrew, Gaetano Borriello, and James Fogarty. 2007. Toward a Systematic Understanding of Suggestion Tactics in Persuasive Technologies. In *Persuasive Technology*, Yvonne de Kort, Wijnand IJsselstein, Cees Midden, Berry Eggen, and B. J. Fogg (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 259–270.

- [3] Apple. 2018. CMMotionActivityManager. <https://developer.apple.com/documentation/coremotion/cmmotionactivitymanager>. (2018). [Online; accessed 11-February-2019].
- [4] Daniel Avrahami and Scott E Hudson. 2006. Responsiveness in instant messaging: predictive models supporting inter-personal communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 731–740.
- [5] Stephanie Bauer, Judith de Niet, Reinier Timman, and Hans Kordy. 2010. Enhancement of care through self-monitoring and tailored feedback via text messaging and their use in the treatment of childhood overweight. *Patient education and counseling* 79, 3 (2010), 315–319.
- [6] Dror Ben-Zeev, Christopher J Brenner, Mark Begale, Jennifer Duffecy, David C Mohr, and Kim T Mueser. 2014. Feasibility, acceptability, and preliminary efficacy of a smartphone intervention for schizophrenia. *Schizophrenia bulletin* 40, 6 (2014), 1244–1253.
- [7] Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, et al. 2008. Activity Sensing in the Wild: A Field Trial of UbiFit Garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 1797–1806.
- [8] Google Developers. 2018. Activity Recognition APL. <https://developers.google.com/location-context/activity-recognition>. (2018). [Online; accessed 11-February-2019].
- [9] Alexandra Ehrenberg, Suzanna Juckes, Katherine M White, and Shari P Walsh. 2008. Personality and self-esteem as predictors of young people’s technology use. *Cyberpsychology & behavior* 11, 6 (2008), 739–741.
- [10] Andreas Filler, Tobias Kowatsch, Severin Haug, Fabian Wahle, Thorsten Staake, and Elgar Fleisch. 2015. MobileCoach: A novel open source platform for the design of evidence-based, scalable and low-cost behavioral health interventions: overview and preliminary evaluation in the public health context. In *Wireless Telecommunications Symposium (WTS), 2015*. IEEE, 1–6.
- [11] Joel E Fischer, Chris Greenhalgh, and Steve Benford. 2011. Investigating episodes of mobile phone activity as indicators of opportune moments to deliver notifications. In *Proceedings of the International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI)*. ACM, 181–190.
- [12] Joel E Fischer, Nick Yee, Victoria Bellotti, Nathan Good, Steve Benford, and Chris Greenhalgh. 2010. Effects of content and time of delivery on receptivity to mobile interruptions. In *Proceedings of the International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI)*. ACM, 103–112.
- [13] David H Gustafson, Fiona M McTavish, Ming-Yuan Chih, Amy K Atwood, Roberta A Johnson, Michael G Boyle, Michael S Levy, Hilary Driscoll, Steven M Chisholm, Lisa Dillenburg, et al. 2014. A smartphone application to support recovery from alcoholism: a randomized clinical trial. *JAMA Psychiatry* 71, 5 (2014), 566–572.
- [14] Joyce Ho and Stephen S Intille. 2005. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 909–918.
- [15] Karen Hovsepian, Mustafa Al’Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment. *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)* (2015), 493–504. <https://doi.org/10.1145/2750858.2807526>
- [16] Abby C King, Eric B Hekler, Lauren A Grieco, Sandra J Winter, Jylana L Sheats, Matthew P Buman, Banny Banerjee, Thomas N Robinson, and Jesse Cirimele. 2013. Harnessing different motivational frames via mobile phones to promote daily physical activity and reduce sedentary behavior in aging adults. *PloS one* 8, 4 (2013), e62613.
- [17] Tobias Kowatsch, Dirk Volland, Iris Shih, Dominik Rügger, Florian Künzler, Filipe Barata, Andreas Filler, Dirk Büchter, Björn Brogle, Katrin Heldt, et al. 2017. Design and Evaluation of a Mobile Chat App for the Open Source Behavioral Health Intervention Platform MobileCoach. In *International Conference on Design Science Research in Information Systems*. Springer, 485–489.
- [18] Jan-Niklas Kramer, Florian Künzler, Varun Mishra, Bastien Presset, David Kotz, Shawna Smith, Urte Scholz, and Tobias Kowatsch. 2018. Investigating Intervention Components and Exploring States of Receptivity for a Smartphone App to Promote Physical Activity: Study Protocol of the ALLY Micro-Randomized Trial. *JMIR Research Protocols*, forthcoming (2018).
- [19] Florian Künzler, Jan-Niklas Kramer, and Tobias Kowatsch. 2017. Efficacy of mobile context-aware notification management systems: A systematic literature review and meta-analysis. In *Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 131–138.
- [20] Afra Mashhadi, Akhil Mathur, and Fahim Kawsar. 2014. The myth of subtle notifications. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 111–114.
- [21] Abhinav Mehrotra, Mirco Musolesi, Robert Hendley, and Veljko Pejovic. 2015. Designing content-driven intelligent notification mechanisms for mobile applications. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 813–824.
- [22] Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. 2016. My phone and me: understanding people’s receptivity to mobile notifications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 1021–1032.
- [23] Varun Mishra, Byron Lowens, Sarah Lord, Kelly Caine, and David Kotz. 2017. Investigating Contextual Cues As Indicators for EMA Delivery. In *Proceedings of the International Workshop on Smart & Ambient Notification and Attention Management (UbiTention)*. ACM,

- 935–940. <https://doi.org/10.1145/3123024.3124571>
- [24] Varun Mishra, Gunnar Pope, Sarah Lord, Stephanie Lewia, Byron Lowens, Kelly Caine, Sougata Sen, Ryan Halter, and David Kotz. 2018. The Case for a Commodity Hardware Solution for Stress Detection. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct (UbiComp'18)*. ACM. <https://doi.org/10.1145/3267305.3267538>
- [25] Varun Mishra, Gunnar Pope, Sarah Lord, Stephanie Lewia, Byron Lowens, Kelly Caine, Sougata Sen, Ryan Halter, and David Kotz. 2019. Continuous Detection of Physiological Stress with Commodity Hardware. *ACM Transactions on Computing for Healthcare (HEALTH)* 1, 1 (2019). <https://doi.org/10.1145/3361562>
- [26] Raul Montoliu, Jan Blom, and Daniel Gatica-Perez. 2013. Discovering places of interest in everyday life from smartphone data. *Multimedia Tools and Applications* 62, 1 (01 Jan 2013), 179–207. <https://doi.org/10.1007/s11042-011-0982-z>
- [27] Leanne G Morrison, Charlie Hargood, Veljko Pejovic, Adam WA Geraghty, Scott Lloyd, Natalie Goodman, Danius T Michaelides, Anna Weston, Mirco Musolesi, Mark J Weal, et al. 2017. The effect of timing and frequency of push notifications on usage of a smartphone-based stress management intervention: An exploratory trial. *PLoS one* 12, 1 (2017), e0169162.
- [28] Inbal Nahum-Shani, Eric B Hekler, and Donna Spruijt-Metz. 2015. Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. *Health Psychology* 34, S (2015), 1209.
- [29] Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. 2016. Just-in-time adaptive interventions (JITAs) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine* 52, 6 (2016), 446–462.
- [30] Heather L. O'Brien and Elaine G. Toms. 2008. What is User Engagement? A Conceptual Framework for Defining User Engagement with Technology. *J. Am. Soc. Inf. Sci. Technol.* 59, 6 (April 2008), 938–955. <https://doi.org/10.1002/asi.v59:6>
- [31] Mikio Obuchi, Wataru Sasaki, Tadashi Okoshi, Jin Nakazawa, and Hideyuki Tokuda. 2016. Investigating interruptibility at activity breakpoints using smartphone activity recognition API. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 1602–1607.
- [32] Tadashi Okoshi, Julian Ramos, Hiroki Nozaki, Jin Nakazawa, Anind K Dey, and Hideyuki Tokuda. 2015. Reducing users' perceived mental effort due to interruptive notifications in multi-device mobile environments. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 475–486.
- [33] Veljko Pejovic and Mirco Musolesi. 2014. InterruptMe: designing intelligent prompting mechanisms for pervasive applications. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 897–908.
- [34] Martin Pielot, Bruno Cardoso, Kleomenis Katevas, Joan Serrà, Aleksandar Matic, and Nuria Oliver. 2017. Beyond interruptibility: Predicting opportune moments to engage mobile phone users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 91.
- [35] Martin Pielot, Tilman Dingler, Jose San Pedro, and Nuria Oliver. 2015. When attention is not scarce-detecting boredom from mobile phone usage. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 825–836.
- [36] Beatrice Rammstedt and Oliver P John. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of research in Personality* 41, 1 (2007), 203–212.
- [37] William Riley, Jami Obermayer, and Jersino Jean-Mary. 2008. Internet and mobile phone text messaging intervention for college smokers. *Journal of American College Health* 57, 2 (2008), 245–248.
- [38] Hillol Sarker, Moushumi Sharmin, Amin Ahsan Ali, Md Mahbubur Rahman, Rummana Bari, Syed Monowar Hossain, and Santosh Kumar. 2014. Assessing the availability of users to engage in just-in-time intervention in the natural environment. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 909–920.
- [39] Ginka Toegel and Jean-Louis Barsoux. 2012. How to Become a Better Leader. *MIT Sloan Management Review* 53, 3 (2012), 51–60.
- [40] Tilo Westermann, Ina Wechsung, and Sebastian Möller. 2016. Smartphone Notifications in Context: A Case Study on Receptivity by the Example of an Advertising Service. In *Proceedings of the CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2355–2361.

Exploring the State-of-Receptivity for mHealth Interventions: Supplementary Material

FLORIAN KÜNZLER*, ETH Zürich
VARUN MISHRA*†, Dartmouth College
JAN-NIKLAS KRAMER, University of St. Gallen
DAVID KOTZ, Dartmouth College
ELGAR FLEISCH, ETH Zürich and University of St. Gallen
TOBIAS KOWATSCH, ETH Zürich and University of St. Gallen

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**; • **Applied computing** → *Health care information systems*.

Additional Key Words and Phrases: Receptivity, Intervention, Interruption, Mobile Health, Engagement

ACM Reference Format:

Florian Künzler, Varun Mishra, Jan-Niklas Kramer, David Kotz, Elgar Fleisch, and Tobias Kowatsch. 2019. Exploring the State-of-Receptivity for mHealth Interventions: Supplementary Material. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 140 (December 2019), 12 pages. <https://doi.org/10.1145/3369805>

S1 EXPECTATION OF DIRECTION OF EFFECTS FROM PRIOR WORK

In this work we are exploring associations from the data collected and participants' receptivity to intervention messages. This analysis for interventions has not been explored before and hence is one of our primary research questions. In the past, researchers have observed associations between different factors and response to notifications or EMA and have observed similar and sometimes conflicting trends. Given that there is no clear consensus for the observed trends in past work – maybe because of the population size, or population demographics or device type or different measure of receptivity/availability – we chose to conduct an exploratory analysis on a representative population using different devices.

However, based on list our expectation of the direction of effects based on observations from previous work. We formulate our hypotheses on the entire receptivity metric: response rate, conversation rate and response delay. Hence higher receptivity would mean higher response rate, higher conversation rate and lower response delay.

*Both authors contributed equally to this research.

†Corresponding author

Authors' addresses: Florian Künzler, ETH Zürich; Varun Mishra, Dartmouth College, varun@cs.dartmouth.edu; Jan-Niklas Kramer, University of St. Gallen; David Kotz, Dartmouth College; Elgar Fleisch, ETH Zürich, University of St. Gallen; Tobias Kowatsch, ETH Zürich, University of St. Gallen.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2019 Copyright held by the owner/author(s).

2474-9567/2019/12-ART140

<https://doi.org/10.1145/3369805>

- Age: Pielot et al. observed that age had a positive correlation with engagement with notifications [3]. Hence, we hypothesize older participants would be more likely to be receptive as compared to younger participants.
- Gender: In the same study, Pielot et al. did not observe any association between gender of the user and engagement with notifications. Hence, we hypothesize gender would not have a significant effect on receptivity.
- Personality: Mehrotra et al. observed that extroversion and neuroticism affect how fast people respond to notifications [2]. Higher extroversion led to longer response time, and higher neuroticism led to faster response time. Hence, we hypothesize these traits should influence receptivity. Higher neuroticism should lead to higher receptivity and higher extroversion should lead to lower receptivity.
- Time of day: There is no conclusive evidence if time of day has a significant effect on receptivity. Westermann et al. found that replying solely on the time of the day for sending notifications had no impact on how fast people engaged with the content [6]. However, Pielot et al. found that notifications posted later in the day had higher engagement rates [3]. In fact, they found that engagement levels remained similar throughout the day and only lowered during early hours of the day (7-8 am). Based on observation by Pielot et al. we hypothesize that receptivity towards interventions would be lower during the early hours of the day, as compared to the rest of the day.
- Phone Battery: Pielot et al. found a weak positive correlation between phone battery level and engagement rates ($r = 0.008, p < 0.05$) [3]. They concluded that participants were more likely to engage if the battery level was higher. However, given the low correlation coefficient (r), we do not expect to observe any significant effect of battery level and receptivity metrics.
- Device Interaction: Pielot et al. found that number of phone unlocks correlate with increased feelings of boredom, which in turn correlated with openness to consume entertaining news articles (i.e., engaging the user) [4]. In a different work, Pielot et al. found that participants were more likely to engage with notifications the less time had passed since their last device interaction (change of screen status) [3]. We formulate our hypotheses for receptivity based on the observation by these works.
- Physical Activity: In previous work, Sarker et al. observed that walking did not indicate availability towards EMA prompts [5]. However, Pielot et al. found that more difficult to predict activities (such as driving, cycling, walking) correlate with higher openness for engagement, as compared to being still [3]. Mehrotra et al. found that certain activities (like cycling) were less opportune moments for delivering notifications, as compared to other activities (like walking). Given the lack of clear consensus on effect of physical activity, we hypothesize that walking would relate with higher receptivity as compared to other activities.
- Location: Similar to physical activity, there is no clear consensus about effect of location. Mehrotra et al. found that type of location (home, work, other) did not have a significant effect on response times to notifications [1]. However, Sarker et al. found that, as compared to other locations, participants at work were less available to respond to EMA [5]. Hence, we hypothesize that participants at work will be less receptive.

S2 CORRELATION ANALYSIS OF FACTORS

There could be correlations and interdependencies amongst the different factors we considered in our work. For the intrinsic factors, we believe, such correlations (if any) could be grossly misleading about the inference we make. For example, if gender and device type is strongly correlated and we make an inference that women are more receptive, it could be because of a particular device type. In this case, device type could have *caused* the increase receptivity, and not because women are more receptive. To this end, we perform a correlation analysis

between the different intrinsic factors; the result is shown in Fig. 1. We did not observe several strong correlations except between gender and age. Women negatively correlated with age, implying in our study, women were younger as compared to men. However, this did not affect receptivity, as we observed that age had an effect on receptivity metrics, but gender did not.

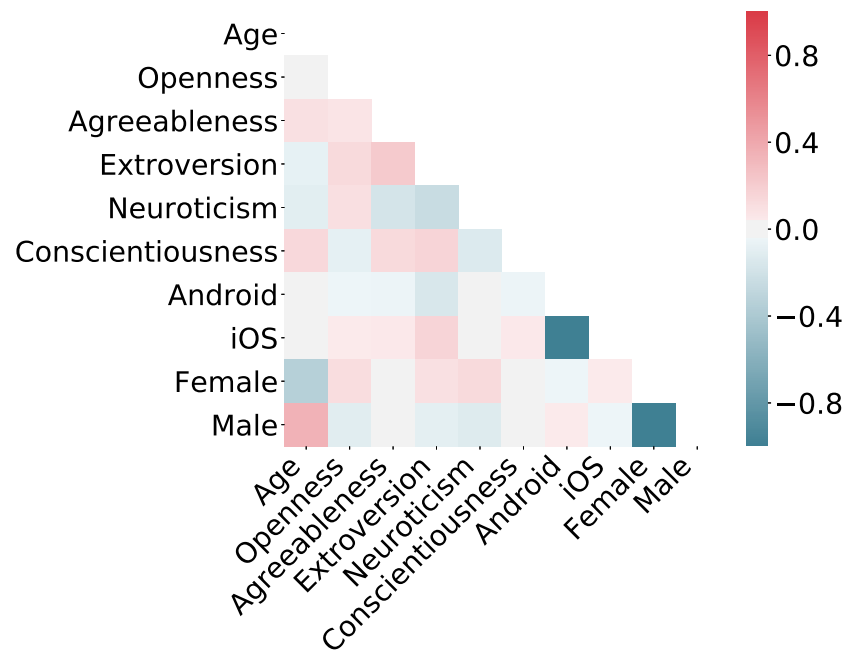


Fig. 1. Correlation analysis of the different intrinsic factors

However, if correlations exist between the contextual factors, it is not a major problem. We argue that all the contextual factors that are sensed from the smartphone are independent of each other, i.e., the inference from the analysis of one sensor would not change regardless of the presence of another 'correlated'. This is because if a sensor signal is associated with a behavioral pattern (in our case, receptivity), there was a 'cause' that led to that behavior. We simply capture an effect of that 'cause' in our various sensor signals. In no circumstance do we imply that the signal caused or led to a particular behavior. As a naive example, let us *assume* people charge their phone when working in office, and because they are working they are less receptive. So here the factor that is 'causing' them to be less receptive is the fact that they are working. Signals like phone battery state, work location and time of day are just signals affected by the 'cause'. In this case, the three signals are correlated, but we cannot just discard two of them because we might think one of the signals caused the behavior. We believe, all three signals are necessary to best capture the 'cause'. This is one of the major reason we considered each factor individually, instead of combining them together in a multi-variate model, where the effect of a correlated signal might be minimized, i.e., a correlated variable might not be significant in a multi-variate model, but significant in a bivariate model.

However, for completeness, we perform correlation analysis on the different contextual factors for iOS and Android users separately, and show the results in Fig. 2 & 3, respectively. We observed several interesting characteristics, e.g., WiFi connected is strongly correlated to users being at home. This is one example of our

reasoning to evaluate each factor individually. Assuming, researchers did not have location information, they could monitor Wi-Fi state to make meaningful inferences. Similarly, we observe that battery full is correlated with time since last screen state change (or lock/unlock time). This suggests that if the battery is full, the phone has been charging for a while and not in use by the participants, thereby making them less receptive, consistent to our analysis in the paper. This is another example, where analyzing the factors individually provides useful information for detecting receptivity.

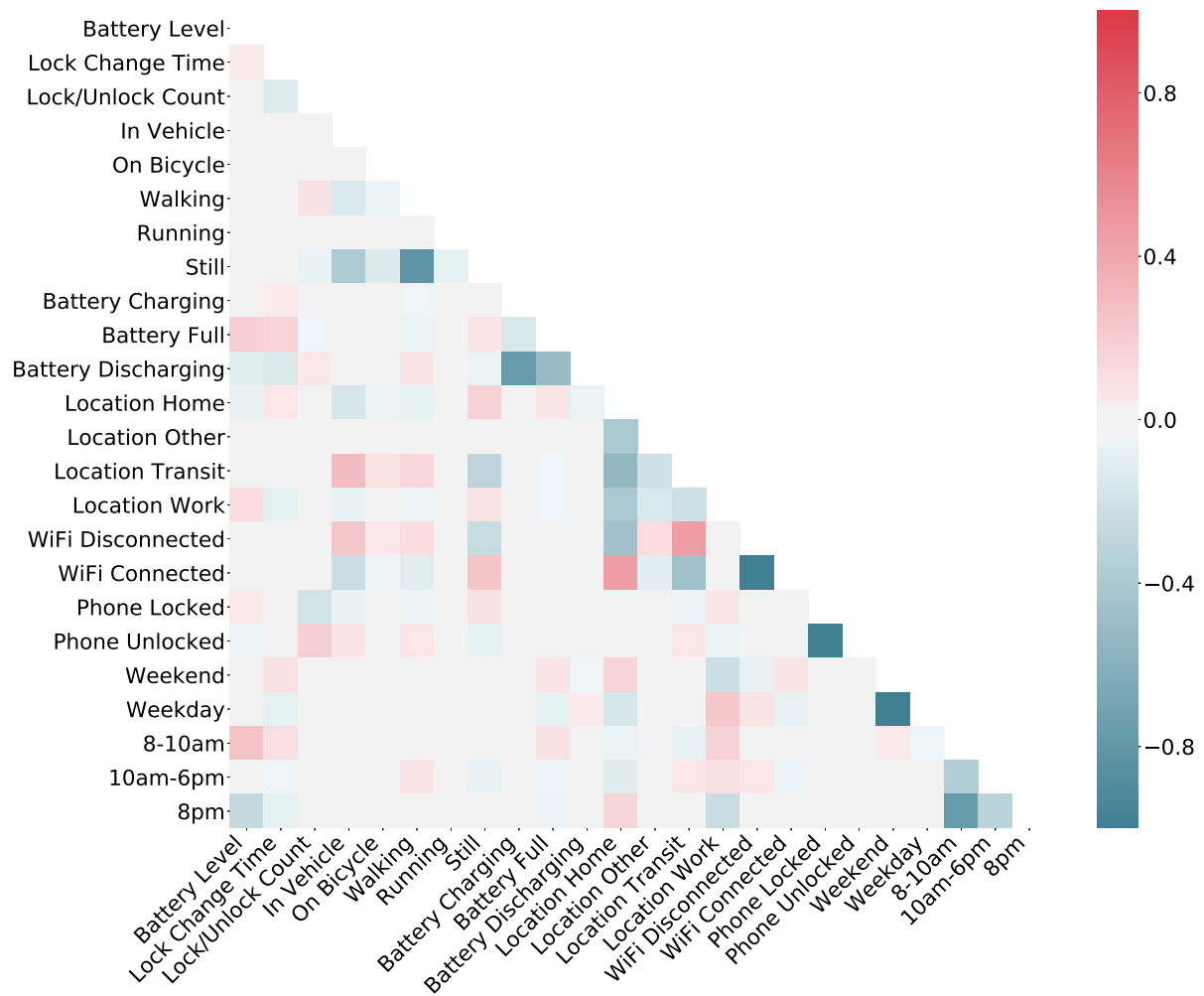


Fig. 2. iOS: Correlation analysis of the different contextual factors

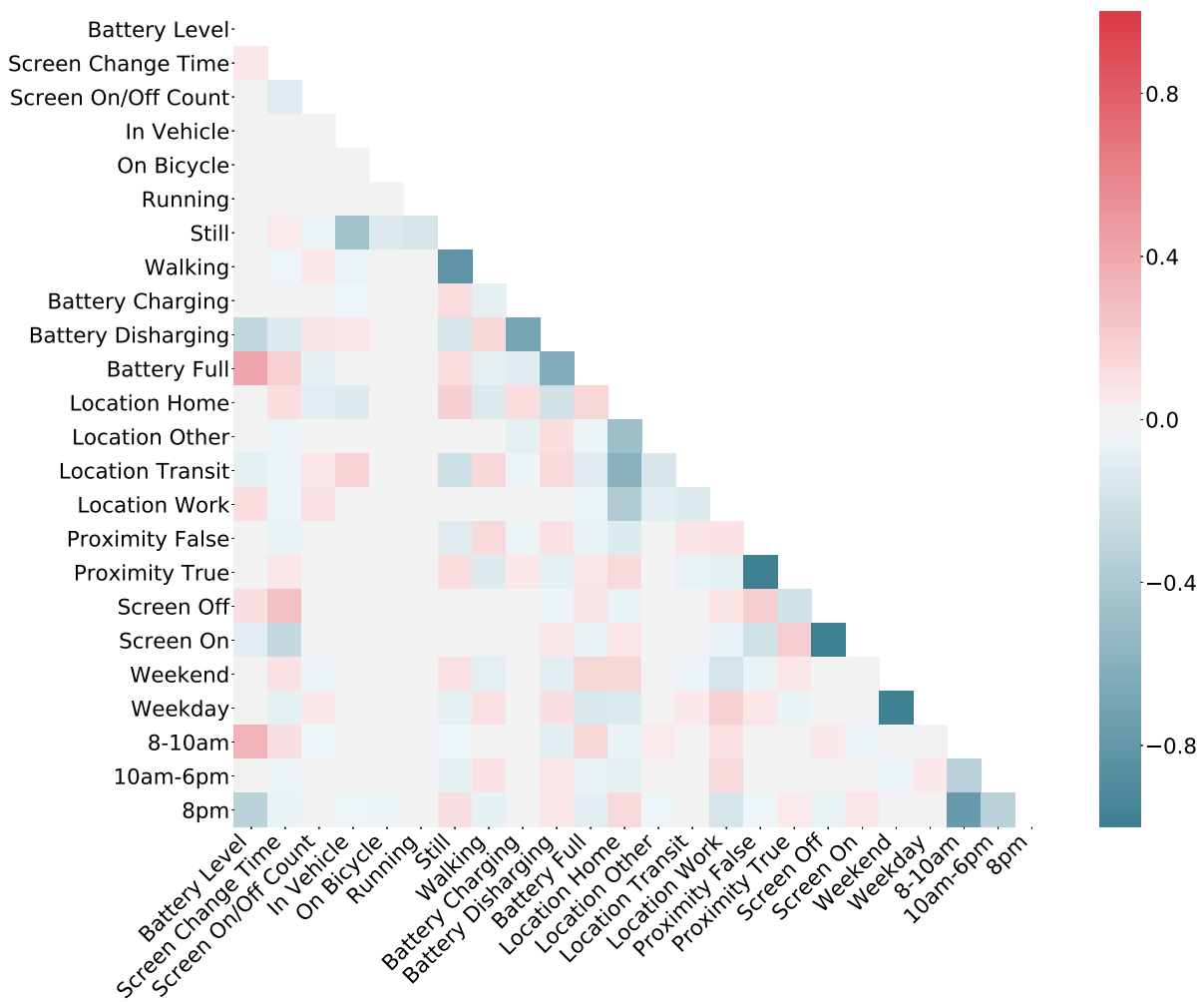


Fig. 3. Android: Correlation analysis of the different contextual factors

Table 1. **iOS**: Coefficients from a multi-linear regression model to understand to understand the influence of the different contextual factors and the different receptivity metrics

	<i>Coefficients</i>					
	<i>Just-in-time response</i>		<i>Conversation engagement</i>		<i>Time delay</i>	
(Intercept)	0.245	***	0.190	***	120.827	***
Activity (in a vehicle)	0.018		0.014		-10.339	
Activity (on a bicycle)	-0.009		-0.002		-9.628	
Activity (on foot)	0.045	***	0.040	***	-15.861	**
Activity (running)	0.020		-0.026		13.877	
Battery level	0.001		0.000		3.134	
Battery status (charging)	0.012		0.014		-0.764	
Battery status (full)	-0.039	**	-0.031	*	10.568	
Time since last lock	-0.011	**	-0.007		3.581	
Lock state (locked)	-0.027	**	-0.025	**	-2.909	
Location (Other)	-0.027	*	-0.018		5.484	
Location (Transit)	-0.008		-0.003		-0.773	
Location (Work)	-0.015		-0.015		19.214	*
Unlock count	0.013	**	0.006		-2.800	
Wi-Fi (not connected)	-0.034	***	-0.026	**	8.890	
Weekend	-0.027	**	-0.027	**	11.376	*
Morning message (8-10am)	-0.015		-0.002		26.884	***
Afternoon message (10am-6pm)	0.078	***	0.013		-25.834	**

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

S3 MULTI-LINEAR REGRESSION ANALYSIS

In this section we perform multi-linear regression analysis to understand which *contextual* factors are significant in a more controlled sense. Since we had different sensors for iOS and Android users, we performed the multi-linear regression analysis for the two device types separately. Table 1 shows the coefficients for the contextual factors with the different receptivity metrics (just-in-time response, conversation engagement and response delay). Table 2 shows the coefficients for the Android users.

Table 2. **Android:** Coefficients from a multi-linear regression model to understand to understand the influence of the different contextual factors and the different receptivity metrics

	<i>Coefficients</i>		
	<i>Just-in-time response</i>	<i>Conversation engagement</i>	<i>Time delay</i>
(Intercept)	0.127 *	0.080	135.176 ***
Activity (in a vehicle)	-0.009	-0.033	17.630
Activity (on a bicycle)	0.140	0.038	-75.178
Activity (running)	-0.043	-0.114	-13.858
Activity (walking)	0.012	-0.005	-18.614
Battery level	-0.001	0.002	-0.858
Battery status (charging)	0.034	0.035	64.955 ***
Battery status (discharging)	0.082 *	0.073 *	-26.642
Location (Home)	0.157 ***	0.149 ***	-18.070
Location (Other)	0.028	0.058	23.536
Location (Transit)	0.152 ***	0.147 ***	-35.109 ·
Location (Work)	0.125 *	0.134 **	-14.514
Proximity (false)	-0.071 ***	-0.070 ***	23.332 **
Screen change time	-0.025 *	-0.019 *	-4.503
Screen state (Off)	-0.030	-0.025	-0.720
Screen on/off count	0.020 *	0.021 *	-2.447
Weekend	-0.018	-0.002	-19.420 *
Morning message (8-10am)	-0.043 ·	-0.019	24.587 *
Afternoon message (10am-6pm)	0.058 ·	0.059 *	-59.742 ***

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, · $p < 0.1$

S4 EXAMPLE OF INTERVENTION CONVERSATIONS

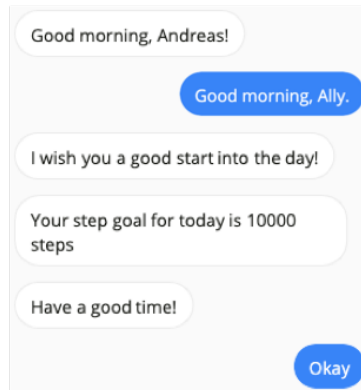
In this section we provide some examples of the actual conversations that took place between the coach (Walkie) and the participants. To maintain user privacy, the names in the conversations are fictitious.

The intervention messages were originally in German. We have translated some of them to English and include them here. It is important to note that we used Google Translate to translate the messages, and being a literal translation, they might not have the same effect or meaning as the original German ones.

We show full day conversations between Walkie and 3 participants – Andreas, Lisa and Sarah (Fig 4, 5 and 6, respectively). Two of these participants received the 'Self-monitoring' prompt (hence a total of 3 conversations in the day). Andreas did not receive the prompt (total of two conversations in the day).

To highlight variations with a single participants, we include examples of the Goal achievement (evening) message on three different days, in Fig. 7. The participant achieved the step goals on two days.

Goal setting (Morning)



Goal achievement (Evening)

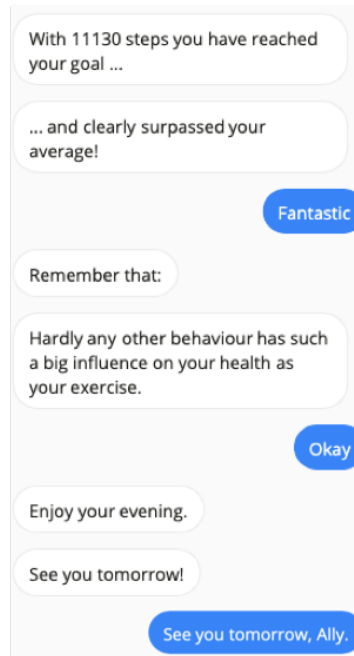
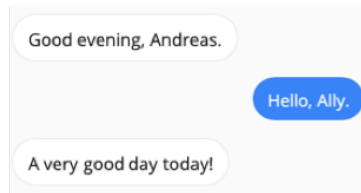


Fig. 4. Conversation with Andreas: Two initiating messages in the day

Goal setting (Morning)

Good morning, Lisa!

Good morning, Ally.

I wish you a good start into the day!

Your step goal for today is 6148 steps

Have a good time!

Okay

Self-monitoring (Afternoon)

Hey Lisa

What about your physical activity today?

Not great

Okay, no problem.

Let's have a look at this together

You've walked 4346 steps so far.

Since your goal is 6148 steps, you need another 1802 steps.

Okay

Ready for a fun fact on physical exercise?

Sure

Exercise can be very helpful against stress!

Take a few conscious steps today when you're stressed out.

It helps you clear your mind and recharge your batteries!

All right.

It works very well!

Try it out as soon as you get a chance!

I'll do

Very good.

Have a nice day, Lisa!

Bye

Goal achievement (Evening)

Good evening, Lisa.

Hello, Ally.

You walked 5417 steps today.

That's below your goal of 6148 steps.

Okay

No problem, Lisa.

Some days you're more active than others.

Do you sometimes feel bad on days when you're less active?

No

There's no reason either.

Days with less activity cannot be avoided.

Right.

That's it, too!

See you tomorrow, Lisa.

See you tomorrow, Ally.

Fig. 5. Conversation with Lisa: Three initiating messages in the day

Goal setting (Morning)

Good morning, Sarah!

Good morning, Ally.

I wish you a good start into the day!

Your step goal for today is 10000 steps

Have a good time!

Okay

Self-monitoring (Afternoon)

Hello Sarah

Are you having a good day?

Yes

I'm glad!

I'm sure you were already active today!

Let's have a look!

You've already walked 9760 steps! So you still need 240 steps to reach your goal.

Cool

Think how you can reach your goal today.

For 240 steps you should plan about 3 minutes.

I'll do

For example, place a bet to spice up your target.

Tell me more.

Tell a friend about your step goal.

Then bet whether you reach your goal or not ...

Okay

Make up a bet for yourself.

If you lose, you'll have to buy a cup of coffee tomorrow, for example!

That's a good idea!

Have fun!

I'll get back to you soon!

So long, Sarah.

So long, Ally.

Goal achievement (Evening)

Good evening, Sarah.

Hello, Ally.

A great day!

You've reached your goal.

And you walked more than an average day!

Nice!

You're on the right track.

Your body is already feeling your activity today.

How's that?

Exercise can immediately lower your blood pressure, your blood fat levels and your blood sugar.

And thus reduces your risk of developing coronary heart disease or diabetes.

Cool

And hey:

You got one Swiss Franc on top!

Money makes the world go 'round

Have a relaxed evening!

See you tomorrow!

See you tomorrow, Ally.

Fig. 6. Conversation with Sarah: Three initiating messages in the day

Goal achievement (Evening 1)

Good evening, Thomas.

Hello, Ally.

I'll have a quick look at your step data ...

You walked 4478 steps.

You haven't been able to reach your goal yet.

Okay

It happens.

Everyone has more active and more inactive days.

Sure

By the way, exercise can not only increase your physical activity ...

... but also your mental performance!

Oh, yeah?

Activity promotes blood circulation in the brain and your nerve cells are better supplied with nutrients.

So take a few steps from time to time when you need to concentrate on your work.

This increases your productivity!

Okay

Have a nice evening.

See you tomorrow, Thomas.

See you tomorrow, Ally.

Goal achievement (Evening 2)

Good evening, Thomas.

Hello, Ally.

Congratulations.

You've reached your step.

Today you were able to increase your movement successfully!

Cool

Any advice if you want to watch your weight?

Yes

Your movement can help you control your weight.

Studies suggest that a daily 1.5 kilometer walk helps lose weight.

Okay, cool.

Have a relaxed evening!

See you tomorrow!

See you tomorrow, Ally.

Goal achievement (Evening 3)

Good evening, Thomas.

Hello, Ally.

A very good day today!

With 8492 steps you have reached your goal ...

... and clearly surpassed your average!

Fantastic

Remember that:

Hardly any other behaviour has such a big influence on your health as your physical activity.

Okay

Enjoy your evening.

See you tomorrow!

See you tomorrow, Ally.

Fig. 7. Conversation with Thomas: Different goal achievement messages on three different days.

REFERENCES

- [1] Abhinav Mehrotra, Mirco Musolesi, Robert Hendley, and Veljko Pejovic. 2015. Designing content-driven intelligent notification mechanisms for mobile applications. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*.

- ACM, 813–824.
- [2] Abhinav Mehrotra, Veljko Pejovic, Jo Vermeulen, Robert Hendley, and Mirco Musolesi. 2016. My phone and me: understanding people’s receptivity to mobile notifications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 1021–1032.
 - [3] Martin Pielot, Bruno Cardoso, Kleomenis Katevas, Joan Serrà, Aleksandar Matic, and Nuria Oliver. 2017. Beyond interruptibility: Predicting opportune moments to engage mobile phone users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 91.
 - [4] Martin Pielot, Tilman Dingler, Jose San Pedro, and Nuria Oliver. 2015. When attention is not scarce-detecting boredom from mobile phone usage. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 825–836.
 - [5] Hillol Sarker, Moushumi Sharmin, Amin Ahsan Ali, Md Mahbubur Rahman, Rummana Bari, Syed Monowar Hossain, and Santosh Kumar. 2014. Assessing the availability of users to engage in just-in-time intervention in the natural environment. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 909–920.
 - [6] Tilo Westermann, Ina Wechsung, and Sebastian Möller. 2016. Smartphone Notifications in Context: A Case Study on Receptivity by the Example of an Advertising Service. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, New York, NY, USA, 2355–2361. <https://doi.org/10.1145/2851581.2892383>