A Multi-perspective Analysis of Social Context and Personal Factors in Office Settings for the Design of an Effective Mobile Notification System

SEYMA KUCUKOZER CAVDAR and TUGBA TASKAYA-TEMIZEL, Middle East Technical University, Turkey
MIRCO MUSOLESI, University College London, United Kingdom, The Alan Turing Institute, United Kingdom, and University of Bologna, Italy
PETER TINO, The University of Birmingham, United Kingdom

In this study, we investigate the effects of social context, personal and mobile phone usage on the inference of work engagement/challenge levels of knowledge workers and their responsiveness to well-being related notifications. Our results show that mobile application usage is associated to the responsiveness and work engagement/challenge levels of knowledge workers. We also developed multi-level (within- and between-subjects) models for the inference of attentional states and engagement/challenge levels with mobile application usage indicators as inputs, such as the number of applications used prior to notifications, the number of switches between applications, and application category usage. The results of our analysis show that the following features are effective for the inference of attentional states and engagement/challenge levels: the number of switches between mobile applications in the last 45 minutes and the duration of application usage in the last 5 minutes before users' response to ESM messages.

CCS Concepts: • **Human-centered computing** → **Mobile computing**; *Human computer interaction (HCI).*

Additional Key Words and Phrases: responsiveness, social context, personal factors, notifications, attentional states

1 INTRODUCTION

Modeling social context in human-computer interaction design studies have recently gained attention, where context is not only conceived as physical objects and people but also as social norms, which influence both individuals and organizations [39]. In ubiquitous systems design, interaction of the actor with other actors, and organization and actor's perception of these relationships might become relevant [68].

An application area of increasing societal and commercial importance is healthcare and well-being. More specifically, health interventions have been increasingly delivered via mobile phones, thanks to their ubiquity and the possibility of inferring contextual information by means of the embedded sensors. Numerous studies have been focusing on when and how to deliver intervention messages effectively [13, 35, 54, 67]. In an office setting, there are many factors that can affect the responsiveness of individuals to these messages, in particular their *engagement* and *challenge* levels [41, 80].

The possibility of responding to a notification may change with the degree of an individual engaged or challenged with his/her work when the notification arrive. Hence, it is important to measure the *in-situ* engagement or challenge levels of office workers during a work day for mobile notifications to be effective on users. *Work engagement* is defined as the degree of involvement/distraction of employees with their work [40], and *challenge level* can be described as the degree of the mental effort that should be exerted to complete a task [41]. In line with the latter study [41], in this paper, we used these labels as referent terms, and they do not fully characterize the definitions. More precisely, challenge level is used to refer to user response for the question regarding how

Authors' addresses: Seyma Kucukozer Cavdar, kseyma@metu.edu.tr; Tugba Taskaya-Temizel, ttemizel@metu.edu.tr, Middle East Technical University, Dumlupinar Bulvari, Ankara, Turkey; Mirco Musolesi, m.musolesi@ucl.ac.uk, University College London, Gower Street, London, United Kingdom, The Alan Turing Institute, London, United Kingdom, University of Bologna, Bologna, Italy; Peter Tino, P.Tino@cs.bham.ac.uk, The University of Birmingham, Birmingham, United Kingdom.

challenged a user (with a Likert scale where 5 indicates very challenged and 1 not at all). Responsiveness to the break reminder notifications in this study is measured based on the response styles to the experience sampling method (ESM) questions, for example whether the user consistently chooses the same answers, i.e., positive or negative. It does not measure the actual adherence to notification content itself. We consider user response patterns on ESM regarding engagement and challenge level questions. When a notification message arrives, a user may respond to questions in a variety of ways: a user (1) might not answer the ESM questions due to her level of challenge related to her work; (2) might frequently rate herself as highly challenged; or (3) might frequently rate herself as scarcely challenged. These response styles were previously linked to users' personality and social factors in the literature. In contrast with the existing studies, we investigate the relation between responsiveness and office related factors i.e. if there is a frequent distraction in an office setting, we might observe high variation or frequent ratings of low engagement in user responses.

Attentional states, as well as engagement and challenge levels of office workers, can indirectly be inferred via mobile phones. For example, an increase in smartphone usage may be a sign of boredom since most users prefer to use mobile phones when they feel bored [10, 43, 61]. Similarly, their interaction with computers (which applications they use and how long) might also reveal boredom [7, 41]. The context and activity of users are other important factors [31, 45, 55, 70]. When people are moving or during an activity transition, they are more likely to respond to messages. However, these factors have mainly been investigated considering all daily activities of individuals without focusing on the specific case of a work environment in the literature. We believe that the effectiveness of systems targeting specific categories of users, such as employees of a company, can be improved through the understanding (and the exploitation) of the characteristics of the individual and social context in these settings.

Responsiveness to well-being related notifications can be affected by the health history of users. It is well-known that discomfort caused by a disorder determines the level of adherence to treatment [69]. In the mobile context, users may be more likely to respond to messages sent by break reminder applications, if they experienced musculoskeletal discomfort due to their sedentary life. Awareness about the consequences can be defined as "a disposition to become aware of the potential consequences of one's acts during the decision-making process" [64], and it is also an important factor for responding to well-being related messages. A recent study showed the effects of self-regulation and habit strength on the sedentary behaviors of knowledge workers [37]. Higher awareness or self-regulation may increase the responsiveness to break-reminder notifications. Besides, social factors, such as subjective norms, have been found as a precursor related to behavioral intention [2, 3]. A subjective norm is the perception towards performing a behavior with influences by others who are important to the one performing the behavior [20]. Recent studies showed that office employees are influenced by their co-workers regarding prolonged sitting behavior [71] or performing physical activity [24]. Hence, office workers might also be influenced to take rest breaks by their colleagues. Finally, the number of colleagues in the same office might be another factor for both responsiveness and work engagement/challenge levels. It has been showed that office type (whether employees work in shared or private offices) has a significant effect on distractions [51] and also on sitting time [52]. We do not directly measure the office type, instead, we will investigate whether the number of colleagues in the same office has also an effect on work engagement/challenge levels and the responsiveness to mobile notifications. The number of colleagues in the same office could be an indirect indicator of office type.

The main contribution of this paper is the design, implementation and evaluation of a framework for the inference of engagement/challenge levels of office workers and their responsiveness to well-being related mobile notifications. The proposed framework can be generalized to other application domains. We investigate responsiveness with several metrics such as acquiescence, disacquiescence, and extreme response style (negative and positive). We conducted a novel user study based on a purpose-built mobile application, which sends break reminder notifications and suggests users exercises, which can be taken during their rest breaks. More specifically, in this paper we investigate the following research questions:

- **RQ1: How is the musculoskeletal discomfort of office workers related to their responsiveness to break-reminder notifications?** It has been shown that the level of pain is a determinant to the adherence or compliance to treatment in clinics [69]. We tried to address a research hypothesis that has been proposed in a more recent study [49] regarding the relationship between the level of musculoskeletal discomfort and the compliance to the stretch breaks.
- **RQ2: How is the awareness of office workers about rest breaks related to their responsiveness to break-reminder notifications?** Previous studies showed that individual control or self-regulation to perform a behavior is also important for taking regular breaks from work or prolonged sitting [37, 49, 71]. Hence, we hypothesize that there is a relation between awareness about rest breaks and responsiveness to break-reminder notifications.
- **RQ3: How are office-related factors associated with office workers' responsiveness to break-reminder notifications?** Previous studies showed how behaviors of office workers could be affected by external factors such as peers and colleagues [24, 48, 71]. In addition, office type (shared or private office) has an effect on distractions observed by employees or their sitting behavior [51, 52, 65]. Hence, we hypothesize that responsiveness to break-reminder notifications can be affected by office-related factors.

We also explored the following research questions, which were proposed in the previous studies but not investigated specifically in office settings or in terms of responsiveness:

- **RQ4a: How is mobile application usage of office workers related to their responsiveness to break-reminder notifications?** It is known that responsiveness to mobile notifications can be inferred with application usage [42, 59]. In those studies, responsiveness was measured as whether users respond to a notification or not. We will explore the relationship between mobile application usage and responsiveness to break-reminder notifications with several responsiveness metrics.
- **RQ4b: Which application usage metrics are related to in-situ engagement/challenge levels of office workers?** The metrics such as duration of application usage, number of applications used or number of switches between applications can be related to in-situ work engagement/challenge levels of office workers since previous studies showed that attentional states could be predicted with such metrics [42, 43, 60]. We will explore the relation in detail by means of a novel technique that incorporates both within- and between-subjects and has not been used previously. We consider the fact that our data comes from a repeated-measures design in this study; hence, apply a more appropriate method for analyzing the relations.
- **RQ4c: How can we build a model for inferring attentional states and engagement/challenge levels of office workers using application usage metrics by considering the "cold start problem", the variety in the number and characteristics of the responses, and repeated-measurement nature of the data? How is this model comparable to individual and general models?** There is a need for personalization for making inferences about attentional states and engagement/challenge levels, since each person has a response bias in questionnaires, which can be defined as the systematic tendency to responding questionnaire items. Hence, each person might have different tendencies for responding to the questionnaire items. Their relative differences need to be considered in the models. In our case, it affects user responses to engagement and challenge levels. However, at the early stages of the design of personalized models, the cold-start problem may occur since the number of data points is relatively small. As the data points increase, models can learn from individual data and get more accurate results. We will discuss a solution for the cold start problem and show its effectiveness by comparing its performance with individual-level models. Specifically, we consider random forest methods for comparison, which have been extensively used in recent studies [74, 81]

Our solution is based on a novel generic *population-* and *individual-level* model for unbalanced and limited amount of data. The experiment was conducted with 31 participants in 10 workdays during their work hours. We sent rest break-reminder notifications to the participants and collected their in-situ engagement and challenge levels with experience sampling method (ESM) questions via the mobile sensing application. The responsiveness metrics were extracted from the answers to the engagement/challenge level questions. The mobile sensing application collected application usage data in the background. Then, we implemented metrics such as duration of application usage, number of applications, number of switches between applications, or application category usage using the mobile application data collected. Musculoskeletal discomfort, awareness, and office-related factors were collected with a questionnaire. We performed bi-variate correlation analyses among the variables.

Then, we focused on modeling in-situ engagement/challenge levels with mobile application usage. Context information such as time and location of the participants and the activity information have an impact on inference on engagement and challenge levels as shown in [45, 59]. However, this paper is focused solely on inference using mobile application information and statistics derived from this data. Unlike the previous studies, we adopted a recently proposed correlation metric called repeated-measures correlation, which is designed specifically for repeated-measures studies [4]. We analyzed both the short-term application usages (e.g. 5-10-15 minutes) and the long-term application usage (e.g. 30-45-60 minutes) for the inference of engagement/challenge levels.

Finally, we modeled engagement/challenge levels, and attentional states of office workers with generalized linear mixed models (GLMMs) [9] using a Markov chain Monte Carlo method. GLMM incorporates fixed and random effects together so that it enables to fit model parameters at both population and individual level. In similar studies in the literature, classification methods such as random forest models or support vector machines are generally used for modeling all participants' data as general (or generalized) models [80]. However, with such approaches, *the relation among measurements from the same participant is ignored*, and the assumption of statistical independence is violated. As a remedy, individual models are built for each participant. As a consequence, a significant amount of data is required for each individual model in order to make them work effectively. In the paper, we present the validation of our analyses using different sub-samples of the data set and a comparison of GLMM with general and individual random forest classifiers.

The main findings of our study can be summarized as follows: (1) musculoskeletal discomfort and awareness about rest breaks have positive effects on the responsiveness to break-reminder mobile notifications; (2) people who take rest breaks more frequently felt less musculoskeletal discomfort while working than those who take rest breaks less frequently; (3) subjective norms on having rest breaks in office environments (i.e. the degree of being affected by colleagues to have rest breaks) are related to the number of office workers in the same office; (4) mobile application usage is negatively related to engagement/challenge levels of office workers; (5) responsiveness to break-reminder notifications are related to mobile application usage; (6) attentional states of office workers can be explained with mobile application usage: a higher amount of application usage might be an indicator of boredom in workplaces or vice versa; (7) GLMMs are effective for individual-level modeling particularly when working with unbalanced and limited amount of data. Overall, we believe that the proposed framework can be applied to a variety of office environments, given its generality.

2 RELATED WORK

In this section, we present the related studies including "Responsiveness and Variability Measures", "Subjective Norm in Office Setting and Office Type", "Musculoskeletal Discomfort and Awareness about Sitting Behavior" "Work Engagement and Challenge", and "Detecting Attentional States".

2.1 Responsiveness and Variability Measures

Responsiveness is commonly used as a synonym of receptiveness, or attractiveness specifically in the mobile computing domain [12]. It is simply the degree of whether a mobile system user answers or reacts to the prompts generated by the system or not. A common method for measuring responsiveness is Experience Sampling Methodology (ESM), which allows capturing and recording human behavior as it happens in their natural settings [34].

Several studies focus on the inference of responsiveness to mobile notifications using mobile phone related features, such as application usage. Pielot *et al.* [60] stated that users are more open to receiving phone notifications if they have recently used their devices. Similarly, Mathur *et al.* [42] investigated the effects of several features, such as the number of applications used and the amount of time spent interacting with phone in the last hour, on predicting user involvement with mobile phones. Their results showed that involvement increased as the number of applications used in the last hour increased.

Responsiveness has mainly been studied with respect to one dimension: whether a user responded to a survey or not. Although users may appear to be attentive to ESM surveys at first, their answers might be inaccurate, repetitive or random. So, response style of users should be understood in order to make more reliable inferences. Response style is defined as "a respondent's tendency to responding systematically to questionnaire items regardless of the content" [56]. The most common response styles are acquiescence or disacquiescence (the tendency to agree or disagree to an item), extreme response style (the tendency to use the extreme categories), and middle response style (the tendency to use the middle category). These measures have been used in this type of studies for a long time in order to test survey validity [15]. Lately, response styles or survey-taking patterns have been found to be related to behavioral measures such as non-cognitive skills [30] or personality [14, 28, 29]. Based on these recent findings, we decided to investigate whether the styles of the responses to the ESM questions are related to the engagement and challenge levels of office workers.

Moreover, entropy is used to quantify the degree of homogeneity of responses. This measure can be quite helpful to understand whether users keep choosing the same items repeatedly (such as choosing the extreme points of a Likert scale) or uniformly. To the best of our knowledge, no study has addressed the relationship between the variety of responses in ESM surveys and social factors in the user environment. Quantifying the influence of the factors behind the heterogeneity in the responsiveness can help researchers to design better measurement scales or to better understand the performance results of models.

2.2 Subjective Norm in Office Setting and Office Type

Subjective norm (social norm) is defined as "person's perception that most people who are important to him/her think he/she should or should not perform the behavior in question" [20]. It can be briefly described as the answer to the question: "Do other people want me to do it?" [27]. For instance, suppose that a social norm exists whereby office workers are expected to spend all their time in their offices during working hours. One of the reasons upholding this social norm is that the company manager appears to criticize office workers if they leave the office premises often. Hence, office workers may not be willing to have rest breaks. This problem can be alleviated if the attitude of the manager changes and the manager encourages office workers to take regular health breaks. It also plays an important role in the intention to use or adopt a new technology [73]. Hence, subjective norm is commonly used in the studies related to technology acceptance. For example, a recent study [48] was conducted in an office environment and focused on the effects of social norms on the adoption of mobile applications for promoting physical activity. The authors found that social influence is an effective factor for using such applications. Another study [71] showed that office employees are affected by their colleagues as far as their sitting behavior is concerned. Similarly, George *et al.* [24] emphasized the importance of being a part of a social group or being able to interact with others as a motivator for performing physical activity in a university

work environment. Hence, in the present work, we aim at analyzing the direct relationship between subjective norm and user responsiveness. We specifically measured to which degree the participants were influenced by their colleagues or managers to have rest breaks.

There have been studies investigating the effects of office type on sedentary behavior [52] and distraction [65]. Office type refers to the fact that participants have either shared (i.e., they share their offices with other colleagues such as open offices) or private (i.e., they work alone) offices. Mullane *et al.* [52] showed that employees in private offices have a higher amount of sitting time compared to those in open offices. They also observe that employees in open offices might be more receptive to social cues than those in private offices. Seddigh *et al.* [65] discussed the office type effects and concluded that the relationship is stronger among employees in open offices than those in cellular offices. Finally, Morrison and Macky [51] showed that distraction is higher in shared offices. Hence, these studies give us an idea about the importance of office-related factors on perceived distraction (which might be considered as a reverse of work engagement), and also on responsiveness to rest break reminders. In our study, we did not directly ask the office type of the participants as in existing studies since we aim at investigating whether the number of people with whom they are sharing their offices has an effect on their responsiveness. The number of colleagues might be an indicator of the office type of the employees: a high number of colleagues might imply a shared office, whereas a low number of colleagues might indicate a more private office, or, more in general, a lower amount of distractions caused by co-workers. The number of colleagues is also a strong indicator for social influence, which was previously used for example for inferring perceived behavior among peers [77]. Our study contributes to the existing literature by showing the relationships between the number of colleagues and responsiveness to notifications.

2.3 Musculoskeletal Discomfort and Awareness about Sitting Behavior

Musculoskeletal discomfort of participants in office settings is discussed in a limited number of studies. Monsey *et al.* [49] found that although the participants of their study were instructed to take breaks, most of the time they did not comply to take breaks. They concluded that there may be different factors that affect the decision of individuals for taking breaks. For example, an interesting point to be investigated is the relationship between musculoskeletal discomfort and compliance to the break-reminder programs/applications.

The number of studies in awareness about sitting behavior is limited in the literature. For example, in a qualitative study [17], participants stated that they were not aware of how much time they have been working most of the time, and reminders from applications for taking breaks could improve their productivity. In another qualitative study, participants stated that lack of awareness related to physical activity affects it [38], or reversely, growing awareness could be a motivator for it [24]. van Dantzig *et al.* [71] also showed that the internal control toward sitting behavior was low for most of the participants of their study. Similarly, Wallmann-Sperlich *et al.* [75] stated that individuals, who believe that sitting for long periods would not be harmful, actually sit for a longer amount of time than individuals who do not. Those results show that personal beliefs and awareness regarding a specific behavior actually affect performing the behavior, and they show the importance of internal factors such as awareness regarding taking breaks. Luo *et al.* [37] recently explored the self-regulation and habit strength for preventing prolonged sitting via a mobile application and found that stronger self-regulation led to quicker responses to notifications. It is the only study that investigates the relation between self-regulation and responsiveness. In our study, we included several metrics that have not been investigated before and made an effort to show how awareness is related to the variability in engagement/challenge level responses.

2.4 Work Engagement and Challenge Levels

Work engagement is described as an active and positive state with a high level of energy, strong involvement, and full concentration [63]. Simply, engagement is the degree of employees involved with or distracted from work [40]. On the other hand, challenge is defined as the mental effort that should be exerted to complete a task [41].

Mark *et al.* [40] investigated the relationship between work engagement and several factors such as face-to-face interactions, Facebook use, e-mail use, application use. They found that the duration of Facebook use and e-mail/calendar use are negatively related to engagement levels, whereas face-to-face interaction are positively related. A very recent work [13] investigated the effects of focus on the adherence to physical activity. They found that when individuals have a high focus on the ongoing task, they are less likely to notice the mobile intervention, or comply with the intervention.

Nduhura and Prieler [53] showed that social media use positively affects workers, energizes them, so that increases their productivity. Although it is not related to work engagement directly, it is worthwhile to mention that Mirjafari *et al.* [47] developed a model for assessing low and high performance in the workplace through mobile sensing. They collected activity, location, phone usage (lock/unlock), and light level through mobile application, heart rate and stress through a wearable device, and time spent in work and time spent at a break through a Bluetooth device. Their results showed that higher performers unlock their phones less but they are more active than low performers. They built a XGBoost classifier for classifying the performers and the performance of the model is presented as AUROC=.83.

As far as work challenge is concerned, Mark *et al.* [40, 41] presented a significant effect of Facebook use, e-mail use, and task switching between computer applications on challenge levels of knowledge workers. According to their results, the duration of Facebook use is negatively related to challenge levels, whereas the duration of e-mail use and the number of application switches are positively related. However, the results of those studies were based only on computer activities. We extend these studies by focusing on mobile application activities and other personal/social factors.

Our study differs from the previous work from a methodological point of view since it is based on the evaluation of work engagement and challenge, which are important to understand the availability of office workers for delivering notifications [13, 41]. We measured *in-situ* work engagement and challenge levels, i.e., we captured the engagement and challenge levels when employees were currently working with a mobile application. Hence, we obtained different levels of work engagement and challenge measures at different time intervals, which is more representative than using single point measures. We investigated the relation among in-situ measurements of work engagement/challenge levels, personal factors and social norms in workplaces.

2.5 Detecting Attentional States

Boredom is described as "lack of stimulation or inability to be stimulated thereto" [18]. It comprises a penetrating deprivation of interest and difficulty of focusing on the ongoing task [21]. Individuals mostly seek a way to escape from the boredom state [25]. To date, several efforts have been made for predicting attentional states, including boredom. Physiological sensors [58] or logging computer activities [7, 41] are some examples of boredom detection techniques widely used in previous studies. More recently, mobile devices have been used for collecting continuous data about users' engagement and attentional states.

Mark *et al.* [41] proposed a theoretical framework representing attentional states in workplaces. They measured engagement and challenge levels of workers in workplaces via ESM questions, then separated the attentional states into four categories: (1) "*rote*" represents highly engaged (the top two of the ratings), not challenged (the bottom two of the ratings); (2) "*focus*" represents highly engaged (the top two of the ratings) and challenged (the bottom two of the ratings); (3) "*bored*" represents low engagement (the bottom two of the ratings) , not challenged (the bottom two of the ratings) ; and (4) "*frustrated*" represents low engagement (the bottom two of the ratings),

high challenge (the top two of the ratings). We normalized the scores and excluded the mid-range values in line with [41] We used a 5-point Likert scale (unlike [41]). Then, they investigated which online activities are related to attentional states and how they are related. Their results showed that the type of online activity determines the attentional states of workers. For example, workers are usually in "bored" or "rote" states when viewing/writing e-mails, whereas they significantly spend less time in "focused" state when using Facebook or Web-browsing.

Pielot *et al.* [61] investigated which mobile phone features are informative for detecting boredom. They stated that users are more likely to use a higher number of applications when they are bored. Similarly, in another study [59], the recency of communication, the intensity of phone usage, proximity and hour of the day were found related to detecting boredom. They also found that when boredom was sensed by mobile phones, sending proactive recommendations significantly attracted users' interests. Matic *et al.* [43] also found that the number of launched applications is a predictive feature for detecting boredom on smartphones. LiKamWa *et al.* [36] developed a mobile phone application, which predicted the moods of users with smartphone usage patterns by fitting individual and general level classifiers. They found that phone calls and categorized application usage were strong predictors for inferring mood. Zenonos *et al.* [80] focused on predicting mood levels with wearable sensors and smartphones in work environments. Instead, in our study we focused on predicting work-related variables, i.e., engagement and challenge levels, and attentional states (focused on work or bored) of knowledge workers during work hours.

With respect to the previous work, our contribution is twofold. First of all, we consider a variety of individual, social and contextual factors, in order to have a more comprehensive understanding of the phenomena under consideration. Secondly, our contribution is above all methodological: we consider a repeated-measures design for modeling without violating any statistical assumptions regarding the nature/distribution of responses. We believe that this is very important in order to ensure the statistical validity of these studies. We compared the results with the models, which are widely used in the literature, i.e., fitting with general models and individual models such as random forest models.

3 METHODOLOGY

We designed an experiment for quantifying the factors related to work engagement and challenge levels of knowledge workers and their responsiveness to the ESM messages. Mobile-based ESM was adopted since the data obtained by ESM tend to have higher validity and less bias compared to other methods [57]. Despite some challenges (e.g. recruiting participants, sampling time, or technical challenges), ESM is a strong and powerful methodology for capturing users' natural feelings and thoughts. The overall framework is shown in Figure 1 and the flowchart of the experiment is given in Figure 2.

3.1 Pre-Experiment Questionnaire

A pre-experiment questionnaire was delivered to the participants in order to collect demographic information and to understand work routines and regular break times, musculoskeletal discomfort, awareness about rest breaks, and office related factors. The demographic questions concern information about age, gender, occupation, job title, and educational background of the participants. The work start/end hours and the participants' favorite times for breaks, type of the breaks (e.g. social, tea/coffee, lunch), location of the breaks (in the same office, outside the office on the same/different floor, or outside the building) and availability for an office exercise in those breaks were also collected in order to understand work routines of the participants.

Subjective norm (SN) was collected with two statements adapted from [78]. The items of the statements were scaled between one ("Strongly disagree") and five ("Strongly agree"):

- My manager(s) influence(s) my intention to have a rest break.
- My colleagues will encourage me to have a rest break.

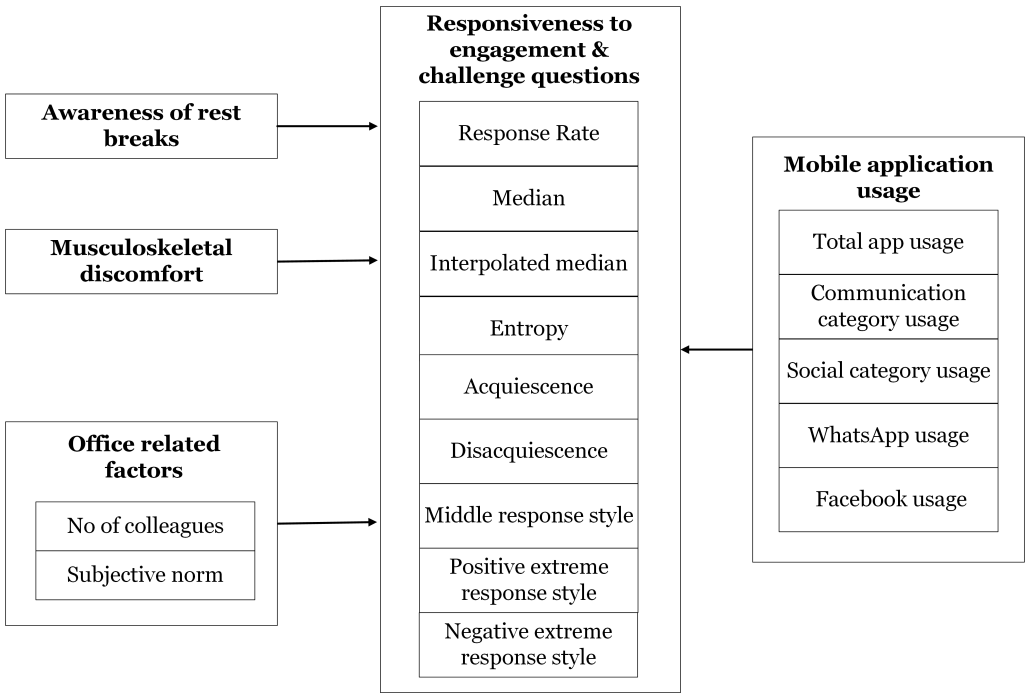

Fig. 1. Proposed research framework.

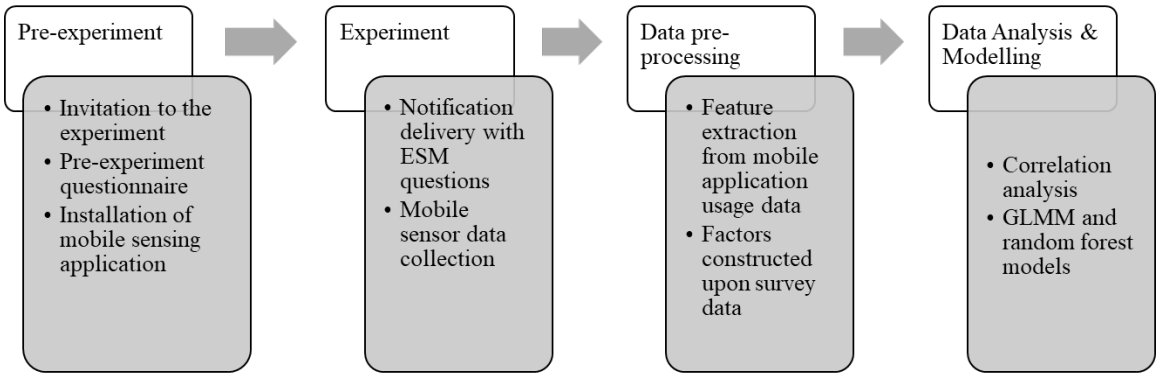

Fig. 2. Steps of the experiment and data analysis.

The questions related to musculoskeletal discomfort, and awareness about rest breaks and doing office exercises were designed with the help of a specialist in behavioral psychology and a physical therapist. The responses were scaled between one ("Never") and three ("Very frequently"). More precisely, the questions were the following:

- Do you feel pain or numbness in your hands, wrists or shoulders while using a computer?
- Do you take breaks in regular periods while working in front of a computer?
- Do you perform any hand, wrist or shoulder exercises or stretching movements in front of your computer?

The first question assesses the musculoskeletal discomfort, while the second and the third assess the awareness about rest breaks. Finally, we requested them to state the number of colleagues who they share their offices with.

3.2 Mobile Sensing Application

A mobile sensing application, which works on the Android operating system version 4.2 or higher, was developed for collecting data from participants and for reminding them to take rest breaks during work hours. The application collected screen activity and application usage information in the background. The application package names were only collected if the user agrees to share this information. The Android operating system requires an additional permission to collect application package names, hence, although it is a part of the informed consent, we had to obtain their additional permission for that purpose. The screen on/off times and user's screen presence were captured at all times. Hence, it was possible to detect when a user started using their mobile phone, and when the interaction with the phone ended.

3.3 Delivery of Notifications and ESM Questions

We prepared four questions related to in-situ engagement and challenge levels of the participants. Five-level Likert scale was used for the ESM questions where one indicating "Not at all" and five indicating "Very much". The questions were adapted from [41]:

- What is the degree of your engagement in the task that you are currently performing?
- What is the degree of your challenge in the task that you are currently performing?

The mobile sensing application was the medium for delivering reminders to participants. Each reminder consisted of ESM questions; hence, we used ESM questions as the reminders themselves. The reminders were dispatched using a delivery algorithm that only considers the work hours of the participants. First, the time slots that the participant has a calendar event were discarded from the work start-end period. Then, two notification times from the preferable time slots of the participant were selected. The remaining four notification times were randomly chosen from the available time slots. If a participant did not state any preferred time in the pre-questionnaire, all six notification times were randomly picked from the work hours of that participant. Finally, when the algorithm selected notification times, it also ensured that there was at least one hour between the two consecutive notification times.

The reminder messages are displayed in the notification bar of the mobile devices as a regular notification. When the user taps on the notification, the reminder message and the ESM questions are displayed. The notification message includes a motivational message for taking a rest break (e.g., "Relax and take a break!", "Action time!"). The ESM messages are displayed below the motivational message on the phone screen. The questions are showed one by one: after the participant answers each question, the next question is displayed. At the end of the questionnaire, the user responses are sent to the main server after clicking the "Submit" button. The design of the notification was adapted from previous work, e.g., [37]. The ESM questions also included questions regarding if the notification message (for a 5, 10, or 30 minutes-break) arrives at an opportune time or not. The results about the timing of the notifications were not included in this study since they were not in the scope of this paper.

The intensity of the messages changes depending on the ringer mode of devices, i.e., it is based on system default settings. For example, if the device is in vibrate mode, the system notifies the user with a vibration. There were no re-prompts if users missed a prompt. A participant may (1) see and respond to the message, (2) see but not respond to the message, or (3) not see the message at all. In either situation, the message was deleted

after 15-minutes of arrival in line with the studies in the literature [46]. Non-interaction with the notification is interpreted as the fact that user is busy at that time, so deleting the notification does not annoy users.

3.4 Pilot Study

A pilot study was conducted in order to evaluate the design and execution of the experiment. Five participants having different mobile phones in terms of brands and Android versions, and working in different workplaces were included in the pilot study.

3.5 Data Collection Procedure

Before delivering the questionnaires and installing the mobile sensing application, the necessary approval was received by the Human Subjects Ethics Committee of the University. The participants were informed about the procedure of the experiment. They were told that the notifications were just for a reminder to take breaks, and they were requested to answer the ESM questions sent with the notifications. They were aware that the study was not aimed at changing their behaviors but at collecting data about their context and status via ESM when the notifications were received.

The duration of the experiment was set to 10 workdays for each participant. During the experiment, the participants received a maximum of six reminder messages per day. The participants who replied to at least 25% of the ESM messages were granted a coffee coupon as the only compensation for taking part in the experiment.

3.6 Participants

Participants were selected with a convenience sampling method: potential participants (who are office workers) were invited to join the experiment through several channels (from social media, by inviting graduate course students in an institute face-to-face, via e-mail, and by delivering leaflets). All invited participants were directed to the experiment web page in which they can access the pre-experiment questionnaire and the download link of the mobile sensing application. None of the participants were affiliated with the research group.

In total, 55 attempts were made for the pre-experiment survey, and 50 individuals responded in full. Forty-two of them successfully installed the mobile application. Eleven participants dropped out of the experiment, resulting in a total of 31 participants. Twenty of the 31 participants (64.52%) were male, and 11 of them (35.48%) were female. The average age of the participants was 31.52, with a minimum of 24 and a maximum of 42. The average work duration per day was 8.5 hours, with a minimum of 7 and a maximum of 10 hours. The job titles of the participants were varied. Eleven of the participants (35.50%) were engineers, nine of them (29.00%) were academics, six of them (19.40%) were specialists, four of them (12.90%) were managers, and one of them (3.20%) was technical personnel. Fourteen participants (45.20%) work in the private sector, 13 of them (41.90%) work in the government sector, 3 of them (9.70%) work as freelancers, and 1 participant (3.20%) was the owner/partner of a company.

Not everyone fully participated in all steps of the experiment: therefore, as a result:

- Nineteen of the 31 participants had a response rate of 25% or higher in ESM questions. These participants are called as *core participants* in the rest of the paper.
- The application package names were obtained from 24 of the 31 participants, and 14 of them were from the core participants.

3.7 Models

We used Kendall's Tau correlation for the analysis concerning RQ1-RQ4a in order to identify the relationships between responsiveness metrics and mobile application usage, awareness about rest breaks, and office-related factors.

For RQ4b, we used each user's ESM responses to assess the relationship between engagement/challenge levels and application usage parameters.

For RQ4c, we employed Generalized Linear Mixed Models using Markov chain Monte Carlo (MCMC) method. The details of the models used in the study are given in Appendix A. In order to illustrate the individual differences in our data set, the box plots of different users showing their engagement levels vs. the number of switches between applications in the last 5 minutes are depicted in Figure 3. Each color shows a different participant whose data points are among the highest in our data set. As it is possible to observe in the figure, individual means might change from participant to participant and there is an inverse relationship between engagement levels and the number of switches (also later confirmed by statistical tests). Those differences should be considered for modeling; hence, we preferred the models that take into consideration the individual inherent differences such as *rmcorr* and GLMM.

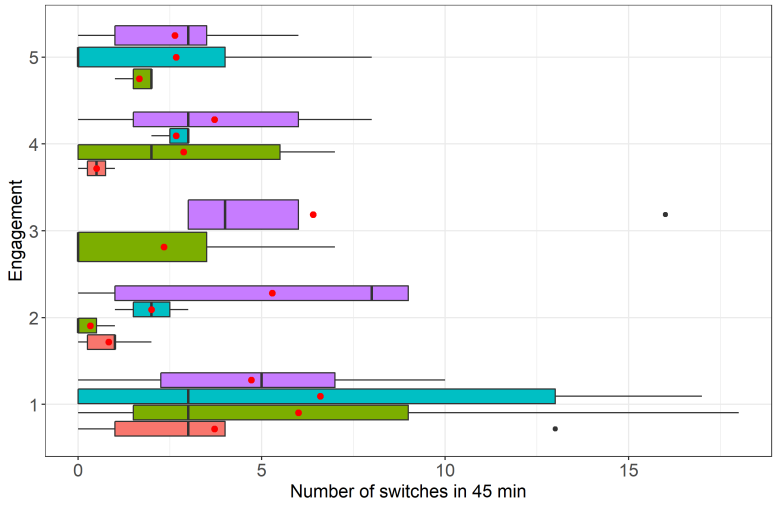

Fig. 3. Box plots of four different users showing their engagement levels vs. number of changes between applications in the latest 5 min. Each box plot with a color corresponds to a user. For example, the participant illustrated with orange color appears to have higher number of application switches when the engagement score is low but the number of application switches decreases when the engagement score increases. Each participant's application usage behavior is different (having different means, which is shown on the plot as red dots).

We performed GLMM analyses with the R package *MCMCglmm* [26]. *MCMCglmm* stochastically calculates the posterior distribution at a particular set of parameters. As far as the MCMC chains were concerned, convergence and consistency were satisfied with the default values for the parameters with a burn-in value of 3000, a thinning interval of 10 and the number of MCMC iterations of 13000. We used weakly informative priors for binary and ordinal response variables. Finally, Gelman-Rubin diagnostics were approximately one, and autocorrelation plots were stationary, which means that autocorrelation between consecutive samples in the chain is low enough for the convergence [26].

One of the aims of our study is to understand whether attentional states and engagement/challenge levels of knowledge workers can be predicted with mobile application usage. For this reason, we collected personal data from our participants. Mobile application usage among individuals may be quite different. For instance, one may spend most of his/her time using mobile phone, whereas another uses for only specific purposes. In such

situations, previous studies [45, 74, 81] stated that general models built upon multiple users' data might not be applied to study individual behavior. Hence, individual models (i.e., trained on solely user's data) can give more accurate results. As a disadvantage, individual models may suffer from the cold-start problem, which means a lack of sufficient individual data for training in the beginning [79, 81]. In order to cover this issue, we selected GLMM which incorporates both *fixed* and *random effects* together, which enables us to model both general and individual characteristics of the data.

In order to compare the results obtained from GLMMs, we used a random forest classifier, which is an ensemble learning method [8]. For training random forest models, we used both general (with all participants) and individual data (i.e., user-specific models). We generated an individual model for each participant by fitting random forest classifier on each participant's data.

We applied the same method for building models for all the classifiers. In particular, we used repeated random sub-sampling validation (i.e. repeated hold-out) [33]. We divided the data set into training and test sets with the proportions of 30-70, 40-60, 50-50, 60-40, and 70-30, and repeated each 20 times. When dividing data into training and test sets, we used stratified sampling, which allows for balancing class proportions in each set. We report the accuracy of the models as the performance metric. The estimated accuracy is obtained by averaging 20 runs. We also report the standard deviation of the accuracy values obtained from the runs. The only difference for the individual random forest classifier is that the data set consisted of one user's responses.

For all random forest classifiers (general and individual); as the parameter of the number of trees, several values (50, 100, 150, 250, 500, 750, 1000) were attempted in the models, which were cross-validated on a data set reserved at the beginning of the experiment and not used later to form test sets. The optimized parameters were 750 for the general attentional states model, 50 for the individual attentional states model, 100 for the general engagement levels model, 200 for the general challenge levels model, and 500 for the individual engagement and challenge levels model. We also calculated the baseline performance with the majority classifier that always predicts the class with the highest number of data points in the training set.

4 COLLECTED DATA AND FEATURE EXTRACTION

In this section, we will provide the descriptive statistics of the collected data. Besides, we will describe the features used in our analyses.

4.1 Application Usage

An application usage session is defined as the time spent between the screen on and off [5, 42]. Based on this definition, we extracted the application usage sessions of each participant. We recorded each session's start-end time, duration and the inter-event time information. Inter-event time refers to the duration between two consecutive usage sessions, i.e., the time difference between the previous session's end time and the current session's start time. We merged the sessions where inter-event time is less than 5 seconds as suggested in previous studies [5, 42]. We investigated usage sessions with a duration of 5, 10, 15, 30, 45 and 60 minutes before each message delivery from the core participants from whom we could obtain application package names.

In Table 1, the descriptive statistics (mean, median and standard deviation) for the application usage in the last 5, 10, 15, 30, 45 and 60 minutes before ESM messages were responded are shown. In the 5-minute case, more than half of the data set consisted of zeros, which means that more than half of the ESM messages were responded when users did not use their mobile phone in the last 5 minutes. As the duration increases, the number of sessions with no application usage decreases (39.47% for the 10-minute case, 29.43% for the 15-minute case, 20.10% for the 30-minute case, 12.68% for the 45-minute case, and 9.22% for the 60-minute case).

The usage of applications per category was also considered. However, such data could be obtained only from 14 number of participants. The application categories were matched using the classification present in the Google

Table 1. The descriptive statistics of the application usage (in seconds) in the last 5, 10, 15, 30, 45 and 60-minutes before ESM messages were responded. $N = 528$ number of sessions were recorded from 19 participants in total (those who had a response rate of 25% or higher in ESM questions).

Time Window	Mean	Std. Dev	Median	Number of sessions with no application usage	Percentage of sessions with no application usage
5 minutes	20.91	35.42	2.00	250	47.34%
10 minutes	31.37	49.93	11.00	200	37.88%
15 minutes	43.04	61.70	21.50	157	29.73%
30 minutes	395.40	508.58	169.50	29	5.49%
45 minutes	596.49	730.86	297.00	22	4.17%
60 minutes	791.29	934.63	448.50	13	2.46%

Play Store on December 2018. The category usage in the last 5, 10, 15, 30, 45 and 60 minutes prior to ESM messages was calculated. The descriptive statistics (mean, median and standard deviation) of each category usage in the last 60 minutes is given in Table 2. We also reported the number of zeros (i.e. non-used categories) and their percentages. As can be seen from the table, the application usage data in several categories is sparse. The densest category (i.e. the category with the highest number of data points other than zero) is communication; 46.17% of the sessions have communication applications usage greater than zero. We are aware that category names obtained from Google Play Store might be limited, however, as can be seen, our data set consisted of usage in a few categories. For this reason, we focused on *communication* and *social* categories in the analyses as they are. The reason of the sparseness in application categories can be attributed to the office workers' attitudes during office hours. Participants may use these applications for a significant amount of time outside office hours but they appear to use them rarely during working hours.

The features regarding the application usage are given below. Each feature is calculated for each participant in the last 5, 10, 15, 30, 45 or 60 minutes before each response to ESM delivery separately:

- *Application usage*: refers to the total duration of application usage (in minutes). For example, AU_{10} refers to the last 10-minute usage.
- *Number of applications*: indicates the number of unique applications used.
- *Number of switches*: is the total number of transitions between mobile applications.
- *Mean application usage*: corresponds to the average duration of application usage (in minutes).
- *Communication category usage*: denotes the duration of application usage in the communication category.
- *Social category usage*: denotes the duration of application usage in the social category.
- *Facebook usage*: is the duration of usage in Facebook application. We selected Facebook from the social category since Facebook usage was higher than half of the social category usage (57.40%).
- *WhatsApp usage*: shows the duration of usage in WhatsApp application. We selected WhatsApp from the communication category since WhatsApp usage was nearly equal to half of the social category usage (47.79%).
- *Messaging applications usage*: indicates the duration of usage in messaging applications (namely WhatsApp, Facebook messenger and SMS usages) belong to the communication category. The usage of messaging applications was equal to 58.54% of the communication category usage.

All the features mentioned above were calculated based on ESM timing, hence resulting in several data points in time per user. We also calculated the cumulative application usage in hours per user during work hours. A total

Table 2. The descriptive statistics of the application category usage (in seconds) in the last 60-min before ESM messages were responded. $N = 418$ number of sessions were recorded from 14 participants in total (those whose application package names could be obtained and had a 25% or higher response rate in ESM questions).

Category Name	Mean usage	Std. Dev.	Median	Number of sessions with no application usage	Percentage of sessions with no application usage
Communication	186.57	273.21	78.5	67	46.17%
Social	69.08	158.44	0	248	59.33%
Tools	14.04	93.77	0	324	77.51%
Productivity	3.28	16.43	0	375	89.71%
Finance	8.54	45.89	0	388	92.82%
Photography	7.11	47.56	0	388	92.82%
personalization	4.14	28.06	0	391	93.54%
Lifestyle	82.46	407.37	0	392	93.78%
Game	23.43	121.45	0	393	94.02%
News and Magazines	13.15	101.47	0	395	94.50%
Food and Drink	3.53	28.40	0	400	95.69%
Music and Audio	2.57	18.16	0	400	95.69%
Travel and Local	4.27	58.50	0	401	95.93%
Video Players and Editors	6.04	70.71	0	403	96.41%
Business	1.56	14.14	0	408	97.61%
Weather	0.41	3.81	0	409	97.85%
Shopping	0.88	11.11	0	410	98.09%
Entertainment	7.63	73.21	0	411	98.33%
Books and Reference	1.58	31.08	0	414	99.04%
Sports	1.04	18.13	0	414	99.04%

of 24 users among all 31 participants were considered since only these participants gave access to our application to collect their mobile application usage details in the background. Note that only 14 had a response rate of 25% or higher in ESM questions. The descriptive statistics of the aggregated mobile application usage parameters are given in Table 3.

- *Total Application Usage*: shows the total duration of application usage.
- *Total Communication Category Usage*: denotes the total duration of application usage in the communication category.
- *Total Social Category Usage*: indicates the total duration of application usage in the social category.
- *Total Facebook Usage*: denotes the total duration of Facebook usage.
- *Total WhatsApp Usage*: indicates the total duration of WhatsApp usage.

4.2 ESM Responses and Responsiveness Metrics

ESM responses consisted of engagement and challenge levels recorded with each ESM questionnaire, and attentional state derived from engagement and challenge scores for the same ESM questionnaire. The attentional states of the participants were classified using challenge and engagement levels of participants as in [41]. For example, if a participant recorded 1 as engagement response and 1 as challenge response, then the attentional

Table 3. The descriptive statistics of the constructs with continuous parameters used in the analyses. SD: Standard Deviation

Construct Name	Parameter Name	*N*	Min.	Max.	Mean±SD	Median
Responsiveness	Response Rate	31	.09	.94	.42±.27	.35
	Engagement					
	Median	31	1.00	5.00	2.82±1.01	3.00
	Interpolated Median	31	1.23	4.63	2.80±.89	2.81
	Entropy	31	.72	2.29	1.79±0.37	1.86
	Acquiescence	31	.00	.80	.36±.22	.37
	Disacquiescence	31	.00	.82	.45±.22	.46
	Middle Response Style	31	.00	.57	.18±.14	.18
	Positive Extreme Response Style	31	.00	.57	.19±.17	.15
	Negative Extreme Response Style	31	.00	.68	.27±.16	.28
	Challenge					
	Median	31	1.00	4.00	2.29±.79	2.00
	Interpolated Median	31	1.18	3.60	2.30±.67	2.38
	Entropy	31	.83	2.23	1.68±.38	1.69
	Acquiescence	31	.00	.56	.20±.16	.19
	Disacquiescence	31	.00	1.00	.58±.25	.52
	Middle Response Style	31	.00	.63	.23±.17	.14
	Positive Extreme Response Style	31	.00	.36	.06±.09	.00
	Negative Extreme Response Style	31	.00	.74	.33±.17	.33
Mobile application usage (in minutes)	Total application usage	31	1.55	65.14	15.16±12.88	13.44
	Total social category usage	24	.01	4.07	1.69±1.35	2.09
	Total communication category usage	24	.36	20.33	7.24±5.34	6.29
	Total Facebook usage	24	.00	4.20	.90±1.07	.75
	Total WhatsApp usage	24	.00	6.00	1.74±1.43	1.59
Office related factors	Number of colleagues in office	31	1.00	50.00	10.13±14.51	3.00
	Subjective Norm	31	2.00	9.00	6.45±1.97	7.00

state of the participant at that moment was labeled as "bored". The details of the classification are given in Section 2.5. The descriptive statistics of engagement and challenge responses obtained from the ESM questions are given in Table 4. Only "focused" and "bored" states were included in the analyses since the number of data points in "rote" and "frustrated" states had a fewer number of data points. The average duration for responding ESM questions is 3.17 minutes with a standard deviation of 2.47[1].

We explored the responsiveness of participants in detail. The indicators reported below were calculated based on the engagement and challenge responses from the ESM questions, then, aggregated. The descriptive statistics of the indicators are presented in Table 3.

- *Response Rate*: This is calculated through dividing the number of each participant's responses that were successfully recorded by the total number of ESM messages sent to that participant. We assumed that the response is missing when the phone is off during an expected prompt.

[1]In our study, the duration was calculated based on the difference between the survey submission time and the ESM message delivery time.

Table 4. The descriptive statistics of the engagement/challenge responses and attentional states obtained from the ESM questionnaires.

Parameter Name	Values	Frequency	Percentage
Engagement	1	132	31.58%
	2	73	17.46%
	3	52	12.44%
	4	66	15.79%
	5	95	22.78%
Challenge	1	146	34.93%
	2	89	21.29%
	3	81	19.38%
	4	60	14.35%
	5	42	10.05%
Attentional states	Bored	197	47.13%
	Focused	175	41.87%
	Rote	38	9.09%
	Frustrated	8	1.91%

- *Median of Engagement and Median of Challenge*: They refer to the median value of the responses of each participant on the ESM questions related to engagement and challenge respectively.
- *Interpolated Median of Engagement and Challenge*: Medians may suffer from ignoring the weights linked to the distributions of responses above or below the median. The interpolated medians take into account the number of data points, which are strictly below or above the median. Hence, in our study, we calculated the interpolated medians of engagement and challenge responses respectively.
- *Entropy of Engagement and Entropy of Challenge*: These refer to the Shannon entropy of the responses of each participant on the ESM questions related to engagement and challenge respectively. The formula of the entropy is given in Equation 1. In the formula, p_i refers to the proportion of item i (where $i = 1, .., 5$ since engagement and challenge levels were measured with 5-level Likert scale). A lower entropy implies higher homogeneity of responses, whereas a higher entropy indicates higher heterogeneity of responses. For example, if a participant's responses belong to only one category (e.g., the user selected the item 3 for all engagement questions), then his/her entropy of engagement will be zero, which indicates pure homogeneity. On the other hand, entropy gets higher values as the responses fall into different categories.

$$Entropy = \sum_{i=1}^{5} -p_i log_2 p_i \tag{1}$$

- *Acquiescence of Engagement and Challenge*: They refer to the tendency to be highly engaged or challenged with work (i.e. responding 4 or 5 to the engagement/challenge level questions). It is calculated by dividing the number of responses recorded as 4 or 5 by the total number of responses.
- *Disacquiescence of Engagement and Challenge*: They refer to the opposite of the acquiescence, which means the tendency to be low engaged or challenged with work (i.e. responding 1 or 2 to the engagement/challenge level questions). It is calculated by dividing the number of responses recorded as 1 or 2 by the total number of responses.

Table 5. The descriptive statistics of the constructs with the indicators (ordinal parameters) used in the analyses.

Construct Name	Parameter Name	Values	Frequency	Percentage
Awareness about rest breaks	Taking regular rest breaks	1	3	9.7%
		2	18	58.1%
		3	10	32.3%
	Doing office exercises	1	20	64.5%
		2	6	19.4%
		3	5	16.1%
Musculoskeletal discomfort	Feeling pain/numbness	1	8	25.8%
		2	5	16.1%
		3	18	58.1%

- *Middle Response Style of Engagement and Challenge*: They refer to the proportion of the responses that received middle (3) response.
- *Positive Extreme Response Style of Engagement and Challenge*: Positive extreme responses imply the responses with the category equal to 5. Hence, the positive extreme response style indicates the proportion of the responses that received positive extreme responses among all responses.
- *Negative Extreme Response Style of Engagement and Challenge*: Negative extreme responses imply the responses with the category equal to 1. Hence, the negative extreme response style indicates the proportion of the responses that received negative extreme responses among all responses.

4.3 Survey Data Set

As discussed above, we collected information regarding office-related factors, musculoskeletal discomfort, and awareness about sitting behavior with the pre-experiment survey. The features and their descriptive statistics obtained from the survey results are given in Table 3 and in Table 5. The total number of participants in the data set is 31. The following features were calculated for each user:

- *Awareness about rest breaks*: This was measured with the degree of participants taking rest breaks and doing office exercises.
- *Musculoskeletal discomfort*: This indicates the degree of feeling pain/numbness when working in front of computers in the office.
- *Office-related Factors*: They refer to the number of colleagues in each participant's office and subjective norm (SN), which identifies the degree of participants affected by their colleagues for taking rest breaks. The number of colleagues in the same office might affect the responsiveness since the distraction level caused by co-workers might vary with different number of people. SN identifies the influence level caused by co-workers, hence, it might affect the responsiveness. Among our *core participants (N=19)*, three of them (15.79%) had no colleagues in their offices, seven of them (36.84%) had up to five colleagues, and the remaining nine of them (47.37%) had a minimum of seven colleagues and a maximum of 50 colleagues. Similarly, seven of the participants (36.84%) had a SN scores between two and six, four of them (21.05%) had SN score of seven, and eight of them (42.11%) had SN scores of eight and nine out of ten.

5 DATA ANALYSIS AND RESULTS

In this section we present the analysis of the data set collected through the studies discussed above. In the first subsection, the relations among the constructs in the survey data set (musculoskeletal discomfort, awareness,

and office-related factors), mobile application usage indicators, and responsiveness metrics are investigated using Kendall's Tau correlation based on users' aggregated values. In the second subsection, the relation between mobile application usage parameters and in-situ engagement/challenge levels is investigated using repeated-measures correlation analysis. Finally, in the third section, a hierarchical model for predicting individuals' engagement and challenge levels is proposed, and its performance is compared with general and individual random forest models.

5.1 Relational Analysis between the Constructs in the Survey Data Set

In the research question RQ1, we investigated *how musculoskeletal discomfort of office workers is related to their responsiveness to the break-reminder notifications*. We assessed the musculoskeletal discomfort as the degree of feeling pain/numbness while working on computers. The results show that the degree of feeling numbness and pain during work is in a weak relationship with the entropy of engagement ($\tau = .25, p = .09, N = 31$). The engagement responses of the participants who felt a higher amount of musculoskeletal discomfort were more heterogeneous than those who felt less musculoskeletal discomfort. In other words, the participants who suffered more from musculoskeletal discomfort responded with a higher number of categories as their engagement levels.

In RQ2, we assessed *how awareness of office workers is related to their responsiveness to the break-reminder notifications*. The awareness was measured with the amount of regular rest breaks and the frequency of doing office exercises. Based on the results, taking regular rest breaks appears to be positively related to the negative extreme response style of engagement ($\tau = .25, p = .09, N = 31$). It can be inferred that the participants who take rest breaks more frequently selected the response item "I am not engaged at all" higher number of times than the ones who take rest breaks less frequently. Even though we did not hypothesize it, an interesting result has been found: taking rest breaks is negatively related to feeling pain/numbness ($\tau = -.50, p = .003, N = 31$). It means that the participants who have rest breaks more frequently felt less musculoskeletal discomfort while working than the ones who take rest breaks less frequently.

We studied *how office-related factors are associated with office workers' responsiveness to the break-reminder notifications* in RQ3. Office-related factors were measured by the number of colleagues and subjective norm. The results showed that the number of colleagues is in a weak positive relationship with the entropy of challenge ($\tau = .23, p = .09, N = 31$). It means that the participants who share their offices with a higher number of colleagues responded to the challenge questions more heterogeneously (i.e., responded with a higher number of item categories) than the ones who share their offices with a lower number of colleagues. At the same time, SN is significantly related to the number of colleagues in office ($\tau = .35, p = .01, N = 31$). Again, the participants who share their offices with a higher number of colleagues stated that they are affected more from their colleagues regarding taking rest breaks than the ones who share their offices with a lower number of colleagues.

We now consider research question RQ4a regarding *the relation between application usage and responsiveness to the break reminder notifications*. The correlation coefficients between the application usage parameters of total application usage, total social category usage, total communication category usage, total Facebook usage, and total WhatsApp usage and responsiveness parameters are depicted in Table 6. The results can be summarized as follows:

- *Total Application Usage*: This is positively related to the median of challenge levels ($\tau = .32, p = .02, N = 31$). Total application usage is also in a positive relationship with the negative extreme response style of challenge ($\tau = .24, p = .02, N = 31$). It means that the participants who used mobile applications for a longer amount of time gave the response "not challenged at all" more than those who used mobile applications for a shorter amount of time.
- *Total Social Category Usage*: This is positively related to the negative extreme response style of engagement ($\tau = .27, p = .06, N = 24$) and challenge ($\tau = .27, p = .07, N = 24$). Those results can be inferred as the participants with a higher amount of application usage in social category tend to select the item "not

Table 6. Kendall's Tau correlation coefficients for the responsiveness variables and application usage parameters ($N = 31$ for Total App, $N = 24$ for communication, social, Facebook and WhatsApp)

		Total App	Communication	Social	Facebook	WhatsApp
Response Rate	τ	.19	-.19	.12	.17	.22
	p	.13	.20	.40	.24	.14
Median of Engagement	τ	.13	.07	.27	.15	.33
	p	.33	.66	.09	.34	.09
Interpolated Median of Engagement	τ	-.10	.09	.04	-.16	-.15
	p	.43	.52	.80	.29	.32
Entropy of Engagement	τ	.18	-.11	.03	.17	.08
	p	.16	.44	.82	.25	.57
Acquiescence of Engagement	τ	-.13	.19	.11	-.14	.01
	p	.31	.20	.46	.36	.96
Disacquiescence of Engagement	τ	.08	.03	.10	.26	.16
	p	.55	.82	.50	.07	.26
Middle Response Style of Engagement	τ	.01	-.09	-.21	-.13	-.14
	p	.95	.57	.15	.37	.33
Positive Extreme Response Style of Engagement	τ	-.02	.11	.28	.02	.03
	p	.86	.45	.06	.90	.84
Negative Extreme Response Style of Engagement	τ	.20	.06	.27	.37	.42
	p	.11	.67	.06	.01	.004
Median of Challenge	τ	.32	.01	.08	.07	.32
	p	.02	.94	.62	.68	.04
Interpolated Median of Challenge	τ	-.07	-.24	-.04	-.28	-.14
	p	.57	.11	.78	.06	.36
Entropy of Challenge	τ	.06	-.24	-.23	-.11	.02
	p	.62	.10	.12	.44	.88
Acquiescence of Challenge	τ	-.11	.01	.12	-.10	.01
	p	.39	.96	.40	.50	.96
Disacquiescence of Challenge	τ	.05	.26	.06	.28	.19
	p	.72	.07	.69	.06	.20
Middle Response Style of Challenge	τ	-.08	-.32	-.10	-.12	-.21
	p	.53	.03	.52	.41	.16
Positive Extreme Response Style of Challenge	τ	.09	.00	.00	-.17	.02
	p	.51	1.00	1.00	.27	.92
Negative Extreme Response Style of Challenge	τ	.24	.34	.27	.36	.31
	p	.06	.02	.07	.02	.04

challenged/engaged at all" more than the participants with a lower amount of application usage in the social category.

- *Total Communication Category Usage*: This is positively associated with the disacquiescence of challenge ($\tau = .26, p = .07, N = 24$) and negative extreme response style of challenge ($\tau = .34, p = .02, N = 24$). It can be stated that the participants who used communication applications more, recorded lower challenge scores than those who used communication applications less.
- *Total Facebook Usage*: This is in a negative relationship with the interpolated median of challenge ($\tau = -.28, p = .06, N = 24$), which means that the participants who used Facebook more tend to respond to the challenge questions with lower scores than the participants who used Facebook less. Similarly, total Facebook usage is positively related to the disacquiescence of engagement ($\tau = .26, p = .07, N = 24$), to the disacquiescence of challenge ($\tau = .28, p = .06, N = 24$), to the negative extreme response style of engagement ($\tau = .37, p = .01, N = 24$), and to the negative extreme response style of challenge ($\tau = .36, p = .02, N = 24$). The results infer that the participants who used Facebook a longer amount of time recorded lower engagement/challenge scores than the participants who used Facebook a lower amount of time.
- *Total WhatsApp Usage*: This is in positive relationship with the median of engagement ($\tau = .33, p = .03, N = 24$) and challenge ($\tau = .32, p = .04, N = 24$). Similar to Facebook usage, WhatsApp usage is positively related to the negative response styles of engagement ($\tau = .42, p = .004, N = 24$) and challenge ($\tau = .31, p = .04, N = 24$). The results infer that the participants who used WhatsApp a longer amount of time during work hours responded with "not challenged/engaged at all" item more than the ones who used WhatsApp a lower amount of time.

Following the study and considering the relations found in the analyses discussed above, our initial research framework in Fig. 1 has been revised as in Fig. 4. The dashed lines show the hypotheses partially supported, the black lines show the hypotheses supported, and the red lines show the relations not hypothesized at first, but found significant.

5.2 Repeated-Measures Correlation Analysis

We now present the results of our investigation regarding *which application usage metrics are related to in-situ engagement/challenge levels of office workers (RQ4b)*. We conducted a repeated-measures correlation analysis due to the reasons mentioned in Section 3.7.

We calculated the repeated-measures correlation coefficients on the ESM response data set between the in-situ engagement/challenge levels and the application usage constructs: *the duration of application usage, the number of applications, the number of switches between applications, mean application usage, the duration of social applications usage, the duration of communication applications usage, the duration of messaging applications usage, the duration of WhatsApp usage* and *the duration of Facebook usage* in order to measure the relation between in-situ engagement/challenge levels and mobile application usage. Repeated-measures correlation (*rmcorr*) results between the application usage constructs and the in-situ engagement levels are given in Table 7. Similarly, *rmcorr* results related to the application usage constructs and the in-situ challenge levels are given in Table 8. Note that *rmcorr* could not be applied as the attentional state is a binary variable.

The results showed that the number of switches between applications and the number of applications are the most related features to challenge levels in decreasing order. In particular, the window size between 30 and 60 minutes appears to be the most determinant one in common. Although the window size between 10 to 30 appears to be statistically significant, the magnitude of the relationships appears to be lower than those having a longer time window. All the relations are in a negative direction. For example, in the case of the 60-minute window size, as the number of switches between application increases, engagement ($r_{rm} = -.11, p = .02$) and challenge levels

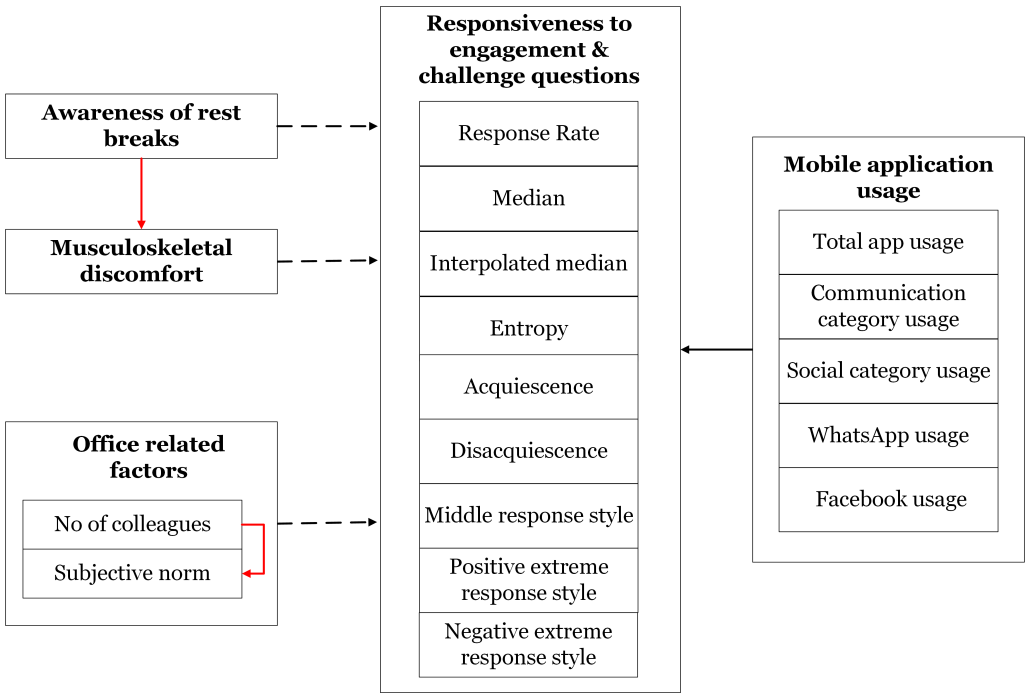

Fig. 4. Revised research framework based on the analysis results.

($r_{rm} = -.17, p < .001$) decrease, or vice versa. Similarly, the number of applications in 60-minute window size has a negative correlation with engagement ($r_{rm} = -.09, p = .08$) and challenge levels ($r_{rm} = -.14, p = .006$). In other words, it means that higher application usage refers to a lower level of engagement/challenge. In terms of the application categories, the use of communication applications is significantly related to the challenge levels in 60-minute window size ($r_{rm} = -0.10, p = .04$) but not to the social category applications.

The total application usage, the number of applications, the number of switches between applications, and the communication applications are the most related features to the engagement levels in decreasing order. Similarly, all the relations are also in a negative direction. However, this time, the window size between 5 and 10 appears to be the most determinant one in common apart from the fact that the number of switches between applications appears to be also significant in longer time windows. As shown previously in Figure 3, each participant had different means in terms of the number of switches with respect to engagement levels. The *rmcorr* results show that individual-level relation from a statistical point of view.

The use of social, Facebook, WhatsApp, and messaging applications were not found as significantly related to the engagement and challenge levels.

5.3 Modeling In-Situ Attentional States, Engagement and Challenge Levels: GLMM

In this section, we developed *a model for inferring attentional states, engagement/challenge levels of office workers using application usage metrics by considering the cold start problem, the variety in the number and characteristics of the responses and repeated-measurement nature of the data (RQ4c).* We also investigated how the model is comparable to individual and general models, which use the random forest method.

Table 7. Repeated-measures correlation results between the application usage variables and *engagement* levels. Rows show the application variables, whereas columns show the time window of the variables ($N = 418$). For example, the *rmcorr* coefficient (r_{rm}) between the engagement levels and the application usage in the last 5 minutes before ESM messages arrived is equal to -.14.

		5-min	10-min	15-min	30-min	45-min	60-min
App Use	r_{rm}	-.14	-.12	-.08	-.05	-.04	-.04
	p	.005	.02	.12	.28	.37	.45
No of Apps	r_{rm}	-.11	-.13	-.08	-.11	-.08	-.09
	p	.02	.009	.11	.03	.11	.08
No of Switch	r_{rm}	-.11	-.11	-.07	-.09	-.12	-.11
	p	.03	.02	.17	.08	.02	.02
Mean App Use	r_{rm}	-.06	-.07	-.05	-.01	.01	.04
	p	.24	.18	.28	.80	.91	.37
Social	r_{rm}	-.04	-.05	-.01	.00	.01	.03
	p	.46	.30	.81	.97	.79	.58
Communication	r_{rm}	-.10	-.11	-.05	-.05	-.07	-.08
	p	.05	.03	.28	.35	.13	.10
Messaging	r_{rm}	-.07	-.09	-.06	-.07	-.07	-.06
	p	.14	.08	.25	.15	.14	.26
Facebook	r_{rm}	-.08	-.05	-.02	.06	.07	.04
	p	.11	.46	.75	.21	.14	.44
WhatsApp	r_{rm}	-.03	-.03	-.01	-.04	-.05	-.04
	p	.54	.53	.85	.38	.35	.48

5.3.1 GLMM Results. Feature and model selection with GLMM for predicting attentional states, engagement and challenge levels is presented in Appendix B. We iteratively selected the model, which gives consistently lower DIC among different sub-samples of the original data set with multiple runs. The details of the model selection is provided in the appendix as to keep the focus of the paper intact. The results obtained from the models selected are discussed in this section. In summary, we included several combinations of the application usage parameters for modeling attentional states, engagement and challenge levels. The pairs of NOS_{60}, AU_5, NOS_{45}, and NOA_{10} were included. Finally, NOS_{45} and AU_5 resulted the lowest DIC for the attentional states, engagement and challenge level models.

In GLMM, we directly modeled the levels of engagement and challenge instead of using a binary conversion of them, since an ESM response does not necessarily imply that the person is engaged/challenged or not.

Table 9 (top) shows the posterior distributions of each parameter with the posterior means and the 95% credible intervals (2.5 and 97.5 percentiles of the posterior distribution) of Model 3 for attentional states. The number of switches between applications in the last 45 minutes has been found statistically significant ($p = .02$) on predicting attentional states. The negative relation means that as the number of switches between applications in the last 45 minutes before ESM messages increases, users are more likely to be "bored"; or as the number of switches between applications decreases, users are more likely to be "focused". Similarly, the application usage in the last 5 minutes before ESM messages has been found negatively related to the attentional states. As the duration of application usage in the last 5 minutes increases, users are more likely to be in the "bored" state, or vice versa. Based on the magnitudes, it can be said that the effect of application usage in the last 5 minutes is higher than the number of switches between applications in the last 45 minutes.

Table 8. Repeated-measures correlation results between the application usage variables and *challenge* levels, where the columns show the time window of the variables ($N = 418$). For example, *rmcorr* coefficient (r_{rm}) between the challenge levels and the application usage in the last 5 minutes before ESM messages were responded is equal to -.11.

		5-min	10-min	15-min	30-min	45-min	60-min
App Use	r_{rm}	-.11	-.09	-.10	-.09	-.10	-.09
	p	.03	.08	.06	.06	.04	.08
No of Apps	r_{rm}	-.09	-.12	-.11	-.17	-.13	-.14
	p	.07	.02	.03	<.001	.008	.006
No of Switch	r_{rm}	-.10	-.11	-.10	-.17	-.18	-.17
	p	.05	.03	.04	<.001	<.001	<.001
Mean App Use	r_{rm}	-.02	-.04	-.05	-.04	-.03	.00
	p	.63	.38	.28	.43	.55	.92
Social	r_{rm}	-.01	-.04	-.04	-.05	-.04	-.03
	p	.87	.45	.38	.29	.37	.51
Communication	r_{rm}	-.09	-.10	-.08	-.07	-.11	-.10
	p	.06	.04	.13	.18	.03	.04
Messaging	r_{rm}	-.05	-.07	-.05	-.07	-.07	-.05
	p	.28	.17	.27	.18	.17	.29
Facebook	r_{rm}	-.05	-.03	-.04	-.02	-.02	-.06
	p	.28	.48	.45	.66	.70	.25
WhatsApp	r_{rm}	-.04	-.04	-.03	-.05	-.05	-.04
	p	.44	.37	.48	.28	.28	.44

Table 9 (middle) shows the posterior distributions of each parameter with the posterior means and the 95% credible intervals (2.5 and 97.5 percentiles of the posterior distribution) of Model 1 for engagement levels. Similar to the attentional states model, the number of switches between applications in the last 45 minutes and the duration of application usage in the last 5 minutes are negatively related to the engagement levels. As the number of switches between applications increases, the participants tend to be lowly engaged with their work or vice versa. Similarly, as the participants use a higher amount of mobile applications in the last 5 minutes, they are more likely to be engaged with their work in a lower degree or vice versa. As in the attentional states model, the magnitude of the application usage is higher than the number of switches between applications on the engagement levels.

Finally, Table 9 (bottom) shows the posterior distributions of each parameter with the posterior means and the 95% credible intervals (2.5 and 97.5 percentiles of the posterior distribution) of Model 3 for challenge levels of users. Again, the effect of the number of switches between applications parameter is significantly negative on challenge levels ($p = .008$). As the number of switches between applications in the last 45 minutes increases, the challenge levels of users decrease, or vice versa. Similarly, the application usage in the last 5 minutes is also in a significant negative relationship with the challenge levels ($p = .01$). The participants who use mobile applications for a longer amount of time in the last 5 minutes tend to be challenged with their work in a lower degree or vice versa.

5.3.2 Comparison of GLMM with General and Individual Models. As stated before, for comparison, we fit random forest models both in general (population) level and individual level. The variables used in the models were the same with the ones reported in the GLMM results: the number of switches between applications in the last 45

Table 9. Posterior means, 95% credible intervals and p values of parameters for Model 3 for attentional states (top), Model 1 for engagement levels (middle) and Model 3 for challenge levels (bottom).

Model 3 for Attentional States			
Parameters	Posterior mean	95% CI	p
(Intercept)	.23	(-.17,.60)	.21
No of switch (45-min)	-.07	(-.12,-.01)	.02
App use (5-min)	-.38	(-.73,-.10)	.01
Model 1 for Engagement Levels			
Parameters	Posterior mean	95% CI	p
(Intercept)	.92	(.68,1.11)	<.001
No of switch (45-min)	-.04	(-.07,-.01)	.008
App use (5-min)	-.21	(-.39,-.04)	.02
Model 3 for Challenge Levels			
Parameters	Posterior mean	95% CI	p
(Intercept)	.84	(.62,1.09)	<.001
No of switch (45-min)	-.06	(-.10,-.02)	.002
App use (5-min)	-.15	(-.31,.03)	.01

minutes and the application usage in the last 5 minutes before ESM messages. We built random forest models for each participant (in total 14 models), then reported the accuracy values by averaging them.

In Table 10, the mean accuracy values obtained from four different classifiers for predicting attentional states, engagement levels, and challenge levels are reported with their standard deviations. In the second column, the percentage of the training set is stated. We also performed 1-person-out cross-validation (CV) on the data set in order to re-validate the performance of the classifiers. The biases caused by individuals that were uniformly distributed in other splits were avoided by performing 1-person-out CV. Remember that attentional states were modeled as a binary response (i.e., as "focused" and "bored"), and engagement and challenge levels as an ordinal response (i.e., as 1-5).

As illustrated in the table, GLMM predicts engagement and challenge levels better than the general random forest model, individual random forest and baseline classifiers for all training percentages and in the 1-person-out CV. As expected, the accuracy obtained using 1-person-out CV is lower since the individual's own data are highly informative and we do not use them in 1-person-out CV. Only for predicting attentional states, individual random forest models are slightly better than GLMM. In order to compare the accuracy values of four classifiers, we conducted statistical tests on the accuracy values obtained from all runs (5 different training percentages x 20 runs = 100 accuracy values for each classifier). The Shapiro-Wilk Test shows that the accuracy values of the models did not distribute normally ($p < .001$). Hence, the Friedman Test was performed [16]. The results of the Friedman Test shows that the average accuracy values obtained from the four classifiers are significantly different for the prediction of attentional states ($\chi^2(3) = 169.07, p < .001, N = 100$), engagement levels ($\chi^2(3) = 206.06, p < .001, N = 100$), and challenge levels ($\chi^2(3) = 208.50, p < .001, N = 100$). Then, Wilcoxon signed rank tests were conducted for binary comparisons of the models as post-hoc tests. For the prediction of attentional states, the accuracy obtained from GLMM is significantly higher than the general random forest's accuracy ($Z = -8.26, p < .001, N = 100$), individual random forest's accuracy ($Z = -5.99, p < .001, N = 100$), and the baseline accuracy ($Z = -4.66, p < .001, N = 100$). Similarly, GLMM significantly outperforms the general

Table 10. Comparison of model accuracy values for predicting attentional states, engagement and challenge levels. The target variable is 2-leveled in the AS models, 5-leveled in Eng. and Chal. models. (AS: Attentional States, Eng.: Engagement, Chal.: Challenge)

Model	Training Percentage	Attentional States $N = 372$	Engagement $N = 418$	Challenge $N = 418$
GLMM	30%	53.53%±3.28%	29.35%±3.29%	33.23%±2.82%
	40%	54.02%±2.09%	31.97%±1.92%	34.17%±3.01%
	50%	53.22%±3.45%	32.03%±2.28%	36.10%±2.05%
	60%	55.51%±2.96%	31.62%±2.94%	36.31%±3.09%
	70%	54.19%±3.17%	34.12%±3.73%	35.97%±2.91%
	1-person-out CV	51.53%±8.40%	30.00%±8.75%	36.58%±8.12%
General RF *(No of trees = 750 for AS 100 for Eng 200 for Chal)*	30%	48.34%±2.31%	21.51%±2.73%	26.56%±3.21%
	40%	48.26%±2.75%	22.61%±2.94%	28.89%±3.11%
	50%	48.57%±2.91%	24.13%±2.40%	31.27%±2.48%
	60%	47.03%±2.95%	22.63%±1.40%	32.68%±2.69%
	70%	49.10%±2.89%	23.08%±3.91%	31.98%±3.60%
	1-person-out CV	43.02%±8.02%	20.73%±6.31%	24.51%±9.77%
Individual RF *(No of trees = 50 for AS 500 for Eng 500 for Chal)*	30%	50.45%±12.95%	23.96%±11.87%	27.07%±13.87%
	40%	51.52%±15.32%	24.51%±12.86%	28.81%±14.38%
	50%	49.73%±18.65%	24.17%±13.11%	30.57%±14.28%
	60%	52.18%±18.44%	24.54%±13.32%	29.43%±15.58%
	70%	49.91%±21.98%	26.40%±16.94%	29.88%±17.71%
Baseline	30%	50.41%±18.26%	32.49%±15.22%	32.95%±14.54%
	40%	51.26%±18.28%	30.12%±13.92%	31.57%±14.47%
	50%	52.67%±19.89%	31.14%±15.71%	30.63%±14.71%
	60%	50.34%±19.67%	30.87%±18.40%	30.92%±16.61%
	70%	51.59%±22.74%	30.34%±21.60%	34.44%±18.02%
	1-person-out CV	50.48%±15.95%	29.85%±11.33%	30.24%±14.25%

random forest ($Z = -8.68, p < .001, N = 100$), and individual random forest ($Z = -8.59, p < .001, N = 100$) for predicting engagement levels. The difference between the GLMM accuracy values and baseline accuracy values is significantly different at the α=0.1-level ($Z = -1.87, p = .06, N = 100$). Finally, for the challenge levels models, GLMM gives significantly the most accurate results among general random forest ($Z = -8.68, p < .001, N = 100$), and individual random forest ($Z = -8.02, p < .001, N = 100$). However, the difference between GLMM and the baseline classifier is not found statistically significant ($Z = -.52, p = .60, N = 100$), which means that baseline classifier and GLMM lead to similar accurate results for predicting challenge levels.

6 DISCUSSION

In this study, we have investigated several research questions concerning the responsiveness of individuals to break-reminder notifications, including engagement/challenge level and relationships between responsiveness and several personal and social factors such as musculoskeletal discomfort, office-related factors, and mobile

application usage. The results of the study show two groups of variables (mobile application usage, and office-related factors) have significant relationships with the responsiveness of knowledge workers regarding their engagement/challenge levels. The remaining two groups of variables (awareness and musculoskeletal discomfort) have a less significant one. In the following, we discuss these results in detail.

6.1 Musculoskeletal Discomfort and Awareness

In RQ1, we investigated the relation between the musculoskeletal discomfort of office workers and their responsiveness to break-reminder notifications. Musculoskeletal discomfort was found positively related to the entropy of engagement responses. The engagement responses of the participants who felt a higher level of musculoskeletal discomfort were more heterogeneous than the ones who experienced a lower level of musculoskeletal discomfort. Surely more evidence is needed, but a high degree of musculoskeletal discomfort could be the reason for not focusing on work, hence the responses varied more.

In RQ2, we investigated the relation between the awareness of office workers about rest breaks and their responsiveness to break-reminder notifications. The participants, who have higher awareness scores related to having rest breaks, responded with the item "I am not engaged at all" more frequently than those than those who have less awareness. The result may be due to the fact that since those participants are more aware of the importance of having rest breaks, they may actually give regular rest breaks more frequently than the ones who are not aware that much. For this reason, our ESM messages may have arrived at the breaks, so that they may have responded as they were not engaged with their work at that moment. We did not find a significant relationship with other responsiveness variables. As van Kenhove *et al.* [72] stated, response behavior can be related to topic involvement, and involvement of individuals must exceed a critical level for the decision of participation. Similarly, in our study, our participants' awareness levels or musculoskeletal discomfort levels may not have exceeded a critical level for themselves, so their responsiveness metrics was not found to be related to their awareness or musculoskeletal discomfort levels.

Although we did not hypothesize the relation between musculoskeletal discomfort and awareness, our results revealed that there is a negative relation between the two. The participants who take rest breaks more regularly stated that they do not feel musculoskeletal discomfort while working as much as the ones who do not take regular rest breaks. On the other hand, the participants who do not take regular rest breaks are the ones who feel musculoskeletal discomfort more. The negative relation between the degree of feeling musculoskeletal discomfort during work and the amount of rest breaks clearly showed that participants who take regular rest breaks from their work reported suffering less from musculoskeletal discomfort. This result on the importance of the rest breaks is in line with the previous studies [6, 22, 50].

6.2 Office-related Factors

In RQ3, we investigated how office-related factors are associated with office workers' responsiveness to break-reminder notifications. The number of colleagues and the subjective norm scores correspond to the office related factors in the study. Based on the results, the number of colleagues was found in a weak positive association with the entropy of challenge responses. Previous studies mostly focused on office types and their effects on distractions [51, 65]. As those studies stated, employees in shared offices stated that they are distracted more easily than those in private offices. The number of colleagues that we included in our study might be a strong indicator of office type. Although we did not ask participants directly their office type, we may make an inference about their office types based on the number of colleagues. For example, in the pre-experiment questionnaire, the maximum and minimum number of people working in the office were reported as 1 and 50, i.e., values that imply a private office and a shared office, respectively. This also shows the variety of office types included in the experiment. We also collected their working locations and they were different.

Based on the results, we might say that employees in the shared offices reported different challenge levels. It may be due to the distraction they perceive in their office environments. As suggested in previous studies [51], distraction in shared offices is higher, which may lead to different challenge levels of employees in such conditions.

Even though we did not hypothesize it, the positive relationship between the number of colleagues and the subjective norm scores was found statistically significant. The participants who shared their offices with a higher number of colleagues stated that they are affected more from their colleagues for taking rest breaks compared to those who share a lower number of colleagues. The result is in line with the previous studies, still, to the best of our knowledge, the relation found in the study has not been stated in any study before. Male employees in a university work environment stated that participating in a social group is a motivating factor [24]. Our result contributes to it by showing a significant direct relationship between the size of that social group and the degree of the social norm. The bigger the social group is, the higher the people are affected by that social group.

6.3 Application Usage

We investigated how mobile application usage of office workers are related to their responsiveness to break-reminder notifications in RQ4a and which application usage metrics are related to in-situ engagement/challenge levels of office workers in RQ4b. The results show that there are significant relations between the responsiveness of participants and application usage parameters. The participants who used mobile applications a higher amount of time during work hours tend to respond to the engagement/challenge questions with lower levels than the ones who used mobile applications lower amount of time. When the results are investigated in detail, it can be seen that disacquiescence or negative extreme response styles of engagement/challenge are in positive relationships with almost all application usage parameters. So, an increase in application usage during work hours may be a signal for a low level of engagement/challenge with work. The studies [40, 41] showed that Facebook usage (on web browsers) is significantly effective on engagement/challenge levels. We contribute to this result by considering Facebook usage as mobile application usage.

Our results show that there is a significant positive relation between total application usage and the median of challenge levels. However, no significant relation was found between interpolated median of engagement/challenge levels and total application usage. Although the relationships are not significant, the direction of the relations was negative. Hence, both results are not consistent. Recall that using median especially for Likert scales has a limitation, therefore we used interpolated median in our study. Besides, those aggregated measures are limited on analyzing the relations at individual level. Every user has a different level of application usage, therefore it is more meaningful to analyze those relationships at individual level. An increase or a decrease at individual level could be seen directly with repeated-measures.

Repeated-measures correlation results show a significant negative relation between *in-situ* engagement/challenge levels and total application usage. As total usage increases, participants' engagement and challenge levels decrease or vice versa. Similar relation occurs with communication applications usage. As participants' usage of communication applications increase, their engagement levels decrease or vice versa. In addition, the number of switches between applications had a negative relation with engagement levels. The more participants switch between applications, the less they are engaged with their work. Similarly, as they use a higher number of applications before 5 or 10 minutes prior to ESM prompts, it shows a decrease in their engagement and challenge levels. Our results are in line with the previous studies that investigated the relation between application usage and boredom [36, 43, 59, 61].

In our study, we investigated the relationship between engagement/challenge levels and application usage both in a short period of time (e.g. 5, 10 and 15 minutes) and in a longer period of time (e.g. 30, 45, 60 minutes). In the previous studies, different periods of time have been discussed. For example, in [42], time window was

set as 60 minutes, whereas in [60] it was set to 10 minutes, and finally in [61] it was 5 minutes. Also, note that those studies did not investigate the work engagement. Our study investigated the effect of different time windows of application usage. More specifically, only communication category, total application usage, number of applications, and number of switches between applications in shorter periods (specifically 5 and 10 minutes) have been found related to engagement/challenge levels.

6.4 GLMM and Individual Models

In RQ4c, we investigated how we can build a model for inferring attentional states and engagement/challenge levels of office workers. Based on the GLMM results, in-situ engagement/challenge levels, the number of application switches in the last 45 minutes and application usage in the last 5 minutes before the ESM delivery are negatively related to the attentional states, engagement, and challenge levels. As the number of application switches and application usage increased, the participants were most likely to be in "bored" state, i.e., work their engagement and challenge levels decreased. The result is in line with previous works in the literature. Similarly, the previous studies [43, 60, 61] showed that an increase in application usage is a sign of boredom. As stated above, the time interval used in those studies differed from our settings. When the time intervals are considered, we conclude that the number of switches between applications in a longer time is effective, whereas the duration of application usage is determinant in a shorter period of time. Since boredom is a state of mind in which one searches for a stimulus, in today's world, most of the mobile phone users engage with their devices when they are in such a mood. As expected, when users switch between mobile applications, it may be a sign of boredom, and that means the user is not engaged or challenged with his/her work. Similarly, when users seek a stimulus, they use their mobile phones for a longer duration.

Our study shows that GLMM fits with an MCMC method may be preferable to using individual models built with random forest model when there is not sufficient data available since GLMM does not require a high number of data points, and it also incorporates random effects itself, which facilitates individual modeling. This model may be a solution for a cold-start problem stated in [81]. Although, the difference in terms of performance between GLMM and individual random forest model appears to be marginal, the difference between them increases varying the target level. As it is possible to observe in Table 10, GLMM outperformed other classifiers for 5-level-engagement and 5-level-challenge. However, the performance difference does not become salient the reasons of which can be attributed to the following:

- Although the overall performances of GLMM and individual random forest were similar in terms of the prediction of the attentional states (approximately %50 accuracy), individual RF failed to predict the users with a relatively low number of data points. As expected, the accuracy increases with the number of individual data points. Because of that, the standard deviation of the accuracy values of the individual RF classifier is higher than GLMM or general RF. Even when the number of data points is low, GLMM successfully makes use of population-level mean in the model itself to predict the outcome variable. Finally, it is possible to observe that the fluctuations in the accuracy values are lower for GLMM.
- Similarly, in [81] the individual random forest model shows similar accuracy as the general random forest model with training set of 45 samples. The individual model provides more accurate results when there are more than 45 data points in the training set. Our maximum number of points per user in this study is around 50. The reason of under-performance of individual models compared to GLMM or general models could be related to the limited number of points. In addition, there might be other features such as user's location or activity, which could be more informative for the prediction of the levels of engagement/challenge and attentional states.

As a result, GLMM might be considered as a good solution specifically when the data points are below 40. GLMM methodologically is designed for modeling longitudinal data. It fits a population-level mean in addition

to an individual-level mean (as illustrated in Figure 3); hence, when a new user is added to the system, at first, population-level predictions could be used. Instead of fitting a general random forest model, and separate individual random forest models, GLMM may handle both.

7 IMPLICATIONS

As stated earlier, we made an effort to understand and model knowledge workers' responsiveness to the questions related to their work engagement/challenge, and their attentional states. As seen in a very recent study [13], focus level of users can be effective on perception of the mobile notifications and adherence to health interventions sent via mobile phones. In this regard, we carried out a data collection study among office workers capturing their in-situ engagement/challenge levels. Then, we extracted responsiveness metrics in order to measure and better understand their responsiveness styles. We also inferred their engagement/challenge levels and attentional states through mobile application usage. Our results provide insights related to personal and social factors effective on work engagement and challenge levels. Our aim was not to intervene in the behavior of the participants, instead, to collect contextual data for understanding their office environments and make predictions related to their responsiveness and engagement/challenge levels in office environments so that implications regarding responsiveness to mobile health applications could be better understood.

We believe that this study provides a series of insights that can inform the design of future positive mobile intervention systems, especially for workplace settings and, at the same time, interesting directions for researchers working in this field. First of all, as the previous studies demonstrated, office environment (e.g., office type, number of colleagues) in which knowledge workers spend their work hours gives clues about their work style. Distractions generated in an office environment, reactions to those distractions, and employees' attentional states caused by those might differ with the office type. For these reasons, office-related factors should be considered before sending well-being related interventions through mobile phones to knowledge workers who mainly have a desk job. For example, an employee in a shared office most probably face multiple sources of distractions; hence, the work engagement level of that employee will differ based on those distractions. Besides, as our study showed, the degree of being influenced by co-workers (in terms of social norms) changes with the number of co-workers around. Employees in private offices might be more receptive to mobile notifications since their office environment does not comprise that much distraction, and the knowledge workers are not influenced by their colleagues as much as the ones in a shared office. We did not investigate those factors in our study comprehensively, but further studies might elaborate on the findings.

Secondly, topic awareness and whether employees have a health problem regarding that topic affect responsiveness. However, as previous studies suggested, awareness and well-being levels should surpass a critical level in order to achieve significant results. In further studies, participants who have very low or very high levels of awareness might be recruited, and more generalized results about the effects of awareness and well-being on responsiveness might be presented.

Thirdly, we have once more shown that mobile application usage is a successful indicator for measuring work-related states such as "bored" or "focused" states. We have shown the effects of various application usage metrics at different time intervals. We have shown that not only considering a longer period for switches between applications but also considering a short period of time for the duration of application usage would be quite useful for the inference of work engagement, challenge, and attentional states. Further studies could benefit from this finding.

Finally, we are in a decade of increasing personalized models in ubiquitous computing. Individual models are quite useful when there is plenty of data. In our study, we obtained better results from GLMM compared to individual random forest classifier. Hence, when dealing with a limited number of data, our suggestion is to use those generalized models at first as a solution to the cold-start problem in future studies.

8 LIMITATIONS

The main limitation of our study is the composition of the participants of the study. The results of this study may not be representative of the general population due to the limitation of the convenience sampling, we could reach only office workers working in one city. In addition, we obtained a limited number of data points from the participants. Responding to ESM messages six times can be troublesome for participants; because of the participation burden, their number of responses was low and also it might have caused the high drop-out rate in terms of participation to the experiment in the first days. For this reason, we had to include the participants with a relatively low threshold of 25% response rate, which is different from the previous studies in the literature that included participant data with at least 33% or 50% response rates. The lower threshold allowed us to include as many participant data as possible to inform our models. In addition, our modeling aimed at handling the "cold-start problem". The number of responses could potentially have been affected by the perceived value of the mobile application. Compliance could be improved with a better design, which reduces the participant burden. It could be better reporting whether non-response to the notifications was because of technical problems or participant burden [76]. With a higher number of participants and a higher number of responses, the correlation results can be re-validated and can be more generalized. In addition, the accuracy values could be improved with a higher number of data points for the GLMM and individual classifiers. The main reason for not achieving high accuracy values even in individual random forest classifier might be the low number of data points for each participant. Note that the number of data points in the categories of the target variables (specifically engagement/challenge levels) was unbalanced (e.g., 132 in category 1 of engagement, 52 in category 3 of engagement) and bimodal for engagement levels. Because of that, classifiers might not be able to learn that category as well as the other categories with a higher number of data. An increasing number of data points may lead to more balanced categories, or, at least, it might enable to improve the performance of the classifier. One possible solution is to merge the categories; however, in this study, we wanted to show the results performed on the original data set without any category aggregation.

It is important to note that adjustments are often necessary when multiple comparisons are carried out. The reason for carrying out a post-hoc test is to correct the inflation rate of the value of α. If the number of tests increases dramatically, the inflated family-wise type I error rate (α multiple) would reach one [32]. One common solution is to control type I error by using the conservative Bonferroni correction method, which divides the raw p values by the number of tests m [32]. However, it has been shown that reducing the type I error for null associations increases the type II error for those associations that are not null [62]. In this study, we did not apply any adjustments since (1) we wanted to avoid type I error; and (2) multiple comparisons were made across the variables from different data sources. The criticism regarding multi-comparison practices in the literature is that studies reporting correlations usually include several tests between different variables measured in the same sample and they fail to evaluate the occurrence of type I errors when many tests were performed using the same data [11].

Another limitation is that we only collected social-norm, number of colleagues, and musculoskeletal discomfort once at the beginning of the experiment. These measures might be different based on the user's context (e.g., location and time), hence they could have been measured dynamically with ESM. However, there is a trade-off between collecting those measurements once and during the study. Since there was already a participation burden, we did not want the participants to be overwhelmed with additional ESM questions keeping the number of ESM questions to a minimum. In a more extensive study, these quantities could be dynamically measured with ESM questions and the results could be compared with the results of this study.

One more limitation of the study is the focus only on mobile application usage variables for predicting attentional states and engagement/challenge levels. The goal of this work was to investigate the feasibility of using a hierarchical model and compare its performance with widely used random forest models on limited data

sets. This work could be extended by considering additional data sources, e.g., other sensor data from mobile and wearable devices.

Finally, it is worth noting that the relations found in the study are not causal. A causal inference study with a larger data set is a very promising direction in our opinion. The data was collected among participants of a single country, and there may be cultural and individual characteristics that have not been considered in the study, and the result might vary for other populations.

9 CONCLUSION

In this study, we have discussed the results of a detailed study on the responsiveness of office workers to the mobile notifications for a well-being behavior intervention, considering their work engagement and challenge levels. We have considered a variety of factors such as musculoskeletal discomfort, awareness about sitting behavior, mobile application usage, and office related factors for inferring the responsiveness and work engagement/challenge levels. Then, we have modeled the attentional states, engagement and challenge levels of knowledge workers using mobile application parameters with generalized linear mixed models with Markov Chain Monte Carlo Method.

We have considered a variety of common response styles in this study. We have emphasized the significant effects of mobile application usage on the work engagement/challenge levels of office employees during work hours and their response styles. We have considered the repeated-measures design of our data set while modeling, which is commonly ignored by previous studies. We have showed that the efficiency of GLMMs for modeling the unbalanced and insufficient number of data and that the accuracy of GLMM can be better than other classifiers even in individual-level. We have also presented GLMM as a solution for cold-start problems at individual-level because GLMM enables to model at both *population-* and *individual-level*.

In general, the key contribution of this work is methodological and we believe that this study might be helpful for researchers working in the area of behavior modeling and intervention based on mobile and ubiquitous systems.

ACKNOWLEDGMENTS

This work is supported by The Scientific and Technological Research Council of Turkey under Tubitak BIDEB-2219 grant no 1059B191500728.

REFERENCES

[1] Emmeke Aarts, Matthijs Verhage, Jesse V. Veenvliet, Conor V. Dolan, and Sophie Van Der Sluis. 2014. A solution to dependency: using multilevel analysis to accommodate nested data. *Nature Neuroscience* 17, 4 (2014), 491.

[2] Icek Ajzen. 1991. The theory of planned behavior. *Organizational Behavior and Human Decision Processes* 50, 2 (1991), 179–211.

[3] Icek Ajzen and Martin Fishbein. 1980. *Understanding Attitudes and Predicting Social Behavior.* Prentice-Hall.

[4] Jonathan Z. Bakdash and Laura R. Marusich. 2017. Repeated measures correlation. *Frontiers in Psychology* 8 (2017), 456.

[5] Nikola Banovic, Christina Brant, Jennifer Mankoff, and Anind Dey. 2014. ProactiveTasks: the short of mobile device use sessions. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services.* ACM, 243–252.

[6] Ronald De Vera Barredo and Kelly Mahon. 2007. The effects of exercise and rest breaks on musculoskeletal discomfort during computer tasks: an evidence-based perspective. *Journal of Physical Therapy Science* 19, 2 (2007), 151–163.

[7] Robert Bixler and Sidney D'Mello. 2013. Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces.* ACM, 225–234.

[8] Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32.

[9] N. E. Breslow and D. G. Clayton. 1993. Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* 88, 421 (mar 1993), 9–25. https://doi.org/10.1080/01621459.1993.10594284

[10] Barry Brown, Moira McGregor, and Donald McMillan. 2014. 100 days of iPhone Use: understanding the details of mobile device use. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services.* ACM, 223–232.

[11] José Manuel Caperos, Ricardo Olmos, and Antonio Pardo. 2016. Inconsistencies in reported p-values in Spanish journals of psychology: The case of correlation coefficients. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 12, 2 (2016), 44.

[12] Yung-Ju Chang and John C. Tang. 2015. Investigating mobile users' ringer mode usage and attentiveness and responsiveness to communication. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 6–15.

[13] Woohyeok Choi, Sangkeun Park, Duyeon Kim, Youn-kyung Lim, and Uichin Lee. 2019. Multi-stage receptivity model for mobile just-in-time health intervention. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 39.

[14] Arthur Couch and Kenneth Keniston. 1960. Yeasayers and naysayers: Agreeing response set as a personality variable. *The Journal of Abnormal and Social Psychology* 60, 2 (1960), 151.

[15] Lee J. Cronbach. 1946. Response sets and test validity. *Educational and Psychological Measurement* 6, 4 (1946), 475–494.

[16] Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* 7, Jan (2006), 1–30.

[17] Daniel A. Epstein, Daniel Avrahami, and Jacob T. Biehl. 2016. Taking 5: Work-breaks, productivity, and opportunities for personal informatics for knowledge workers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 673–684.

[18] Otto Fenichel. 1951. On the psychology of boredom. In *Organization and Pathology of Thought*. Columbia University Press New York, NY, 349–361.

[19] W. Holmes Finch, Jocelyn E. Bolin, and Ken Kelley. 2014. *Multilevel Modeling Using R*. CRC Press, Taylor & Francis Group, Boca Raton, FL, USA.

[20] Martin Fishbein and Icek Ajzen. 1975. *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*. Addison-Wesley, MA.

[21] Cynthia D. Fisherl. 1993. Boredom at work: a neglected concept. *Human Relations* 46, 3 (1993), 395–417.

[22] Traci Galinsky, Naomi Swanson, Steven Sauter, Robin Dunkin, Joseph Hurrell, and Lawrence Schleifer. 2007. Supplementary breaks and stretching exercises for data entry operators: a follow-up field study. *American Journal of Industrial Medicine* 50, 7 (2007), 519–527.

[23] Andrew Gelman, Hal S. Stern, John B. Carlin, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis*. Chapman and Hall/CRC.

[24] Emma S. George, Gregory S. Kolt, Richard R. Rosenkranz, and Justin M. Guagliano. 2014. Physical activity and sedentary time: male perceptions in a university work environment. *American Journal of Men's Health* 8, 2 (2014), 148–158.

[25] Thomas Goetz, Anne C Frenzel, Nathan C. Hall, Ulrike E. Nett, Reinhard Pekrun, and Anastasiya A. Lipnevich. 2014. Types of boredom: an experience sampling approach. *Motivation and Emotion* 38, 3 (2014), 401–419.

[26] Jarrod D. Hadfield et al. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software* 33, 2 (2010), 1–22.

[27] Marija Ham, Marina Jeger, and Anita Frajman Ivković. 2015. The role of subjective norms in forming the intention to purchase green food. *Economic Research* 28, 1 (2015), 738–748.

[28] Jia He and Fons J.R. Van de Vijver. 2013. A general response style factor: Evidence from a multi-ethnic study in the Netherlands. *Personality and Individual Differences* 55, 7 (2013), 794–800.

[29] Matthew V. Hibbing, Matthew Cawvey, Raman Deol, Andrew J. Bloeser, and Jeffery J. Mondak. 2017. The relationship between personality and response patterns on public opinion surveys: The Big Five, extreme response style, and acquiescence response style. *International Journal of Public Opinion Research* 31, 1 (2017), 161–177.

[30] Collin Hitt, Julie Trivitt, and Albert Cheng. 2016. When you say nothing at all: The predictive power of student effort on surveys. *Economics of Education Review* 52 (2016), 105–119.

[31] 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*. ACM, 909–918.

[32] Mohieddin Jafari and Naser Ansari-Pour. 2019. Why, when and how to adjust your P values? *Cell Journal (Yakhteh)* 20, 4 (2019), 604.

[33] Ron Kohavi et al. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, Vol. 14. Montreal, Canada, 1137–1145.

[34] Reed Larson and Mihaly Csikszentmihalyi. 1983. The experience sampling method. *New Directions for Methodology of Social & Behavioral Science* 15 (1983), 41–56.

[35] Neal Lathia, Veljko Pejovic, Kiran K. Rachuri, Cecilia Mascolo, Mirco Musolesi, and Peter J. Rentfrow. 2013. Smartphones for large-scale behavior change interventions. *IEEE Pervasive Computing* 12, 3 (2013), 66–73.

[36] Robert LiKamWa, Yunxin Liu, Nicholas D. Lane, and Lin Zhong. 2013. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 389–402.

[37] Yuhan Luo, Bongshin Lee, Donghee Yvette Wohn, Amanda L. Rebar, David E Conroy, and Eun Kyoung Choe. 2018. Time for break: understanding information workers' sedentary behavior through a break prompting system. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 127.

[38] Ruth M. Mabry, Zakiya Q. Al-Busaidi, Marina M. Reeves, Neville Owen, and Elizabeth G. Eakin. 2013. Addressing physical inactivity in Omani adults: perceptions of public health managers. *Public Health Nutrition* 17, 3 (2013), 674–681.
[39] Giuseppe Mantovani. 1996. Social context in HCI: A new framework for mental models, cooperation, and communication. *Cognitive Science* 20, 2 (1996), 237–269.
[40] Gloria Mark, Shamsi Iqbal, Mary Czerwinski, and Paul Johns. 2014. Capturing the mood: Facebook and face-to-face encounters in the workplace. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1082–1094.
[41] Gloria Mark, Shamsi T. Iqbal, Mary Czerwinski, and Paul Johns. 2014. Bored mondays and focused afternoons: the rhythm of attention and online activity in the workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3025–3034.
[42] Akhil Mathur, Nicholas D. Lane, and Fahim Kawsar. 2016. Engagement-aware computing: modelling user engagement from mobile contexts. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 622–633.
[43] Aleksandar Matic, Martin Pielot, and Nuria Oliver. 2015. Boredom-computer interaction: boredom proneness and the use of smartphone. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 837–841.
[44] Charles E. McCulloch and Shayle R. Searle. 2004. *Generalized, Linear, and Mixed Models*. John Wiley & Sons.
[45] Abhinav Mehrotra, Mirco Musolesi, Robert Hendley, and Veljko Pejovic. 2015. Designing content-driven intelligent notification mechanisms for mobile applications. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 813–824.
[46] 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 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1021–1032.
[47] Shayan Mirjafari, Kizito Masaba, Ted Grover, Weichen Wang, Pino Audia, Andrew T. Campbell, Nitesh V. Chawla, Vedant Das Swain, Munmun De Choudhury, Anind K. Dey, Sidney K. D'Mello, Ge Gao, Julie M. Gregg, Krithika Jagannath, Kaifeng Jiang, Suwen Lin, Qiang Liu, Gloria Mark, Gonzalo J. Martinez, Stephen M. Mattingly, Edward Moskal, Raghu Mulukutla, Subigya Nepal, Kari Nies, Manikanta D. Reddy, Pablo Robles-Granda, Koustuv Saha, Anusha Sirigiri, and Aaron Striegel. 2019. Differentiating higher and lower job performers in the workplace using mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–24. https://doi.org/10.1145/3328908
[48] Hazwani Mohd Mohadis and Nazlena Mohamad Ali. 2018. Smartphone application for physical activity enhancement at workplace: would office workers actually use it?. In *2018 International Conference on Information and Communication Technology for the Muslim World (ICT4M)*. IEEE, 144–149.
[49] Melissa Monsey, Irina Ioffe, Angela Beatini, Betsy Lukey, Andrea Santiago, and Anne Birge James. 2003. Increasing compliance with stretch breaks in computer users through reminder software. *Work* 21, 2 (2003), 107–111.
[50] Dan Morris, A.J. Brush, and Brian R. Meyers. 2008. SuperBreak: using interactivity to enhance ergonomic typing breaks. In *Proceedings of the 2008 SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1817–1826.
[51] Rachel L. Morrison and Keith A. Macky. 2017. The demands and resources arising from shared office spaces. *Applied Ergonomics* 60 (2017), 103–115.
[52] Sarah L. Mullane, Meynard J. L. Toledo, Sarah A. Rydell, Linda H. Feltes, Brenna Vuong, Noe C. Crespo, Mark A. Pereira, and Matthew P. Buman. 2017. Social ecological correlates of workplace sedentary behavior. *International Journal of Behavioral Nutrition and Physical Activity* 14, 1 (2017), 117.
[53] Dominique Nduhura and Michael Prieler. 2017. When I chat online, I feel relaxed and work better: exploring the use of social media in the public sector workplace in Rwanda. *Telecommunications Policy* 41, 7-8 (2017), 708–716. http://dx.doi.org/10.1016/j.telpol.2017.05.008
[54] Harri Oinas-Kukkonen. 2010. Behavior change support systems: a research model and agenda. In *International Conference on Persuasive Technology*. Springer, 4–14.
[55] Tadashi Okoshi, Kota Tsubouchi, Masaya Taji, Takanori Ichikawa, and Hideyuki Tokuda. 2017. Attention and engagement-awareness in the wild: a large-scale study with adaptive notifications. In *2017 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 100–110.
[56] Delroy L. Paulhus. 1991. Measurement and control of response bias. In *Measures of Personality and Social Psychological Attitudes*. Elsevier, San Diego, 17–59.
[57] Veljko Pejovic, Neal Lathia, Cecilia Mascolo, and Mirco Musolesi. 2016. Mobile-based experience sampling for behaviour research. In *Emotions and Personality in Personalized Services*. Springer, Cham, 141–161.
[58] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. 2001. Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 10 (2001), 1175–1191.
[59] Martin Pielot, Linas Baltrunas, and Nuria Oliver. 2015. Boredom-triggered proactive recommendations. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. ACM, 1106–1110.
[60] 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.

[61] 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 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 825–836.
[62] Kenneth J. Rothman. 1990. No adjustments are needed for multiple comparisons. *Epidemiology* (1990), 43–46.
[63] Wilmar B. Schaufeli, Arnold B. Bakker, and Marisa Salanova. 2006. The measurement of work engagement with a short questionnaire: a cross-national study. *Educational and Psychological Measurement* 66, 4 (2006), 701–716.
[64] Shalom H. Schwartz. 1968. Awareness of consequences and the influence of moral norms on interpersonal behavior. *Sociometry* 31, 4 (1968), 355–369.
[65] Aram Seddigh, Erik Berntson, Loretta G. Platts, and Hugo Westerlund. 2016. Does personality have a different impact on self-rated distraction, job satisfaction, and job performance in different office types? *PloS ONE* 11, 5 (2016).
[66] David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64, 4 (2002), 583–639.
[67] Piiastiina Tikka and Harri Oinas-Kukkonen. 2016. RightOnTime: the role of timing and unobtrusiveness in behavior change support systems. In *International Conference on Persuasive Technology*. Springer, 327–338.
[68] Minh H. Tran, Jun Han, and Alan Colman. 2009. Social context: Supporting interaction awareness in ubiquitous environments. In *2009 6th Annual International Mobile and Ubiquitous Systems: Networking & Services, MobiQuitous*. IEEE, 1–10.
[69] Dennis C. Turk and Thomas E. Rudy. 1991. Neglected topics in the treatment of chronic pain patients – relapse, noncompliance, and adherence enhancement. *Pain* 44, 1 (1991), 5–28.
[70] Gašper Urh and Veljko Pejović. 2016. TaskyApp: inferring task engagement via smartphone sensing. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 1548–1553.
[71] Saskia van Dantzig, Gijs Geleijnse, and Aart Tijmen Halteren. 2013. Toward a persuasive mobile application to reduce sedentary behavior. *Personal and Ubiquitous Computing* 17, 6 (2013), 1237–1246.
[72] Patrick Van Kenhove, Katrien Wijnen, and Kristof De Wulf. 2002. The influence of topic involvement on mail - survey response behavior. *Psychology & Marketing* 19, 3 (2002), 293–301.
[73] Viswanath Venkatesh and Fred D. Davis. 2000. A theoretical extension of the technology acceptance model: four longitudinal field studies. *Management Science* 46, 2 (2000), 186–204.
[74] Aku Visuri, Niels van Berkel, Chu Luo, Jorge Goncalves, Denzil Ferreira, and Vassilis Kostakos. 2017. Predicting interruptibility for manual data collection: a cluster-based user model. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 12.
[75] Birgit Wallmann-Sperlich, Jens Bucksch, Sven Schneider, and Ingo Froboese. 2014. Socio-demographic, behavioural and cognitive correlates of work-related sitting time in German men and women. *BMC Public Health* 14, 1 (2014), 1259.
[76] Cheng K. Fred Wen, Stefan Schneider, Arthur A. Stone, and Donna Spruijt-Metz. 2017. Compliance with mobile ecological momentary assessment protocols in children and adolescents: a systematic review and meta-analysis. *Journal of Medical Internet Research* 19, 4 (2017).
[77] Marc C. Willemsen, Hein de Vries, Gerard van Breukelen, and Brian Oldenburg. 1996. Determinants of intention to quit smoking among Dutch employees: the influence of the social environment. *Preventive Medicine* 25, 2 (1996), 195–202.
[78] Ping Yu, Haocheng Li, and Marie-Pierre Gagnon. 2009. Health IT acceptance factors in long-term care facilities: a cross-sectional survey. *International Journal of Medical Informatics* 78, 4 (2009), 219–229.
[79] Fengpeng Yuan, Xianyi Gao, and Janne Lindqvist. 2017. How busy are you?: predicting the interruptibility intensity of mobile users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 5346–5360.
[80] Alexandros Zenonos, Aftab Khan, Georgios Kalogridis, Stefanos Vatsikas, Tim Lewis, and Mahesh Sooriyabandara. 2016. HealthyOffice: mood recognition at work using smartphones and wearable sensors. In *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. IEEE, 1–6.
[81] Manuela Züger, Sebastian C. Müller, André N. Meyer, and Thomas Fritz. 2018. Sensing interruptibility in the office: a field study on the use of biometric and computer interaction sensors. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 591.

A DESCRIPTION OF THE MODELS

For the Kendall's Tau correlation analyses regarding RQ1-RQ4a, the data set consisted of each user's responses, i.e. it is not a repeated-measures design. Since the normality of the parameters could not be satisfied, Kendall's tau is selected for the correlation analysis. For the repeated-measures correlation analysis regarding RQ4b, our ESM data set consisted of the participants' responses measured at different time points throughout the experiment. Hence, it comprised dependent repeated-measures of individuals with varying sizes. The use of simple correlation statistics such as Pearson to study the potential relations between the constructs on such data set may produce biased and

erroneous results due to the violation of statistical independence and/or differing patterns between-participants versus within-participants [4]. Repeated-measures correlation, an atypical application of ANCOVA, is a recently proposed statistics for determining the common within-individual association for paired measures assessed on two or more occasions for multiple individuals [4]. For these reasons, we conducted repeated-measures correlation analysis for RQ4b.

We employed GLMM analysis for RQ4c because our data set consisted of multiple responses of participants so that the data points could not be considered as independent. In addition, the number of responses is not equal for each participant. Moreover, the response variables were binary (attentional states) and ordinal (engagement and challenge levels), which means that their distributions were not Gaussian. For these reasons, the assumptions of approaches such as ANOVA have been violated. In this case, it is suggested to use approaches such as generalized linear mixed models (i.e., hierarchical modeling or multilevel modeling) [1, 4].

Generalized linear models (GLM) extend linear models by handling response variables with non-normal distribution [26, 44]. Generalized linear mixed model (GLMM) incorporates random effects to GLMs. Random effects are mostly individuals, population, species, or vials with a large number of levels [26], which impact the dependent variable because of the variations among the levels [19]. In its simplest form, a GLMM can be written as in Equation 2, where y is a $N \times 1$ column vector, the target variable (N is the number of data points); x is the matrix of fixed predictors with the dimension of $N \times p$ (p is the number of predictors); β is a $p \times 1$ column vector of the fixed-effects coefficients; z is the design matrix with the dimension of $N \times q$ corresponding to the q random predictors (accounting for the random complement to the fixed x); b is a $q \times 1$ vector of the random effects (the random complement to the fixed β); and β_0 is a $N \times 1$ residuals vector. For our study, y is the engagement/challenge level as the target variable, x represents the predictors, namely the mobile application usage parameters, e.g., number of switches, total application usage in the last 5 minutes, and z indicates the participants as the random predictors. Since GLMM incorporates random effects (i.e. individuals), it is possible to obtain an individual prediction from GLMM. GLMM simply fits separate linear regression lines for each random effect included.

$$y = \beta_0 + x\beta + zb \tag{2}$$

For model selection, Deviance Information Criterion (DIC) is generally used. DIC, as other information criteria, is defined based on the principle of leveraging goodness of fit and model complexity. Unlike information criteria based on point estimates of the model parameters (e.g. Akaike's Information Criterion), DIC makes use of the posterior distribution over the model parameters, given the data. This is relevant in the Bayesian setting where the model posterior is often characterized in terms of a sample obtained through e.g. MCMC approaches. Lower DIC values should be preferred for model selection [66]. Furthermore, the convergence of the Markov chains is checked using the Gelman-Rubin diagnostic criterion, where 1.002 and below indicates the convergence [23].

B MODELING AND FEATURE SET SELECTION USING GLMM

Repeated-measures correlation results gave an idea about which variables of application usage are related to engagement/challenge levels. However, we are not still sure which of these features will indeed be effective for modeling. As *rmcorr* shows the linear association between the variables and GLMM is inherently a linear model, we took the correlation results as a basis in feature set selection and model building instead of variable selection with Gini or other metrics.

As can be seen from the Tables 7 and 8; app use (5-min), no of apps (5-min, 10-min, 30 min, 45-min, and 60-min), no of switch (5-min, 10-min, 30-min, 45-min, and 60-min), communication category usage (5-min, 10-min, 45-min, and 60-min) are the most related variables to the engagement and challenge levels. Hence, it was planned to include a combination of those variables in the GLMM analysis. However, before fitting a GLMM, we calculated

the correlations between the predictor variables in order to detect possible multicollinearity issues. Strong correlations were observed between some of the variables such as the number of switches between applications (45-min) and the number of apps (45-min) ($r_{rm} = .91, p < .001, N = 418$) signaling a multicollinearity problem. Then, we took into account the pairs, which are no of switch (60-min), no of apps (10-min), app use (5-min), and no of switch (45-min) having a correlation less than .70. The combination of those variables were given to the GLMM trials.

Then, several GLMMs were built for modeling attentional states, engagement levels and challenge levels of users. During this step, we took a different approach from the studies in the literature. We subsampled the data set five times rather than solely using the original one. This is due to the fact that the model and feature selection should not be affected by the number of responses or the highest/lowest amount of application usage. If the majority of data comes from a few users in the data set, selected features and model might not be representative for all users. In the subsampling approach, in each subsample, we eliminated the data points of the participants with the highest and lowest response rates incrementally. To be more specific, the first sample is the full original data set consisting of all the responses of users. In the second data set, two users were removed from the first data set. These users were those with the highest and lowest number of ESM responses. In the third data set, we excluded two more users with the highest and lowest number of ESM responses in the second data set. In each iteration, we ended up with a data set, which is more uniform in terms of participants' responses than the data set from the previous step. At the end of this process, we selected the model, which gave consistently the lowest DIC, which is the preferred metric for Bayesian model selection [66] on all the data sets.

The response variables are binary (attentional states) and ordinal (engagement and challenge levels). The model fitting process was carried out incrementally by adding constructs one at a time according to their geometric mean performance across the data sets. Each model was run five times. In the end, the mean and standard deviation of the overall performance of each model was reported.

For modeling attentional states, first, no of switch (60-min), no of apps (10-min), app use (5-min) and no of switch (45-min) variables were given to GLMMs separately. Then, those variables were given to the model by pairs. In total, 11 combinations of those variables were used for predicting attentional states. Table 11 (top) presents the top three model runs which gave the lowest DIC for predicting attentional states on five different data sets with the mean and standard deviation of the DIC values for each run. Similarly, engagement and challenge levels of users were modeled with GLMM. Same 11 models for each were built. The middle and bottom tables in Table 11 summarize DIC values obtained for the top three models of engagement and challenge models, respectively. For all target variables (attentional states, engagement and challenge levels), the models with no of switch (45-min) and app use (5-min) predictors gave the lowest DIC; hence, those models were selected.

Table 11. Deviance Information Criteria (DIC) estimates for the GLMMs used to predict the attentional states (top), engagement levels (middle) and challenge levels (bottom). Three runs with different windows sizes were performed in each subsampled data set and the mean and standard deviation of these runs are summarized below. NOS_x: No of switch in the last x minutes, NOA_x: No of apps in the last x minutes, AU_x: App use in the last x minutes.

Attentional States Model Runs						
Model No	Covariates	Data set 1 (N = 372)	Data set 2 (N = 313)	Data set 3 (N = 257)	Data set 4 (N = 297)	Data set 5 (N = 224)
1	NOS_{60} + AU_5	503.27±.78	421.26±1.51	337.21±.89	391.30±.79	303.24±1.27
2	NOS_{45} + NOA_{10}	504.54±2.00	423.71±1.16	339.11±1.94	393.32±2.55	300.45±3.19
3	NOS_{45} + AU_5	502.55±1.05	421.78±1.62	335.60±2.49	391.24±3.74	299.49±1.19
Engagement Levels Model Runs						
Model No	Covariates	Data set 1 (N = 418)	Data set 2 (N = 355)	Data set 3 (N = 292)	Data set 4 (N = 323)	Data set 5 (N = 255)
1	NOS_{45} + AU_5	1235.46±3.41	1040.74±1.80	862.34±1.78	949.52±3.71	754.79±1.61
2	NOS_{45} + NOA_{10}	1238.83±1.43	1037.72±3.88	858.34±3.59	953.77±2.05	758.83±1.17
3	NOS_{60} + AU_5	1239.64±1.04	1042.00±2.43	864.96±1.80	951.23±2.83	753.02±4.86
Challenge Levels Model Runs						
Model No	Covariates	Data set 1 (N = 418)	Data set 2 (N = 355)	Data set 3 (N = 292)	Data set 4 (N = 323)	Data set 5 (N = 255)
1	NOS_{60} + NOA_{10}	1199.97±3.93	1013.63±2.92	821.35±4.55	914.07±2.90	711.49±3.87
2	NOS_{45} + NOA_{10}	1192.73±6.67	1011.47±4.51	824.88±1.90	913.99±5.04	707.31±9.00
3	NOS_{45} + AU_5	1199.25±1.39	1009.10±3.52	820.65±5.54	913.86±6.63	710.52±5.16