Optimizing for Happiness and Productivity: Modeling Opportune Moments for Transitions and Breaks at Work

Harmanpreet Kaur[1], **Alex C. Williams**[2], **Daniel McDuff**[3], **Mary Czerwinski**[3], **Jaime Teevan**[3], **Shamsi T. Iqbal**[3]
[1]University of Michigan, [2]University of Waterloo, [3]Microsoft Research
harmank@umich.edu, alex.williams@uwaterloo.ca, {damcduff,marycz,teevan,shamsi}@microsoft.com

ABSTRACT
Information workers perform jobs that demand constant multitasking, leading to context switches, productivity loss, stress, and unhappiness. Systems that can mediate task transitions and breaks have the potential to keep people both productive and happy. We explore a crucial initial step for this goal: finding opportune moments to recommend transitions and breaks without disrupting people during focused states. Using affect, workstation activity, and task data from a three-week field study ($N = 25$), we build models to predict whether a person should continue their task, transition to a new task, or take a break. The R^2 values of our models are as high as 0.7, with only 15% error cases. We ask users to evaluate the timing of recommendations provided by a recommender that relies on these models. Our study shows that users find our transition and break recommendations to be well-timed, rating them as 86% and 77% accurate, respectively. We conclude with a discussion of the implications for intelligent systems that seek to guide task transitions and manage interruptions at work.

Author Keywords
Affect; Productivity; Workplace; Recommendations

CCS Concepts
•**Human-centered computing** → **User models;** *User studies;* •**Computing methodologies** → **Model development and analysis;**

INTRODUCTION
Information workers operate in an environment where multitasking is common [18, 45] and task priorities shift constantly [66]. In practice, multitasking often leads to context switching as people try to manage different tasks and communication channels at once [20, 28]. As a result, information workers may switch context at inopportune moments—when they have maximum context about their current task and are in a state of flow [16]—resulting in high task-resumption costs and loss of productivity [45]. Switching out of unproductive states, though, is important, since these can lead to stress and unhappiness at work [30], which also leads to loss of productivity [63]. This vicious cycle is hard to break if we consider productivity and affect in isolation: a person's affective state is crucial to their workplace effectiveness [33, 58]. Indeed, the *happy-productive worker* hypothesis claims that information workers cannot be their most productive selves, or do their best work, without first being happy [82].

One way to keep information workers both happy and productive is to recommend state changing actions (such as, "transition to a different task" or "take a break") at times when we believe people to be in unhappy or unproductive states. Such recommendations must be well-timed, as prior work suggests that intelligent systems can do more harm than good if people are interrupted at the wrong times [5, 26, 46, 50]. It is challenging to identify the ideal moments for such recommendations without a fine-grained understanding of people's affective state and work context. Most prior work has relied on wearable sensors to gain some understanding of these factors (e.g., [84]), but several of these sensors are challenging to wear continuously, and are subject to technological failure [11].

Our goal is to identify opportune moments for guiding people towards effective states at work in a minimally invasive way. We rely on a tool that logs workstation activity, daily task information, and affect derived from facial expressions in a privacy-preserving way; and conduct a four-week field study with 25 participants at a large technology company. Our work has two phases: (1) we use three weeks of data collected via our tool to build predictive models that jointly optimize people's positive affect and productivity, and (2) we deploy these models to make real-time recommendations of transitions and breaks for our participants, and obtain their feedback on the timing of these recommendations.

Our results show that we can jointly model positive affect and productivity with reasonable goodness-of-fit (R^2 0.2–0.7) and low error (<15%). While all our logged features are important for the models, the importance of each feature varies by individual. When applied in practice, these models identify opportune moments for transitions and breaks in real-time with 85.7% and 77% accuracy, respectively. Our user study shows that participants appreciate timely reminders about these actions, follow them to replenish their energy, and are more reflective about their work as a result. We end with implications for building intelligent systems for workplace well-being.

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376817

RELATED WORK

Our work relies on principles from HCI, ubiquitous computing, organizational behavior, and psychology, as described below.

The Happy-Productive Worker

Organizational behavior studies show that people who have a happy disposition at work tend to have higher rated (i.e., more productive) performance measures [33, 82]. Coined the "happy-productive worker hypothesis," this has been studied in several organizational settings with different operationalizations of happiness (e.g., job satisfaction, lack of emotional exhaustion) and productivity (e.g., meeting monthly targets, ratings from manager). The hypothesis has been supported by prior work in specific domains [15, 73, 75, 81].

Observing 42 software developers—an important class of information workers—Graziotin et al. [31] find that happy developers are indeed better at analytical problem solving and critical thinking. Similarly, several researchers have mined publicly available issue-tracking data from software repositories (e.g., Apache Jira) to find that positive emotions are correlated with shorter issue fixing time [21, 55, 65]. In a more recent paper, Graziotin et al. suggest that it is more cost-effective to study unhappiness and reduce it—this automatically reduces stress and improves productivity [30]. It is evident from this work that productivity and happiness are intertwined; thus, we consider both in our study.

Multitasking and Interruption Management

Humans are prone to multitasking because they are cognitively capable of switching between tasks [71, 79] and technology supports this practice [9, 18, 28, 64]. However, multitasking often results in switching tasks at inopportune moments, due to internal [2, 49] and external [17, 41] interruptions. Information workers switch windows every 40 seconds [59] and working spheres every 3 minutes [28]. Once interrupted, they can take $\sim$15 minutes to resume their task [45]. An interruption at the wrong time, e.g., when people are in a state of flow [16], can result in lower task productivity [63] and increased frustration, anxiety and annoyance [4, 5, 26, 43, 46, 56].

Theoretical Studies of Breaks and Transitions

Several papers in the psychology literature have studied people's behavior in the context of breaks. Strongman and Burt find that people often take breaks due to internal states of tiredness, boredom, or hunger; and for activities such as walking, socializing, or exercising [74]. Observing 107 employees from diverse industries, Kühnel et al. also find that mental exhaustion is a key reason for taking a break [52]. Prior work is split on what is the most helpful way of taking a break: several studies note that taking shorter breaks more often is ideal for productivity [38, 74], whereas others show that short breaks lead to more fragmented work and thus reduce overall productivity [19]. One reason for these differences could be that people's practices for breaks vary by task and timeline requirements [10, 22, 66]. We explore this hypothesis by combination of task and affect factors in our models.

Task transitions are harder to ascertain for information work: there are minimal changes in the environment, but people perform deliberate attentional reorientation when they switch tasks [69]. McFarlane proposes four methods of interrupting a user for switching tasks: immediate, negotiated, mediated, and scheduled [62]. Our work relies on a mediated strategy, where the system uses contextual information to decide when to recommend a task switch or a break to the user, thereby reducing the burden on the user to pick an optimal moment.

Sensor-based Affect and Productivity Monitoring

Prior work has employed sensor-based monitoring to identify opportune moments for a task switch: [47, 51] use pupillary response to measure cognitive load; [37, 39] study heart rate variability (HRV) as a proxy for focus; and several other studies, including [12, 34, 35, 67, 70, 77] use electromyogram, accelerometry data, electrocardiogram data, skin conductance, sleep and circadian rhythms, mobile phone context, and other signals to measure stress and cognitive load.

A comprehensive overview of sensor-based psychological, physiological, behavioral, and contextual measurements of stress can be found in [3]. Most related to our work is Züger et al.'s prediction of interruptible moments in people's workdays based on a combination of sensor-based data outlined above [84]. They collect ground truth self-reports of interruptibility from people and train personalized models that use data from several sensors to predict if an individual is interruptible at a given time. Complementary to their approach, our models are based on a joint optimization of productivity and happiness, and we build on their work by using data from an emotion, activity, and task logging tool, but without any wearable sensors. Our work extends sensor-based monitoring studies, specifically, those that demonstrate that even simple sensors are valuable for modeling interruptability [25].

Tool-Based Productivity Mediation

Researchers have leveraged different sources of user data to develop systems that help people better manage their attention spans, todos, and overall productivity. For example, systems like Active Progress bar [42], Busybody [40], Groupbar [72], Lilsys [8], Oasis [44], and several others rely on various forms of log data analysis to reduce interruptions and support easy task switching for productive outcomes at work.

Most related to our overarching goal, several papers in the HCI literature have studied work-related breaks and built tools to support them. Epstein et al. conducted an extensive analysis of people's definition of a break and found that these are subjective, though the desired outcomes (e.g., feeling relaxed) are consistent [24]. They designed 13 visualizations to support learning and reflection of people's unique break patterns. Cambo et al. introduced BreakSense, a multi-device application that employs location-based challenges to promote mobility in the workplace [11]. Similarly, Luo et al. designed "Time for Break," a break-prompting system aimed at combating prolonged sedentary behavior, and found that pre-existing habits play an important role in system adoption [54]. Most recently, Tseng et al. developed and studied UpTime, a conversational system built into Slack that improves transition between breaks and work time by blocking distractions (e.g., social media sites) for a fixed period of time [76].

While the goal of these systems is to support taking breaks and maximizing productivity, leveraging user affect remains relatively unexplored in this context. As a first step, we leverage affect, workstation activity, and task information to predict opportune moments for task transitions and breaks for people, to help them become happy-productive workers.

RESEARCH GOALS

Our broad research goal is to help people achieve their work-related goals while also optimizing positive affect in the workplace. To approximate this, we use predicted emotion labels for people's facial expressions, their workstation activity, and their daily task list, to recommend actions for productivity and positive affect at any given time—specifically, switching to a different task or taking a break. Our research question is:

RQ. Can we identify opportune times for transitioning tasks and taking breaks for people during their workday?

We study this in two phases: in Phase 1, we develop models to predict opportune moments for intervention using a jointly optimized value for positive affect and productivity; in Phase 2, we study how people respond to recommendations of transitions and breaks based on the jointly optimized value.

PHASE 1: MODEL DEVELOPMENT

To guide people towards positive affect and productivity at work, we performed optimization over data collected about people's emotions, workstation activity, and tasks. Here, we describe how we collected this data, followed by the specifics of our features, the models used for prediction, and the metrics we used for evaluating our prediction models.

Tracking Emotion, Workstation Activity, and Tasks

We collected 9 categories of data for our prediction task: (1) emotion, (2) heart rate, (3) physical movement, (4) interaction data, (5) time of day, (6) day of week, (7) task information, (8) digital actions being performed, and (9) productivity and affect reports. We used an existing Emotion and Activity Logging Software [61] to collect categories 1–6 and 8, and built an interface, FLOWZONE (Figure 1), on top of this software for categories 7 and 9. Categories 1–8 are used as inputs for our models and 9 is used to compute the output.

Emotion and Activity Logging Software

We got emotion expressions and workstation activity by processing data collected by the software via a standard webcam (participant privacy was preserved by never storing raw data). The software [61] analyzes people's facial expressions while at their desk. It consists of a visual and an activity pipeline.

Visual Pipeline. The software processes video data from a webcam. First, it detects faces in the video and extracts landmark positions of key facial features. The distance of the user's face from the camera is extracted using the inter-ocular distance calculated from the facial landmarks. Next, the facial regions of interest are analyzed using an emotion detection algorithm, returning eight probabilities for each of the following basic emotional expressions—anger, disgust, fear, joy, sadness, surprise, contempt, and neutral [23]—with an accuracy of $\sim$87%. It uses Microsoft's publicly-available EmotionAPI to detect emotion expression (for more information on its classification of facial expressions, see [7]). Using image frames, the software also extracts heart rate via the photoplethysmographic signal [60, 68].

Activity Pipeline. The software logs information about the open applications and interactions with computer peripherals. Each time applications are opened, closed, in focus (the front application), minimized, or maximized, it records these activities with the corresponding timestamp. The software only logs the title of the window—indicating the page or application—and these values are hashed before storing. It also logs mouse movements and clicks and keyboard inputs.

FLOWZONE: An Interface to Collect Task Information and Self-Reports on Productivity and Affect

We developed FLOWZONE, a user interface on top of the aforementioned Emotion and Activity Logging Software [61] to collect additional information on people's daily tasks, and self-reports of task progress, productivity, and affect. FLOWZONE is comprised of two components: the Task Tracker, and the Productivity and Affect Self-report interface. [1] The data collected through FLOWZONE is temporally aligned with the data collected by the Emotion and Activity Logging Software.

Task Tracker. The Task Tracker is a simple to-do list interface which asks people about the type of activities involved in doing each task on the to-do list. These activities can be selected from a list of eight: reading, writing, coding, digital communication, brainstorming, paper-based reading/writing, creating spreadsheets, and online information search. The interface also asks for each to-do item's urgency and difficulty, and an estimate for the anticipated completion time for it (Figure 1). Prior work shows that the emotion and activity-based markers can change based on the task being performed [10, 22], making this task information critical for our models.

Productivity and Affect Self-Reports. The Emotion and Activity Logging Software captures 8 emotion labels based on people's facial expressions. Prior work also relies on self-reported affect, noting that these values (*emotion* from facial expressions and *affect* from self-reports) are similar but unique signals of people's affective state [83]. The relationship between the two is an ongoing topic of research (e.g., [6, 27, 32] describe the challenges in determining this relationship).

Given this prior work, we collected self-reports of affect and task progress in addition to emotion from the logging software. People reported affect via 6 variables derived from the Positive and Negative Affect Scale (PANAS) [78]. Of the 6, 3 are positive items (inspired, enthusiastic, determined) and the other 3 are negative items (irritable, nervous, upset) from the original 20 on the scale. People selected values for these using sliders ranging from 0–10. We used a smaller subset of PANAS items to minimize time spent filling out the report (reducing interruption costs)—a practice that has been seen in prior work with similar goals of reducing self-report costs [57, 80]. People also reported on how productive and busy they felt (range: 0–10), and their progress per task (range: 0–100).

[1] Pictures of the interface are included in the supplementary material.

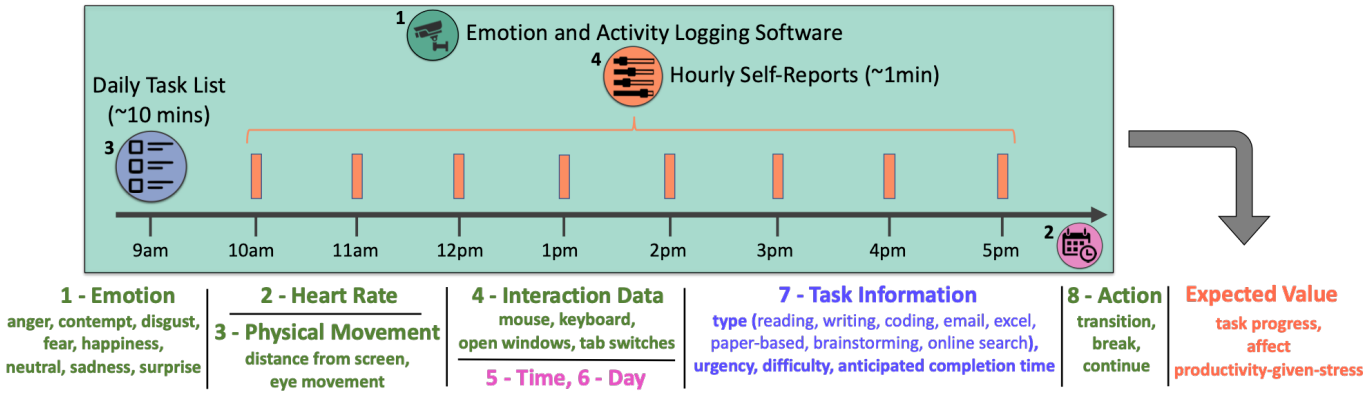

Figure 1. Emotion, Workstation Activity, and Task Tracking setup. The different components are: (1) a logging software that tracks rich emotion and workstation activity data via a webcam [61]; (2) time and day tracking; (3) a daily task list interface where people enter information about the task type, urgency, difficulty, and anticipated completion time; and (4) hourly self-reports of task progress, overall affect, and feeling of productivity–given–stress.

Data Collection for Model Building

We recruited 30 participants from a large technology company and asked them to install our data collection tool on their desktop computers for four consecutive weeks. We used data from the first 3 weeks to build models and the last week for validation (see Phase 2). 5 participants had incomplete data due to incorrect setup, insufficient self-reports, or taking time off. Our dataset thus comprised of data from 25 people (F=6, M=19) with job roles: Software Engineer (8), Senior Software Engineer (5), Designer (3), Data Scientist (2), Finance Manager (2), Senior Program Manager (1), Senior Content Developer (1), Principal Development Manager (1), Applied ML Engineer (1), and Project Manager (1). Participants were compensated with $150 post-study. They were asked to engage in their regular activities on their computers; the only change to their routine was filling out the Task Tracker at the start of their day, and the hourly productivity and affect self-reports (Figure 1). The Emotion and Activity Logging Software collected data in the background, with informed consent.

Input: Features Categories

We generated a list of 24 input features in eight categories from the collected data (see Figure 1 for a full list of features).

1-Emotion. Classified into eight emotion categories—anger, contempt, disgust, fear, happiness, neutral, sadness, surprise—by the logging software, with probabilities that represent the magnitude of each emotion at a given time, adding up to 1.

2-Heart rate (HR). Prior work shows that a low heart rate and low heart rate variability is reflective of focus [37, 39]. Heart rate variability cannot yet be calculated without using wearable sensors; we used heart rate captured by the logging software to observe if the magnitude accounts for any importance.

3-Physical Movement. This includes eye movement and distance from screen, captured via the logging software.

4-Interaction Data. Also captured by the logging software, this includes mouse and keyboard activity, number of tab switches, and number of open windows.

5-Time, 6-Day. We encoded time (hours at work) and day of week (categorical variable using 7 binary features, one for each day) as two feature proxies for circadian rhythms.

7-Task Information. Includes eight features: task *urgency*; task *difficulty*; anticipated task *completion time*; and task *type* as binary coded values for reading, coding, content creation, digital communication, brainstorming, paper-based reading or writing, creating spreadsheets, and searching for information online. Each feature vector includes this information for tasks that show progress between self-reports at different time intervals. If multiple tasks show progress, task *type* information is a union of the type values, *difficulty* and *anticipated time* are added, and *urgency* is an argmax over the individual values.

8-Potential Actions. At any given time, a person can take one of three actions to change their work environment: (1) *transition* to a different task, (2) take a *break*, or (3) *continue* their current task (i.e., take no action). Breaks may be digital (e.g., visit social media) or physical (e.g., walk away from their computer). Without wearable sensors, we do not have data for what people do during physical breaks; we encode these when there is absence of data. A time sample is considered a digital break if people visit one of the following websites during that time: Facebook, Twitter, LinkedIn, Instagram, Reddit, YouTube, Twitch. It is considered a transition if the foreground windows and tabs being used change completely in that timeframe. All other time samples fall under "continue".

Output: Joint Productivity–Happiness Expected Value

We defined the output of this model as the Joint Productivity–Happiness Expected Value (Expected Value going forward), which considers both productivity and positive affect. Expected Value is computed from the hourly self-report data from FLOWZONE, thus including self-reported task progress and affect, normalized and scaled to be in the range of 0–100.

Prior work shows that people's productivity and affect are correlated. Mark et al. present a framework for this interaction effect using engagement and challenge as axes for attentional states [58]. They classify the quadrants as *rote, focus, bored* and *frustrated* work. For example, people can be happy doing rote work, which may or may not be productive, or people can be focused but also stressed because of a deadline. We scaled people's self-reports of feeling productive (task progress scale) and busy (modified PANAS slider) from -5–5 each, and multiplied these to obtain a value that matches [58]'s quadrants (e.g., low challenge and high engagement means rote work

in [58]; we used productivity 0–5 and feeling busy -5–0 to represent this, which gave the interaction effect a negative value). We normalized and scaled this multiplied value to range from 0–100 as well, and called this *productivity–given–stress*.

For our study, Expected Value was calculated as a cumulative sum that weights each of the three components equally:

$$ExpectedValue = (\tfrac{1}{3} \times TaskProgress) + (\tfrac{1}{3} \times Affect) + (\tfrac{1}{3} \times Productivity|stress)$$

We elaborate on how alternative weightings could be used (e.g., prioritizing a particular component) in our Discussion.

Datasets

Original Dataset. Our original dataset is comprised of data collected over three weeks ($N = 25$). Emotion and workstation activity were logged at a microsecond-level granularity; self-reported data was collected at hourly intervals. Since we did not force people to answer self-reports (to minimize disruptions), there were some hours with missing datapoints. On average, there were 7 hourly self-reports collected per participant, per day (min=4, max=15). Expected Value is dependent on self-reports; thus our original *complete* dataset only had instances for which the self-report value was available—hourly at best. This gave us 62.44 datapoints per participant, on average (min=33, max=122).

Original + Simulated Dataset. Our two data sources—log data and self-reported data—operate at different time intervals (microsecond and hourly, respectively). To better align these and get a complete picture of a user's day, we up-sampled the self-reported hourly data using growth and decay functions, thus getting Expected Value at more granular time intervals. Given a value at hour *h1* and another at *h2* (where $h1 < h2$), we applied a growth function to value at *h1* and a decay function to that at *h2*, and took the max value for every time interval *t* between *h1–h2*. We experimented with several growth and decay functions: exponent with $\gamma = \frac{1}{25}, \frac{1}{5}$, 1, 5, 25; natural log (*Ln*); and Log_{10}. We also tested different time samples for up-sampling: $1, 2, 3, 4, 5, 7, 8, 10, 15, 20, 25, 30$ *mins*. The microsecond-level data obtained from the logging software was similarly down-sampled to the same time samples by applying aggregation functions.

Models

We built models that took as input all of our data sources and predicted an Expected Value. Our data effectively represents a timeseries per person, and our output variable's continuous nature called for regression models. We thus modeled our setup as a classic timeseries forecasting problem using Auto-Regressive Integrated Moving Average (ARIMA) models. ARIMA models have 3 main components: (1) *the Auto-Regressive part*, the number of prior (*lagged*) values of the dependent variable to be used for each new training and prediction datapoint; (2) *the Integrated part*, the degree of differencing required to convert a non-stationary variable into a stationary timeseries; and (3) *the Moving Average part*, the number of random errors of the past to be used to account for current datapoint's errors. ARIMA models traditionally use only one timeseries: the main variable being regressed (here, Expected Value). We build ARIMAX models—ARIMA models with eXogenous variables—to account for input features (e.g., emotion labels, task information) which are potential explanatory variables, called exogenous in ARIMA terminology.

While ARIMAX models are the best representation of our timeseries data, they are complex and thus expensive to compute. With our deployment goal for Phase 2, we also modeled our data using other regression models. We tried several different ones (e.g., Support Vector Regression–SVR and Multiple Linear Regression–MLR), and finally picked Random Forest Regression (RFR) models for our real-time recommendation task because these had the best performance (metrics for gauging performance are explained below) and assigned feature importances similar to ARIMAX. We relied on this similarity of feature importances between the two types of models as a form of validation for using the less computationally demanding RFR models for deployment in Phase 2.

Cross-validation. When using the original dataset, we applied leave-one-out cross-validation (LOOCV), training on n-1 datapoints and testing on 1, and averaging the results of all possible model combinations done this way. For the original + simulated dataset, we used holdout cross-validation, using 60% of data for training, and 20% each for validation and testing. For both these methods of cross-validation, we followed day forward chaining to ensure that future values are never used to predict past values. That is, for each day, we treated each future datapoint as a new test case and used all prior ones as our training set. Similarly for the train–validation–test dataset split, we used ordered splitting such that no future datapoints were in the training or validation sets.

Metrics. For the ARIMAX models, we used Akaike Information Criterion (AIC) values to find the best-fitting model. AIC values are used for timeseries models because they represent goodness-of-fit for past and future of the timeseries data; lower AIC values indicate better fit. We used R^2 and Adjusted-R^2 to evaluate the goodness-of-fit of our RFR models. These metrics are used to report how well the selected independent features explain the variability of our dependent variable (Expected Value). For example, an R^2 value of 0.X is read as "the model explains X% of variance in the data." R^2 values can be biased to the addition of new features, even when the features do not add any explanatory power. Adj-R^2 handles this bias, and thus is a better measure for model comparison. We report both for our RFR models, but pick the best models using the Adj-R^2 values. We also computed Root Mean Square Error (RMSE) values for both types of models to quantify the difference between actual and predicted Expected Values.

PHASE 1: MODEL EVALUATION

We used emotion, workstation activity, and task data to model our output variable—the Expected Value of people's workday (which is designed to jointly capture their productivity and affect). We tested several regression models on our original and original + simulated datasets. We ultimately relied on the original + simulated dataset for all our model-building after validating this dataset against the original dataset: there is no significant difference ($p > 0.1$) in model performance or feature importances between the two datasets.

	AR,I,MA	Feature (Estimate)	AIC
P1	(2,0,4)	Time (1.3)***, Tab Switches(-0.7)***, Continue(-0.4)**	2650
P2	(5,1,2)	Complete Time (-0.7)***, Mon (-0.4)**, Transition (0.6)**	2218
P3	(1,0,4)	Screen Distance (1.5)**, Surprise (-0.3)**, Continue(-0.1)*	4902
P4	(3,1,5)	Keyboard (1.07)***, Coding (0.26)**, Continue (0.72)***	2291
P5	(2,1,5)	Keyboard (0.89)***, Surprise (-1.3)***, Break (1.22)**	1880
P6	(5,2,3)	Screen Distance (1.1)**, Continue (0.35)**	3105
P7	(3,1,4)	Tab Switches (-1.33)**, HR (-0.13)*, Continue (-1.07)**	1284
P8	(4,0,4)	HR (-1.78)**, Transition (-1.2)**	4171
P9	(1,1,5)	Mouse (0.94)***, Anger (-0.42)**, Break (1.05)**	1255
P10	(2,0,3)	Happiness (0.8)***, Urgency (-1.23)***, Continue (-0.28)*	2260
P11	(5,0,4)	Mouse (0.63)**, Continue (0.35)**	1315
P12	(5,1,4)	Eye Movement (0.23)*, Wed (1.53)**, Surprise (0.18)*	3147
P13	(1,0,5)	Sadness (0.77)**, Surprise (0.14)*, Transition (-0.9)*	3357
P14	(3,2,4)	Time (-1.6)***, Mon (-0.66)**, Continue (-0.83)**	2469
P15	(1,1,5)	Keyboard (-0.33)*, Difficulty (-0.92)**	2195
P16	(2,1,5)	Screen Distance (0.72)**, Break (0.59)**	1109
P17	(1,0,5)	Brainstorming (0.31)*, Continue (0.41)**	3692
P18	(4,0,3)	Tab Switches (-0.39)***, Screen Distance (-0.62)**	2774
P19	(2,1,4)	Sadness (0.25)**, Reading (-0.51)*, Break (1.51)**	2053
P20	(2,1,3)	Coding (0.79)***, Continue (1.27)**	1529
P21	(1,0,5)	Urgency (-1.25)**, Monday (-0.68)**, Break (1.04)***	1893
P22	(2,1,5)	Keyboard (0.93)**, Difficulty (0.17)**, Continue (0.65)*	3041
P23	(3,2,4)	Writing (-1.64)***, Break (0.49)**, Continue (0.11)*	2063
P24	(2,1,5)	Tab Switches (0.16)*, Reading (0.79)**, Transition (1.14)**	1148
P25	(4,0,4)	Complete Time (-1.29)***, Tues (-0.09)*, Anger (-0.17)*	1547

Table 1. Results from the ARIMAX models per participant: AR, I, MA denote values for the auto-regressive, integrated, and moving average components of the model. Features presented are those with significant estimate values, and AIC values represent goodness-of-fit. Significance levels: *=p<0.05 **=p<0.01 ***=p<0.001

ARIMAX Model Performance

ARIMAX models are commonly applied to timeseries data like ours. Since each timeseries is unique to the context it was collected in, we treated all participants' data separately, and built personalized ARIMAX models for all of them. The core AR, I, MA features of an ARIMAX model rely on this unique context per timeseries (see AR, I, MA values in Table 1). ARIMAX models output results in the form of estimates for each independent variable along with p-values for significance.

We find that ARIMAX models output 2-3 significant features per participant. To better understand the broader categories of features that are important, we binned our 24 features into the 8 categories in our setup, each representing a different data source (Figure 1): Emotion, Heart Rate (HR), Physical Movement, Interaction Data, Time of Day, Day of Week, Task Information, and Action. Noting the categories with at least one significant feature per participant, Action is the most popular category (21 out of 25 participants show at least one of break, transition, or continue as having a significant estimate), followed by Task Information (12 out of 25), Interaction Data (10 out of 25), Emotion (8 out of 25), Physical Movement (5 out of 25), Day of Week (5 out of 25), Time of Day (2 out of 25), and Heart Rate (2 out of 25).

ARIMAX models consistently return significant estimates for a feature in the *Action* class: whether someone has recently taken a break, transitioned tasks, or has been continuing the same task, is important for predicting future actions. Timeseries models are well-known for capturing such historical nuance. We find that samples aggregated at 7- and 10-minutes (time sample variable t used in Original + Simulated dataset) provide the best results for these models, with average AR and MA values being 3 and 4, respectively. This means that the ARIMAX models consider the past 21-30 minutes (3x7 and 3x10) of data in forecasting the Expected Value for a given time interval, and do this with an average RMSE of 8.6% and AIC value of 2374. The validation split highlights $Exp(\frac{-1}{25}x)$ as the time decay function for the best performing model.

Random Forest Regression Model Performance

ARIMAX models are complex and computationally demanding (processing time of $\sim$15 mins per participant), making it hard to use them in real-time settings. We tested other regression models (e.g., SVR, MLR), settling on Random Forest Regression (RFR) models because they have the best performance. Since ARIMAX models are more naturally suited to our timeseries data, we relied on the results of the ARIMAX models to validate the performance of the RFR models.

We built RFR predictive models at three levels: general, per participant, and per cluster, where clustering is done based on job role. A general model with good performance has the potential of being applied at a larger scale, because it indicates that people's data can be used interchangeably. Personalized models per participant with good performance can help us understand which features matter most when modeling different individuals. Models for different job role clusters can highlight whether people's work practices, productivity, and affect are defined by something specific about their job role.

Table 2 presents results for all RFR models using the metrics explained above. It also includes the distribution of data (mean and S.D.) for each participant and cluster, to better contextualize our R^2 and Adj-R^2 results. Further, Table 2 highlights the best values of the constants used for modeling via the holdout validation set. All models with the best validation set performance use $Exp(\frac{-1}{25}x)$ as their time decay function; the best values for Time Sample per model are indicated in Table 2. Below, we share results from each of these models, and then compare the feature importances seen across them.

General Model Performance

Given prior work that suggests that people have unique patterns of affect, activity, and daily to-dos at work, it comes as no surprise that our general model that includes all participants as one data source has mediocre performance. With an R^2 and Adj-R^2 value of 0.2, the general model is able to explain 20% variance in data, making it a moderate fit. The Root Mean Squared Error (RMSE) for this model is 26.5, on a scale of 0–100; RMSE values share the same scale as the output variable, Expected Value (Table 2, header "All").

Personalized Model Performance

Our personalized models have high R^2 and Adj-R^2, especially when considering the wide distributions of data per participant. R^2 values range from $0.2-0.7$, with an average of 0.52, and Adj-R^2 values range from $0.2-0.7$, with an average of 0.47 (Table 2, header "Participants"). High values for both these metrics indicate that our models are a good fit for people's data, and a large percentage (up to 70% in the best case) of the variance in data is explained by the models. The RMSE values range from 3.5 to 13.2, the average value being 7.1. Overall, these models perform extremely well both in terms of goodness-of-fit and low error values.

	Participants																								Clusters					All
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	1	2	3	4	5	
Job Role	Software Engineers								Senior Software Engineers					Designers			Data Scientists		Finance Managers												
Mean	26.2	39.7	35.9	21	29.4	19.9	33	31.8	33.8	33.8	27.5	26	29.1	29	30.5	30.7	35.4	23.8	25.7	22.6	23.3	32.9	33.6	35.6	17.6	29.6	30	30	29.6	24.2	29.1
S.D.	23.6	22.5	31.4	19.9	25.5	12.3	27.8	25.5	22.2	31.4	24.8	25.5	18.2	22.5	20	19.5	21.9	20.9	19.3	16.5	19	27.5	25.8	24.3	13.8	23.5	24.4	20.7	21.4	17.9	22.4
RMSE	4.7	4.9	5.4	3.6	5.6	10.3	7.6	5.6	5.2	4.6	7.6	9.7	9	13.2	6.1	6.5	3.5	9.6	7	10	11.7	5.4	8.3	6.4	5.6	5.2	5.1	5.1	4.7	3.9	26.5
R^2	0.4	0.7	0.4	0.5	0.2	0.6	0.6	0.6	0.7	0.4	0.3	0.3	0.6	0.5	0.6	0.6	0.7	0.2	0.3	0.6	0.7	0.6	0.6	0.5	0.7	0.7	0.6	0.7	0.4	0.5	0.2
Adj R^2	0.4	0.7	0.3	0.4	0.2	0.6	0.5	0.6	0.6	0.3	0.3	0.2	0.5	0.4	0.6	0.4	0.7	0.2	0.3	0.5	0.7	0.6	0.6	0.5	0.6	0.7	0.5	0.7	0.3	0.5	0.2
Time Sample	7min	2min	4min	7min	1min	5min	4min	8min	3min	7min	4min	1min	8min	1min	3min	1min	3min	1min	1min	8min	5min	2min	1min	7min	4min	2min	2min	1min	4min	2min	2min

Table 2. Results of Random Forest Regression models for all individual Participants (P1–25) and Clusters (C1–5), and a Generalized Model for "All" Participants. Participants are color-coordinated according to their cluster membership. E.g., P1–8 belong to cluster C1.

Cluster Model Performance

Our cluster models have similar performance to the personalized models, with R^2 values ranging from $0.4 - 0.7$, and RMSE values between $3.9 - 5.2$. In fact, in some cases, these models perform better than the personalized models for the participants in the cluster. Since the clusters are formed based on job role, this suggests that people doing similar jobs have similar task progress, affect, and productivity-given-stress rates. In a cold-start setting—when we do not yet have enough data from a participant to build personalized models for them immediately—modeling based on data from their job role cluster could be a viable alternative. The clusters we chose here were based on the official job roles of our participants: Software Engineer, Senior Software Engineer, Designer, Data Scientist, Finance Manager, Other (which included Senior Program Manager, Senior Content Developer, Principal Development Manager, Applied ML Engineer, and Project Manager).

Understanding Feature Importance

We binned our 24 features into the same 8 categories to understand the importance of each category. The feature importances of all categories sum up to 1; the maximum importance value assigned to any individual category is 0.60 (Figure 2). [2]

We find interaction data to be the most important feature category on average, followed by task information, emotion, physical movement, time, heart rate, day, and potential actions. The average feature importances for different categories across all participants were: interaction data=0.22, task information=0.19, emotion=0.17, physical movement=0.14, time=0.13, heart rate=0.08, day=0.07, and potential action=0.02. Even though interaction data is the most important feature category on average, it is not the most important feature for each participant. For example, emotion is the most important category for P13, task information for P2 and P20, and combinations of other categories are equally important for other participants. The order of feature importance remains the same if we look at the frequency at which each feature is most important.

Interaction data—the most important feature category for personalized models, on average—is not as important for the clusters or general model (Figure 2, Clusters start with "C" and general model under "All"). On average, the feature importances per cluster are not aligned with their members' personalized models, and instead have job-based patterns: task information and interaction data are the most important categories for C1 (Software Engineers) and C2 (Senior Software Engineers), whereas time of day is crucial for C5 (Finance Managers). The important categories intuitively match the requirements of the job role (e.g., finance managers might have less collaborative roles than software engineers, keeping them at their desks and following a 9am–5pm day). The general model's feature importances are more spread out across all feature categories, as expected in an aggregated model.

Overall, we find that different features are important for personalized, cluster, and general models. This is interesting given that the R^2, Adj-R^2, and RMSE values are not too different across these, especially for personalized and cluster models. Indeed, it seems that an important consideration when applying these models in a real-world setting is the eventual need for personalized models. While starting with cluster-based models might rid one of the cold-start problem, no general or cluster model represents the participant and what is important for their Expected Value in the same way as their own data.

[2]Expanded version of Figure 2 with all 24 features is included in supplementary material.

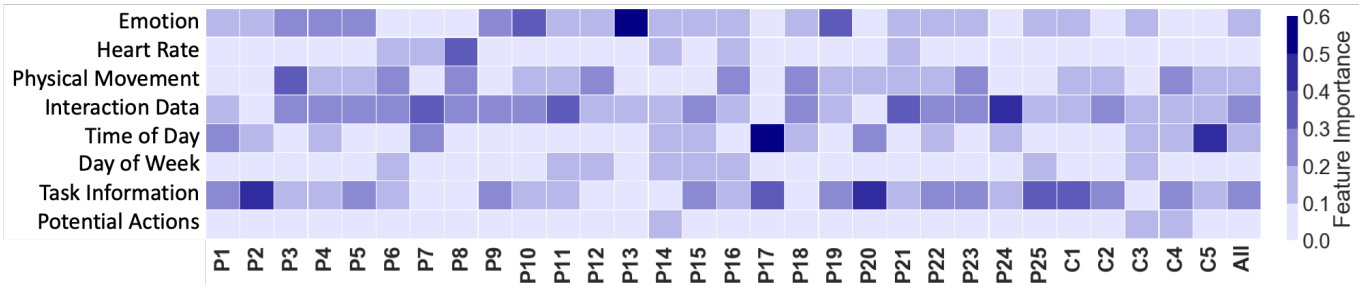

Figure 2. Feature importance output from the random forest regression models for all participants, clusters, and the general model.

Comparing ARIMAX and RFR Models

At least one of the features with significant estimates in the ARIMAX models also consistently belonged to an important feature category in the RFR results. For example, task information is the most important feature class for P2, and anticipated completion time (a feature that falls under the task information category) has a significant estimate from the ARIMAX model for P2. The primary difference between the two models is in the *Action* class: ARIMAX models consistently return significant estimates for a feature in the Action class whereas RFR models do not. We hypothesize that this is due to the nature of the action variable: whether someone has recently taken a break, transitioned tasks, or has been continuing the same task becomes a more important consideration over time. Timeseries models capture exactly this nuance, whereas RFR models do not consider these historical values.

Performance Tradeoffs. We used RFR models for deployment; while we tested ARIMAX models in this setting, the high processing time for ARIMAX ($\sim$15 mins per participant compared to $\sim$2 mins for RFR models) made it infeasible to use them in a real-time context. Even though ARIMAX models better capture the nuances of past actions in predicting future ones, we made this tradeoff because all other feature categories have similar importances in both ARIMAX and RFR models. More nuanced engineering efforts could reduce processing times to make ARIMAX models also work in real-time settings—we leave these explorations to future work.

PHASE 2: MODEL DEPLOYMENT

In Phase 2, we built a system that uses the models from Phase 1 to recommend transitions and breaks in real time. We deployed this system to observe how people perceive the timing of our recommendations—whether we were able to find opportune moments for transitions and breaks.

FLOWZONE v2: Real-time Recommendations

We added frontend and backend components to FLOWZONE to recommend transitions and breaks in real time.

Frontend Modifications. We added two Windows forms that appear for transition and break recommendations; nothing appears for "continue". Each form showed the recommendation along with an explanation (e.g., for breaks, it said "*Wow, you've been working hard! FlowZone thinks a break right now will replenish your energy and keep you going!*"). We did not provide personalized explanations for the recommendations. The forms asked participants to select one of the following options for each recommendation: (1) "Yes, going to take a break," (2) "Yes, it's time for a break, but I can't take one right away," (3) "Yes I just took a break," and (4) "No, this is not a good time for a break." The granularity in "Yes" options supports our goal of observing whether the recommendation is an interruption or comes at an opportune time.

Model-based Backend. Our backend enabled real-time queries to the logging databases and the models built in Phase 1. We hosted a webserver that interacted with these components using API calls; the logging databases were hosted on Azure Table Service, and the model files were hosted on our webserver after being converted to a compressed format. Our backend pipeline was: (1) logging software stored data every microsecond (as before); (2) for each participant's chosen time sample t (i.e., the Time Sample parameter, in minutes, with the best performance in Phase 1), FLOWZONE pinged the server to get the last t minutes of data; (3) the backend computed three feature vectors by aggregating t minutes of data and adding a binary encoding for transition, break, and continue to each of these vectors, respectively; (4) the backend computed an argmax over the output Expected Value for the three vectors, one each for transition, break, and continue; (5) the potential action with the maximum Expected Value was returned as a recommendation to the frontend, where it was shown to the participant with the corresponding form.

Study Design

We deployed our updated FLOWZONE app for three days during the fourth week of our study (model condition, *M* going forward). To ensure that our participants' responses about the recommendations were not biased by system novelty, we added a control condition (*C* going forward) which used the same system setup and outputs, but relied on pseudo-random, heuristics-based rules for recommending transitions and breaks. Our goal was not to compare the two conditions; rather, to validate that people were rating the timing, and not rating favorably because of the novelty of the system.

For Condition *M*, participants received recommendations for transitions or breaks using the predictive models built in Phase 1. For Condition *C*, we did not use models; we assigned heuristics-based probabilities to the potential actions: transition and break were assigned $\frac{1}{6}$th probability each, and continue was assigned $\frac{2}{3}$rd probability because continuing a task is more common than task transitions or breaks. At every 30-minute interval, the Condition *C* recommender picked one out of the three options based on the probabilities assigned, and recommended that to the participant. We set recommendation checks at 30-minute intervals for Condition *C* because this is traditionally the smallest time interval on people's work calendars. Both *M* and *C* condition participants were shown the same interface and explanations.

Post-Study Survey. All participants took a post-study survey that included: (1) open-text questions about people's opinion of FLOWZONE—whether the transition and break recommendations were well-timed or not, appropriately frequent or not, examples of cases of good and bad recommendations (and why), and if they felt better after following a recommendation than the state they were in before; and (2) two Likert questions on whether FLOWZONE made them feel more productive and happy at work (range: strongly disagree–strongly agree, 1–5).

The survey also included questions about the idea of intelligent systems guiding people at work to jointly optimize their happiness and productivity. We asked an open-text question on what people thought would be good or bad about this idea, and 4 Likert questions on whether they thought this tool would: (1) be useful for their work practices, (2) make them feel positive at / about work, (3) make them feel negative at / about work, and (4) be helpful for their productivity at work (all ranged: strongly disagree–strongly agree, 1–5).

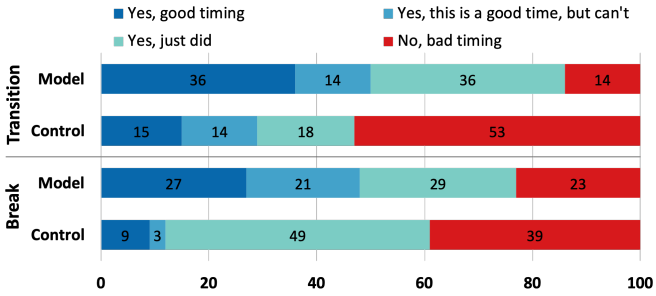

Figure 3. The percentage of times transition and break recommendations fell under each of the categories provided to participants.

Data Collection

We set up the study with 15 participants in Condition *M* and 10 in Condition *C*. The 10 participants for Condition *C* had relatively low data volume in Phase 1—there were some gaps in their data due to frequent meetings away from their desk (the tool was recording data only at their desk), remote work days, or unexpected vacation time. The Adj-R^2 values for their models ranged from $0.2 - 0.4$ (moderate to low variance in data explained by the model). We worried that these participants may not be naturally receptive towards recommendations given their low response rate during data collection. We gave them the option to opt out before Phase 2, but they wanted to continue and assured us that the low data collection was due to external factors, not our setup.

We designed Condition *M* to represent the best case modeling we could do, and Condition *C* to counteract any novelty effects from the FLOWZONE. Participants were not aware of any system differences, and their data was only included in the Phase 2 results if they stayed active at their workstations, per our request. 11 out of 15 people in Condition *M*, and eight out of 10 in Condition *C* continued to use FLOWZONE with the updated real-time recommendations. Since this was the fourth week of a field study, the dropouts were not surprising.

Deployment Results

Participants in Condition *M* received on average 1.2 recommendations for transitions and 5.6 for breaks (s.d. transition=0.56, break=1.6), and those in Condition *C* received 3.5 transition and 4 break recommendations every day (s.d. transition=0.72, break=2.7). Thus, participants in Condition *M* saw relatively more recommendations for breaks over transitions, whereas the number was even in Condition *C*. This was not unexpected given Condition *C* setup—equal probabilities for transitions and breaks. The large difference in the number of transitions for Condition *M* vs. *C* shows that our models have a nuanced understanding of the number of tasks performed in a day and when people should transition between them.

Condition *M* recommendations had a high chance of being at opportune times. We calculated accuracy via summation of responses for all "Yes" categories divided by the total number of responses, per condition. Using this descriptive metric, we find model-based recommendations of transitions and breaks to be 85.7% and 77% accurate, respectively (Figure 3). In comparison, Condition *C* transitions and breaks were only 47% and 60.6% accurate, respectively, indicating that people's evaluation in favor of Condition *M* was due to different, more opportune-timed recommendations, rather than a novelty bias.

When asked about whether FLOWZONE made them more productive at work, Condition *M* participants leaned positive (Agree=6, Neutral=2, Disagree=4), whereas Condition *C* participants had mostly neutral responses (A=2, N=4, D=1). When asked the same question in the context of happiness at work, Condition *M* participants were positive (Strongly Agree=1, A=6, N=2, D=3), and those in Condition *C* remained neutral or negative (N=5, D=2).

When asked to rate if this future intelligent system would be useful at work, most people responded positively (SA=1, A=12, N=4, D=2). They similarly had a positive response for whether this tool could help them feel positive about work (SA=2, A=10, N=6, D=1) and help their productivity (A=12, N=4, D=3). Most people were appreciative of the idea behind task tracking, productivity, and positive affect at work, and were excited about tools like this becoming commonplace in the future: "*it can be a digital assistant looking after you and your well-being, what more could you want?!*" (P13).

DISCUSSION AND FUTURE WORK

We have shown that it is possible to build models that jointly optimize happiness and productivity at work, using emotion, workstation activity, and task-based data (R^2 $0.2 - 0.7$; RMSE $<15\%$). When deployed, these models allowed us to recommend transitions and breaks to people at opportune times (people evaluated the accuracy of the timing as: transitions=85.7%, breaks=77%). Below, we discuss several design implications and considerations that resulted from our studies.

Frequency and Timing of Recommendations is Crucial

We began our exploration of recommending transitions and breaks with the intuition that finding opportune times is important to avoid disrupting people's focused work times. Our participants highlighted the same need for well-timed recommendations, making it an important design consideration for systems in this space. Participants in our model condition felt that the frequency of recommendations was "*just right*" (P3) or "*frequent, but good for my health*" whereas those in the control condition felt that the frequency "*very rarely matched my own assessment*" (P11). In addition to frequency, the timing of recommendations is important for people's decision to follow through: "*I found the timing to be surprisingly good. Following the recommendations did help me feel happier and more productive because I took more breaks that I realized after the fact that I needed. Hours turn to blurs without something to break them up so taking breaks helped the day seem more full*" (P4). When the timing was not right, the pop-up served as an interruption. Critically, the notion of "opportune" timing is personalized—future work should consider further analysis of factors that make a moment "opportune" for an individual.

Intervention Design Needs Personalization

FLOWZONE was meant to study the *timing* of recommendations, but looking ahead, this is simply a starting point for designing interventions that might help people follow through a recommended action. These intervention-style applications that might apply our recommender need to design personalized strategies. Indeed, prior work also suggests that people's definitions of breaks and transitions are subjective and task-dependent [11, 24]. We built personalized models to support

this subjectivity; future work can extend these models to be suited to particular task settings, individuals' moods, personality traits, or workplace behavior trends. Our participants agreed, and mentioned some preferences for what a break recommendation could look like: "*if you can provide a joke instead of asking me to take a break, or sending me analytics about what other employees are doing at this moment, or how many people are suffering at the same problem I have may help me feel better*" (P12). Prior work has evaluated other forms of personalization via gamified, guided breaks [11], visualizations and analytics about workdays and break-time behavior [24], or conversational agents [76]—more opportunities exist for combining these complementary approaches.

Keeping Control with the User

People in our deployment study were generally positive about an intelligent system that guided them towards happy and productive states, but they wanted control as needed. Knowledge of deadlines, meetings, and other external factors affect people's workday. Unless the tool let them manipulate these meta-level factors, some people felt that an intelligent system could not guide their workflow effectively. Technology can never completely meet the fluid social needs of users. Ackerman calls this the "social-technical gap" and suggests that instead of attempting to build the impossible perfect solutions, we should build first order approximations of them [1]. As such, we added opportunities for user-control in our models via the design of the output variable which assigns weights to productivity, affect, and their interaction variable. We use equal weights, but this could be user-controlled, as seen in related recommender systems work [36]. Understanding how parameter weights—user-controlled or learned by the system—impact system utility is an interesting avenue of future work.

Understanding Context in the Workplace

People rarely work in isolation—they are a part of teams within organizations, often collaborating on a daily basis [13]. Several participants mentioned a desire for a team-centric version of FLOWZONE. This requires context about how a team works together: group coherence, communication, and reliance become important. It is not as simple as jointly optimizing happiness and productivity for each individual team member—when people rely on each other in a team setting, their productivity and happiness are dependent on that of other members. While we wait for technical advances that can enable an understanding of team context, we apply cluster-based data aggregation as a starting point. We were able to achieve reasonable goodness-of-fit using clusters (R^2 values ranged from $0.4 - 0.7$), but the feature importances that were unique to an individual were lost. Future work should consider more nuanced clusters by conducting studies to surface the tacit roles people perform under the umbrella of an official job title.

Ethical Considerations

While emotion, workstation activity, and task data logging can help build accurate models for happiness and productivity, there are concerns about worker privacy, both from us and our participants: *"feeling like you're being watched all the time would just be bad"* (P3). This is an important consideration, as prior work (e.g., [29]) cautions us of the privacy breaches that are impossible to manage once tracking becomes a required or coerced aspect of work. Beyond privacy, building tools for productivity and efficiency is often seen as supporting Taylorism, where employees' effort is optimized for the most output, with no consideration of the individuals [53]. Our efforts oppose this, instead aiming to keep employees happy while completing fulfilling work. We believe in the "happy-productive worker"—being happy at work is what causes people to be more productive [82]—thus our focus is to optimize happiness, while recognizing that getting things done is also necessary.

LIMITATIONS

One limitation of our study is that, similar to prior work (e.g., [84]), our setup relies on hourly self-reports of productivity and affect. Even though we make it easy to dismiss these reporting forms, the hourly requests can be a form of interruption. However, note that these reports are only for model building purposes. Once the model is deployed, these would only be required periodically for model updates. Another limitation is that we validated our approaches via a deployment study lasting three days. Since users can take time to adapt to suggestions and integrate them in their work patterns, longitudinal studies of such recommendations may provide additional insights. Further, we evaluated our models against a simple, heuristics-based control, to account for any novelty bias. In future work, we hope to compare our results to other existing approaches (e.g., the Pomodoro approach [14]) in a longitudinal study. Finally, our models use simulated datasets (in combination with real user data) to enable complex modeling techniques such as timeseries forecasting. While simulated datasets are commonplace in other domains (e.g., natural language processing [48]), they are new to domains like workplace recommendations. We validated the integrity of our original + simulated dataset via comparison tests with the original dataset, but hope that future work will consider other ways to acquire and validate these simulated datasets.

CONCLUSION

We explore how affect, workstation activity, and task data can be used to develop predictive models for recommending task transitions or breaks, with the goal of guiding information workers towards more productive, happy work. We find these models to be highly personalized, though some commonalities exist across the same job roles. Validation of our models with real-time recommendations shows 86% accuracy in predicting opportune moments for transitions and 77% accuracy for breaks. While open research questions remain around how to support users in following through with the recommendations and how to support team / collaborative settings, our work is a crucial first step towards building intelligent systems that consider both: people's happiness and their productivity.

ACKNOWLEDGMENTS

We would like to thank our reviewers for their thoughtful feedback. We are also grateful to Paul Bennett, Kael Rowan, Shane Williams, and the IDEAS team at Microsoft Research for their support and feedback. This work was initiated while the first two authors were interns at Microsoft.

REFERENCES

[1] Mark S. Ackerman. 2000. The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility. *Human-Computer Interaction* 15, 2–3 (2000), 179–203. DOI: http://dx.doi.org/10.1207/S15327051HCI1523_5

[2] Rachel F. Adler and Raquel Benbunan-Fich. 2013. Self-interruptions in discretionary multitasking. *Computers in Human Behavior* 29, 4 (2013), 1441–1449. DOI:http://dx.doi.org/10.1016/j.chb.2013.01.040

[3] Ane Alberdi, Asier Aztiria, and Adrian Basarab. 2016. Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review. *Journal of Biomedical Informatics* 59 (2016), 49–75. DOI: http://dx.doi.org/10.1016/j.jbi.2015.11.007

[4] Brian P. Bailey and Shamsi T. Iqbal. 2008. Understanding Changes in Mental Workload during Execution of Goal-Directed Tasks and Its Application for Interruption Management. *ACM Trans. Comput.-Hum. Interact.* 14, 4, Article 21 (Jan 2008), 28 pages. DOI:http://dx.doi.org/10.1145/1314683.1314689

[5] Brian P. Bailey and Joseph A. Konstan. 2006. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior* 22, 4 (2006), 685–708. DOI: http://dx.doi.org/10.1016/j.chb.2005.12.009

[6] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest* 20, 1 (2019), 1–68. DOI: http://dx.doi.org/10.1177/1529100619832930

[7] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI '16)*. Association for Computing Machinery, New York, NY, USA, 279–283. DOI: http://dx.doi.org/10.1145/2993148.2993165

[8] James "Bo" Begole, Nicholas E. Matsakis, and John C. Tang. 2004. Lilsys: Sensing Unavailability. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work (CSCW '04)*. Association for Computing Machinery, New York, NY, USA, 511–514. DOI: http://dx.doi.org/10.1145/1031607.1031691

[9] Raquel Benbunan-Fich and Gregory E. Truman. 2009. Multitasking with Laptops during Meetings. *Commun. ACM* 52, 2 (Feb. 2009), 139–141. DOI: http://dx.doi.org/10.1145/1461928.1461963

[10] Derek S Burkland. 2013. *The effects of taking a short break: task difficulty, need for recovery and task performance*. Ph.D. Dissertation. University of Wisconsin–Stout.

[11] Scott A. Cambo, Daniel Avrahami, and Matthew L. Lee. 2017. BreakSense: Combining Physiological and Location Sensing to Promote Mobility During Work-Breaks. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 3595–3607. DOI:http://dx.doi.org/10.1145/3025453.3026021

[12] Daniel Chen, Jamie Hart, and Roel Vertegaal. 2007. Towards a Physiological Model of User Interruptability. In *Human-Computer Interaction – INTERACT 2007*, Cécilia Baranauskas, Philippe Palanque, Julio Abascal, and Simone Diniz Junqueira Barbosa (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 439–451.

[13] Jan Chong and Rosanne Siino. 2006. Interruptions on Software Teams: A Comparison of Paired and Solo Programmers. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work (CSCW '06)*. ACM, New York, NY, USA, 29–38. DOI: http://dx.doi.org/10.1145/1180875.1180882

[14] Francesco Cirillo. 2006. The pomodoro technique (the pomodoro). *Agile Processes in Software Engineering and* 54, 2 (2006).

[15] Russell Cropanzano and Thomas A Wright. 2001. When a "happy" worker is really a "productive" worker: A review and further refinement of the happy-productive worker thesis. *Consulting Psychology Journal: Practice and Research* 53, 3 (2001), 182–199. DOI: http://dx.doi.org/10.1037/1061-4087.53.3.182

[16] Mihaly Csikszentmihalyi. 1997. *Finding flow: The psychology of engagement with everyday life.* Basic Books.

[17] Edward B. Cutrell, Mary Czerwinski, and Eric Horvitz. 2000. Effects of Instant Messaging Interruptions on Computing Tasks. In *CHI '00 Extended Abstracts on Human Factors in Computing Systems (CHI EA '00)*. ACM, New York, NY, USA, 99–100. DOI: http://dx.doi.org/10.1145/633292.633351

[18] Mary Czerwinski, Eric Horvitz, and Susan Wilhite. 2004. A Diary Study of Task Switching and Interruptions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 175–182. DOI: http://dx.doi.org/10.1145/985692.985715

[19] Awwad J. Dababneh, Naomi Swanson, and Richard L. Shell. 2001. Impact of added rest breaks on the productivity and well being of workers. *Ergonomics* 44, 2 (2001), 164–174. DOI: http://dx.doi.org/10.1080/00140130121538

[20] Laura Dabbish, Gloria Mark, and Víctor M. González. 2011. Why Do i Keep Interrupting Myself? Environment, Habit and Self-Interruption. In *Proceedings of the SIGCHI Conference on Human*

Factors in Computing Systems (CHI '11). Association for Computing Machinery, New York, NY, USA, 3127–3130. DOI: http://dx.doi.org/10.1145/1978942.1979405

[21] Giuseppe Destefanis, Marco Ortu, Steve Counsell, Stephen Swift, Michele Marchesi, and Roberto Tonelli. 2016. Software development: do good manners matter? *PeerJ Computer Science* 2 (July 2016), e73. DOI: http://dx.doi.org/10.7717/peerj-cs.73

[22] Belinda H.W. Eijckelhof, Maaike A. Huysmans, Birgitte M. Blatter, Priscilla C. Leider, Peter W. Johnson, and Jaap H. van Die}. 2014. Office workers' computer use patterns are associated with workplace stressors. *Applied Ergonomics* 45, 6 (2014), 1660–1667. DOI: http://dx.doi.org/10.1016/j.apergo.2014.05.013

[23] Paul Ekman, E Richard Sorenson, and Wallace V Friesen. 1969. Pan-cultural elements in facial displays of emotion. *Science* 164, 3875 (1969), 86–88. DOI: http://dx.doi.org/10.1126/science.164.3875.86

[24] 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 (CHI '16)*. ACM, New York, NY, USA, 673–684. DOI: http://dx.doi.org/10.1145/2858036.2858066

[25] James Fogarty, Scott E. Hudson, Christopher G. Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny C. Lee, and Jie Yang. 2005. Predicting human interruptibility with sensors. *ACM Trans. Comput.-Hum. Interact.* 12, 1 (March 2005), 119âĂŞ146. DOI: http://dx.doi.org/10.1145/1057237.1057243

[26] Pamela S Galluch, Varun Grover, and Jason Bennett Thatcher. 2015. Interrupting the workplace: Examining stressors in an information technology context. *Journal of the Association for Information Systems* 16, 1 (2015), 1. DOI:http://dx.doi.org/10.17705/1jais.00387

[27] Malcolm Gladwell. 2002. The naked face. *The New Yorker* 5 (2002), 38–49.

[28] Victor M. González and Gloria Mark. 2004. "Constant, Constant, Multi-Tasking Craziness": Managing Multiple Working Spheres. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. Association for Computing Machinery, New York, NY, USA, 113–120. DOI: http://dx.doi.org/10.1145/985692.985707

[29] Nanna Gorm and Irina Shklovski. 2016. Sharing Steps in the Workplace: Changing Privacy Concerns Over Time. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4315–4319. DOI: http://dx.doi.org/10.1145/2858036.2858352

[30] Daniel Graziotin, Fabian Fagerholm, Xiaofeng Wang, and Pekka Abrahamsson. 2017. On the Unhappiness of Software Developers. In *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering (EASE'17)*. Association for Computing Machinery, New York, NY, USA, 324–333. DOI:http://dx.doi.org/10.1145/3084226.3084242

[31] Daniel Graziotin, Xiaofeng Wang, and Pekka Abrahamsson. 2014. Happy software developers solve problems better: psychological measurements in empirical software engineering. *PeerJ* 2 (2014), e289. DOI:http://dx.doi.org/10.7717/peerj.289

[32] James J Gross. 1998. The emerging field of emotion regulation: An integrative review. *Review of general psychology* 2, 3 (1998), 271–299. DOI: http://dx.doi.org/10.1037/1089-2680.2.3.271

[33] Barry Gruenberg. 1980. The happy worker: An analysis of educational and occupational differences in determinants of job satisfaction. *American journal of sociology* 86, 2 (1980), 247–271. DOI: http://dx.doi.org/10.1086/227238

[34] Fangfang Guo, Yu Li, Mohan S. Kankanhalli, and Michael S. Brown. 2013. An Evaluation of Wearable Activity Monitoring Devices. In *Proceedings of the 1st ACM International Workshop on Personal Data Meets Distributed Multimedia (PDM '13)*. Association for Computing Machinery, New York, NY, USA, 31–34. DOI:http://dx.doi.org/10.1145/2509352.2512882

[35] Eija Haapalainen, SeungJun Kim, Jodi F. Forlizzi, and Anind K. Dey. 2010. Psycho-Physiological Measures for Assessing Cognitive Load. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp '10)*. Association for Computing Machinery, New York, NY, USA, 301–310. DOI: http://dx.doi.org/10.1145/1864349.1864395

[36] F. Maxwell Harper, Funing Xu, Harmanpreet Kaur, Kyle Condiff, Shuo Chang, and Loren Terveen. 2015. Putting Users in Control of Their Recommendations. In *Proceedings of the 9th ACM Conference on Recommender Systems (RecSys '15)*. Association for Computing Machinery, New York, NY, USA, 3–10. DOI: http://dx.doi.org/10.1145/2792838.2800179

[37] Jennifer Healey, Rosalind W Picard, and others. 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems* 6, 2 (2005), 156–166. DOI: http://dx.doi.org/10.1109/tits.2005.848368

[38] Robert A Henning, Pierre Jacques, George V Kissel, Anne B Sullivan, and Sabina M Alteras-Webb. 1997. Frequent short rest breaks from computer work: effects on productivity and well-being at two field sites. *Ergonomics* 40, 1 (1997), 78–91. DOI: http://dx.doi.org/10.1080/001401397188396

[39] Nis Hjortskov, Dag Rissén, Anne Katrine Blangsted, Nils Fallentin, Ulf Lundberg, and Karen Søgaard. 2004. The effect of mental stress on heart rate variability and blood pressure during computer work. *European Journal of Applied Physiology* 92, 1 (01 Jun 2004), 84–89. DOI: http://dx.doi.org/10.1007/s00421-004-1055-z

[40] Eric Horvitz, Paul Koch, and Johnson Apacible. 2004. BusyBody: Creating and Fielding Personalized Models of the Cost of Interruption. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work (CSCW '04)*. ACM, New York, NY, USA, 507–510. DOI:http://dx.doi.org/10.1145/1031607.1031690

[41] James M. Hudson, Jim Christensen, Wendy A. Kellogg, and Thomas Erickson. 2002. "I'd Be Overwhelmed, but It's Just One More Thing to Do": Availability and Interruption in Research Management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '02)*. ACM, New York, NY, USA, 97–104. DOI: http://dx.doi.org/10.1145/503376.503394

[42] Christophe Hurter, Benjamin R. Cowan, Audrey Girouard, and Nathalie Henry Riche. 2012. Active Progress Bar: Aiding the Switch to Temporary Activities. In *Proceedings of the 26th Annual BCS Interaction Specialist Group Conference on People and Computers (BCS-HCI '12)*. BCS Learning Development Ltd., Swindon, GBR, 99–108.

[43] Shamsi T. Iqbal, Piotr D. Adamczyk, Xianjun Sam Zheng, and Brian P. Bailey. 2005. Towards an Index of Opportunity: Understanding Changes in Mental Workload during Task Execution. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*. Association for Computing Machinery, New York, NY, USA, 311–320. DOI: http://dx.doi.org/10.1145/1054972.1055016

[44] Shamsi T. Iqbal and Brian P. Bailey. 2011. Oasis: A Framework for Linking Notification Delivery to the Perceptual Structure of Goal-Directed Tasks. *ACM Trans. Comput.-Hum. Interact.* 17, 4, Article Article 15 (Dec. 2011), 28 pages. DOI: http://dx.doi.org/10.1145/1879831.1879833

[45] Shamsi T. Iqbal and Eric Horvitz. 2007. Disruption and Recovery of Computing Tasks: Field Study, Analysis, and Directions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 677–686. DOI: http://dx.doi.org/10.1145/1240624.1240730

[46] Shamsi T. Iqbal and Eric Horvitz. 2010. Notifications and Awareness: A Field Study of Alert Usage and Preferences. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW '10)*. Association for Computing Machinery, New York, NY, USA, 27–30. DOI: http://dx.doi.org/10.1145/1718918.1718926

[47] Shamsi T. Iqbal, Xianjun Sam Zheng, and Brian P. Bailey. 2004. Task-Evoked Pupillary Response to Mental Workload in Human-Computer Interaction. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems (CHI EA '04)*. Association for Computing Machinery, New York, NY, USA, 1477–1480. DOI: http://dx.doi.org/10.1145/985921.986094

[48] Youxuan Jiang, Jonathan K. Kummerfeld, and Walter S. Lasecki. 2017. Understanding Task Design Trade-offs in Crowdsourced Paraphrase Collection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 103–109.

[49] Jing Jin and Laura A. Dabbish. 2009. Self-interruption on the Computer: A Typology of Discretionary Task Interleaving. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 1799–1808. DOI: http://dx.doi.org/10.1145/1518701.1518979

[50] Pamela Karr-Wisniewski and Ying Lu. 2010. When more is too much: Operationalizing technology overload and exploring its impact on knowledge worker productivity. *Computers in Human Behavior* 26, 5 (2010), 1061–1072. DOI: http://dx.doi.org/10.1016/j.chb.2010.03.008

[51] Ioanna Katidioti, Jelmer P. Borst, Douwe J. Bierens de Haan, Tamara Pepping, Marieke K. van Vugt, and Niels A. Taatgen. 2016. Interrupted by Your Pupil: An Interruption Management System Based on Pupil Dilation. *International Journal of HumanâĂŞComputer Interaction* 32, 10 (2016), 791–801. DOI: http://dx.doi.org/10.1080/10447318.2016.1198525

[52] Jana Kühnel, Hannes Zacher, Jessica De Bloom, and Ronald Bledow. 2017. Take a break! Benefits of sleep and short breaks for daily work engagement. *European Journal of Work and Organizational Psychology* 26, 4 (2017), 481–491. DOI: http://dx.doi.org/10.1080/1359432X.2016.1269750

[53] Craig R Littler. 1978. Understanding taylorism. *British Journal of Sociology* (1978), 185–202.

[54] 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 (CHI '18)*. ACM, New York, NY, USA, Article 127, 14 pages. DOI:http://dx.doi.org/10.1145/3173574.3173701

[55] Mika Mäntylä, Bram Adams, Giuseppe Destefanis, Daniel Graziotin, and Marco Ortu. 2016. Mining Valence, Arousal, and Dominance: Possibilities for Detecting Burnout and Productivity?. In *Proceedings of the 13th International Conference on Mining Software Repositories (MSR '16)*. Association for Computing Machinery, New York, NY, USA, 247–258. DOI: http://dx.doi.org/10.1145/2901739.2901752

[56] Gloria Mark, Daniela Gudith, and Ulrich Klocke. 2008. The Cost of Interrupted Work: More Speed and Stress. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 107–110. DOI: http://dx.doi.org/10.1145/1357054.1357072

[57] Gloria Mark, Shamsi Iqbal, Mary Czerwinski, and Paul Johns. 2015. Focused, Aroused, but So Distractible: Temporal Perspectives on Multitasking and Communications. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 903–916. DOI: http://dx.doi.org/10.1145/2675133.2675221

[58] 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 (CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3025–3034. DOI: http://dx.doi.org/10.1145/2556288.2557204

[59] Gloria Mark, Shamsi T. Iqbal, Mary Czerwinski, Paul Johns, and Akane Sano. 2016. Neurotics Can't Focus: An in Situ Study of Online Multitasking in the Workplace. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1739–1744. DOI: http://dx.doi.org/10.1145/2858036.2858202

[60] Daniel McDuff, Sarah Gontarek, and Rosalind W Picard. 2014. Improvements in remote cardiopulmonary measurement using a five band digital camera. *IEEE Transactions on Biomedical Engineering* 61, 10 (2014), 2593–2601. DOI: http://dx.doi.org/10.1109/tbme.2014.2323695

[61] Daniel McDuff, Kael Rowan, Piali Choudhury, Jessica Wolk, ThuVan Pham, and Mary Czerwinski. 2019. A Multimodal Emotion Sensing Platform for Building Emotion-Aware Applications. *arXiv preprint arXiv:1903.12133* (2019).

[62] Daniel McFarlane and Kara Latorella. 2002. The Scope and Importance of Human Interruption in Human-Computer Interaction Design. *Human-computer Interaction* 17 (03 2002), 1–61. DOI: http://dx.doi.org/10.1207/S15327051HCI1701_1

[63] Wesley P. McTernan, Maureen F. Dollard, and Anthony D. LaMontagne. 2013. Depression in the workplace: An economic cost analysis of depression-related productivity loss attributable to job strain and bullying. *Work & Stress* 27, 4 (2013), 321–338. DOI: http://dx.doi.org/10.1080/02678373.2013.846948

[64] Brid O'Conaill and David Frohlich. 1995. Timespace in the Workplace: Dealing with Interruptions. In *Conference Companion on Human Factors in Computing Systems (CHI '95)*. ACM, New York, NY, USA, 262–263. DOI: http://dx.doi.org/10.1145/223355.223665

[65] Marco Ortu, Bram Adams, Giuseppe Destefanis, Parastou Tourani, Michele Marchesi, and Roberto Tonelli. 2015. Are Bullies More Productive? Empirical Study of Affectiveness vs. Issue Fixing Time. In *Proceedings of the 12th Working Conference on Mining Software Repositories (MSR '15)*. IEEE Press, 303–313.

[66] Leslie A Perlow. 1999. The time famine: Toward a sociology of work time. *Administrative science quarterly* 44, 1 (1999).

[67] June J. Pilcher, Douglas R. Ginter, and Brigitte Sadowsky. 1997. Sleep quality versus sleep quantity: Relationships between sleep and measures of health, well-being and sleepiness in college students. *Journal of Psychosomatic Research* 42, 6 (1997), 583–596. DOI: http://dx.doi.org/10.1016/S0022-3999(97)00004-4

[68] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. 2011. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering* 58, 1 (2011), 7–11. DOI: http://dx.doi.org/10.1109/tbme.2010.2086456

[69] Michael I Posner, Mary Jo Nissen, and William C Ogden. 1978. Attended and unattended processing modes: The role of set for spatial location. *Modes of perceiving and processing information* 137, 158 (1978), 2.

[70] Peter Richter, Thomas Wagner, Ralf Heger, and Gunther Weise. 1998. Psychophysiological analysis of mental load during driving on rural roads-a quasi-experimental field study. *Ergonomics* 41, 5 (1998), 593–609. DOI: http://dx.doi.org/10.1080/001401398186775

[71] Joshua S. Rubinstein, David Meyer, and Jeffrey E. Evans. 2001. Executive Control of Cognitive Processes in Task Switching. 27 (Sept 2001), 763–97. DOI:http://dx.doi.org/10.1037/0096-1523.27.4.763

[72] Greg Smith, Patrick Baudisch, George Robertson, Mary Czerwinski, Brian Meyers, Daniel Robbins, and Donna Andrews. 2003. GroupBar: The TaskBar Evolved. In *Proceedings of OZCHI 2003*. 34–43.

[73] Barry M Staw. 1986. Organizational psychology and the pursuit of the happy/productive worker. *California Management Review* 28, 4 (1986), 40–53.

[74] Kenneth T Strongman and Christopher DB Burt. 2000. Taking breaks from work: An exploratory inquiry. *The Journal of psychology* 134, 3 (2000), 229–242. DOI: http://dx.doi.org/10.1080/00223980009600864

[75] Toon W. Taris and Paul J.G. Schreurs. 2009. Well-being and organizational performance: An organizational-level test of the happy-productive worker hypothesis. *Work & Stress* 23, 2 (2009), 120–136. DOI: http://dx.doi.org/10.1080/02678370903072555

[76] Vincent W.-S. Tseng, Matthew L. Lee, Laurent Denoue, and Daniel Avrahami. 2019. Overcoming Distractions During Transitions from Break to Work Using a Conversational Website-Blocking System. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 467, 13 pages. DOI: http://dx.doi.org/10.1145/3290605.3300697

[77] Antoine U Viola, Lynette M James, Luc JM Schlangen, and Derk-Jan Dijk. 2008. Blue-enriched white light in the workplace improves self-reported alertness, performance and sleep quality. *Scandinavian journal of work, environment & health* (2008), 297–306.

[78] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063.

[79] Christopher D. Wickens. 2008. Multiple Resources and Mental Workload. *Human Factors* 50, 3 (2008), 449–455. DOI: http://dx.doi.org/10.1518/001872008X288394

[80] Alex C. Williams, Harmanpreet Kaur, Gloria Mark, Anne Loomis Thompson, Shamsi T. Iqbal, and Jaime Teevan. 2018. Supporting Workplace Detachment and Reattachment with Conversational Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article Paper 88, 13 pages. DOI: http://dx.doi.org/10.1145/3173574.3173662

[81] Thomas A Wright and Russell Cropanzano. 1997. Well-being, Satisfaction and Job Performance: Another Look at the Happy/Productive Worker Thesis. In *Academy of Management Proceedings*, Vol. 1997. Academy of Management Briarcliff Manor, NY 10510, 364–368. DOI: http://dx.doi.org/10.5465/ambpp.1997.4988986

[82] Thomas A Wright and Barry M Staw. 1999. Affect and favorable work outcomes: two longitudinal tests of the happy–productive worker thesis. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 20, 1 (1999), 1–23. DOI: http://dx.doi.org/10.1002/(SICI)1099-1379(199901)20:1<1::AID-JOB885>3.0.CO;2-W

[83] Biqiao Zhang and Emily Mower Provost. 2019. Automatic recognition of self-reported and perceived emotions. In *Multimodal Behavior Analysis in the Wild*. Elsevier, 443–470. DOI: http://dx.doi.org/10.1016/B978-0-12-814601-9.00027-4

[84] 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 (CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article Paper 591, 14 pages. DOI:http://dx.doi.org/10.1145/3173574.3174165