Sensing Behavioral Change over Time: Using Within-Person Variability Features from Mobile Sensing to Predict Personality Traits

WEICHEN WANG, Dartmouth College, USA
GABRIELLA M. HARARI, Stanford University, USA
RUI WANG, Dartmouth College, USA
SANDRINE R. MÜLLER, University of Cambridge, United Kingdom
SHAYAN MIRJAFARI, KIZITO MASABA, and ANDREW T. CAMPBELL, Dartmouth College, USA

Personality traits describe individual differences in patterns of thinking, feeling, and behaving ("between-person" variability). But individuals also show changes in their own patterns over time ("within-person" variability). Existing approaches to measuring within-person variability typically rely on self-report methods that do not account for fine-grained behavior change patterns (e.g., hour-by-hour). In this paper, we use passive sensing data from mobile phones to examine the extent to which within-person variability in behavioral patterns can predict self-reported personality traits. Data were collected from 646 college students who participated in a self-tracking assignment for 14 days. To measure variability in behavior, we focused on 5 sensed behaviors (ambient audio amplitude, exposure to human voice, physical activity, phone usage, and location data) and computed 4 within-person variability features (simple standard deviation, circadian rhythm, regularity index, and flexible regularity index). We identified a number of significant correlations between the within-person variability features and the self-reported personality traits. Finally, we designed a model to predict the personality traits from the within-person variability features. Our results show that we can predict personality traits with good accuracy. The resulting predictions correlate with self-reported personality traits in the range of $r = 0.32$, MAE $= 0.45$ (for Openness in iOS users) to $r = 0.69$, MAE $= 0.55$ (for Extraversion in Android users). Our results suggest that within-person variability features from smartphone data has potential for passive personality assessment.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**; • **Applied computing → Psychology**;

Additional Key Words and Phrases: Mobile Sensing, Personality Traits, Within-Person Variability

ACM Reference Format:
Weichen Wang, Gabriella M. Harari, Rui Wang, Sandrine R. Müller, Shayan Mirjafari, Kizito Masaba, and Andrew T. Campbell. 2018. Sensing Behavioral Change over Time: Using Within-Person Variability Features from Mobile Sensing to Predict Personality Traits. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 141 (September 2018), 21 pages. https://doi.org/10.1145/3264951

The research reported in this article is supported the National Science Foundation (NSF) Award BCS-1520288.
Authors' addresses: Weichen Wang, Dartmouth College, Computer Science, Hanover, NH, 03755, USA, wang@cs.dartmouth.edu; Gabriella M. Harari, Stanford University, Department of Communication, Stanford, CA, 94305, USA, gharari@stanford.edu; Rui Wang, Dartmouth College, Computer Science, Hanover, NH, 03755, USA, rui.wang.gr@dartmouth.edu; Sandrine R. Müller, University of Cambridge, Department of Psychology, Cambridge, United Kingdom, srm77@cam.ac.uk; Shayan Mirjafari, shayan@cs.dartmouth.edu; Kizito Masaba, kizito.masaba.gr@dartmouth.edu; Andrew T. Campbell, Dartmouth College, Computer Science, Hanover, NH, 03755, USA, campbell@cs.dartmouth.edu.

2474-9567/2018/9-ART141 $15.00
https://doi.org/10.1145/3264951

1 INTRODUCTION

Personality psychology focuses on examining individual differences in people's thoughts, feelings, and behaviors. Compared to the amount of research on people's thoughts and feelings, considerably less research has examined how people behave in the context of everyday life (e.g., daily levels of physical activity, sociability, places visited). Traditionally, research examining individual differences has focused on *between-person* variability in mean levels of such behaviors. For example, people who are more extroverted can be characterized by higher mean levels of talkativeness, compared with people who are less extroverted. However, people also vary in the extent to which their own behavioral patterns change over time, which is referred to as *within-person* variability. For example, a person who is described as being more extroverted, may show great variability (e.g., socializing little during the week, but a great amount during the weekend) or stability (e.g., socializing a similar amount every day of the week) in their behavior patterns when considered over time.

Past research has shown within-person variability to be linked to various psychological characteristics, such as a person's affective states [35], mental well-being [47] and personality trait ratings [19]. However, past research often relied on a person's capacity to accurately recall their daily experiences retrospectively [52, 53], a task that is challenging and time-consuming. Individuals may also intentionally under-report or over-estimate some of their behaviors [64]. Fortunately, smartphone sensing methods [39] are set to overcome these barriers by unobtrusively measuring behavioral patterns continuously over time and thereby allowing us to understand the fine-grained within-person variability in behavioral patterns.

In this paper, we present a new approach to capture within-person variability in behaviors using mobile sensing with the goal of assessing and predicting self-reported personality traits. We use the Big Five model [34] , which describes 5 major personality trait dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Openness is a personality trait that describes the extent to which a person is imaginative and insightful. Conscientousness is a trait that describes the extent to which a person is thoughtful, shows impulse control, and engages in goal-directed behaviors. Extraversion is a trait that describes the extent to which a person is excitable, social, talkative, and exhibits emotional expressiveness. Agreeableness is a trait that describes the extent to which a person is trusting, altruistic, kind, and engages in prosocial behaviors. Neuroticism is a trait that describes the extent to which a person is moody and emotionally unstable.

To examine how within-person variability in behavioral patterns are related to trait ratings, we consider the following everyday behaviors inferred using mobile sensing [27]: social interactions (i.e., how much a person socializes), physical activity and mobility (i.e., how physically active they are, and how many different places they spend time in), daily activities (i.e., how often they use their phone), and situational information (i.e., noisiness of their environment). To quantify within-person variability in lifestyle behaviors, we compute the following within-person variability metrics for each of the four everyday behaviors: standard deviation, circadian rhythm, regularity index, and flexible regularity index. We examine the connections between within-person variability metrics and personality traits; specifically, we pose the following broad research question: *To what extent do day-to-day behavioral patterns of stability and change reveal a person's personality traits?* To answer this research question, we collected self-reported personality trait ratings as our measure of ground truth, along with the following sensing data as our measures of lifestyle behaviors: ambient audio amplitude levels (which indicates how quiet or noisy the environment is), exposure to human voice (relating to how social the user is), physical activity, phone usage and location data from 646 students using their Android and iPhones at the University of Texas at Austin (UTA) over a course of up to 2 weeks. The contributions of the paper are as follows:

- We demonstrate for the first time how within-person variability patterns collected passively by a comprehensive cross-platform (i.e., Android and iOS) mobile sensing app can be used to predict personality traits. Specifically, we measure and assess everyday behaviors including social interactions, physical movement,

daily activity and situational information. Furthermore, we propose using the following measures of within-person variability: standard deviation, circadian rhythm, regularity index, and flexible regularity index. These measures capture behavioral variability from different perspectives (these measures are explained in more detail in Section 3).

- We identify a number of important associations between the within-person variability features and self-reported personality traits. Furthermore, we predict personality traits solely based on within-person behavior change features. Our results show that our proposed personality prediction model based on within-person variability features provides good estimation of personality traits, particularly for extraversion and agreeableness. For example, for Android users the leave-one-out model for predicting the extraversion trait achieves 0.55 of MAE, which is 0.24 (30%) lower than the average baseline and 0.5 (48%) lower than the random baseline; for iOS users the leave-one-out model for predicting the extraversion trait achieves 0.61 of MAE, which is 0.11 (15%) lower than the average baseline and 0.39 (39%) lower than the random baseline.

To the best of our knowledge, we are the first to explore how within-person variability patterns can be used to predict personality traits using features derived from mobile sensing. Our results pave the way for future research on this psychological topic. The structure of the paper is as follows. First, we present related work on personality and mobile sensing research in Section 2, followed by a detailed description of the "within-person variability" measurements in Section 3. We describe the sensing system, study design, and the dataset in Section 4. Following that, we discuss the correlations between the within-person variability features and personality traits in Section 5.1. In Section 5.2, we present the results of using the within-person variability features to regress on and predict personality scores. We discuss limitations of our methods in Section 6. Finally, we present concluding remarks and the implications in Section 7.

2 RELATED WORK

In recent years, mobile sensing has demonstrated its potential as a tool for tracking and modeling human behavior [28, 29, 51, 62]. Equipped with unobtrusive sensors, smartphones can collect continuous sensing data that reveal individuals' behavioral patterns and psychological states over long periods of time. For example, several studies have used smartphone sensing to continuously assess people's mental health [5, 39, 58]. The StudentLife study [68] investigated the relationship between many types of smartphone data (e.g., conversation, sleep, activity, and co-location) and mental health outcomes (e.g., depression, stress, loneliness, and flourishing) in Dartmouth students during an academic term. Using the same dataset, Harari et al. [26] analyzed the changes of students' activity and sociability behaviors over a term via the accelerometer and microphone sensors. Ben-Zeev et al. [6] used mobile phones to collect passive sensing data from smartphones and find schizophrenia relapse signals in location, activity, and exposure to conversation prior to patients experiencing relapses. Saeb et al. [62] reported depressive symptom severity correlates with mobility patterns and phone usage derived from smartphone data. They replicated their findings using the StudentLife [68] dataset [61]. Canzian and Musolesi [13] proposed a location routine index computed from smartphone location data, which was predictive depression severity. Abdullah et al. [1] reported using location features computed from smartphone data, such as distance traveled, conversation frequency, and non-stationary duration, to infer the social rhythm metric (SRM) [53] score, a widely used lifestyle regularity metric.

Several other studies have focused on inferring stability or variability in lifestyle behavioral patterns inferred from smartphone sensing data. For example, Abdullah et al. [2] computed daily rhythms related to sleep to measure well-being. Saeb et al [62] explored using circadian rhythm inferred from smartphone location data to assess depressive symptom severity. Ghandeharioun et al. [21] computed a sleep regularity index (SRI) using accelerometer data and show that SRI differs significantly between days with good and poor mental health.

Mehtrotra and Musolesi [51] proposed a movement digital biomarker for monitoring emotional state , which measures the similarity between the sequences of visited places in a day.

There has also been an increasing amount of research focused on predicting personality traits from digital media data. Researchers have sought to predict personality using social network structures and interactions , showing that the behavioral data collected from social media platforms such as Facebook [44, 54, 66, 72] or Twitter [22, 23, 57] can be informative in the prediction of people's personalities. Youyou et al. [72] showed that using Facebook Likes, a computer can predict participants' personality more accurately than their Facebook friends. Moreover, the computer-made personality judgments had higher external validity when predicting life outcomes (e.g., substance use, political attitudes, and physical health). Park et al. [54] built a predictive model of personality based on 66,732 Facebook users' written language. The predicted personality scores correlate with the ground truth. Golbeck et al. [22] shows that Twitter users' language use, sentiment, and Twitter use can be used to predict personality traits. However, these methods usually require access to extensive information about people's online social networks.

Researchers have also utilized the information from mobile phones in personality assessment studies. Early studies tried to find relationships between personality traits and phone communications (i.e., phone calls and text messages) [11, 14, 16]. For instance, Montjoye et al. [15] used standard mobile phone logs (i.e., calls and texts) to predict users' personality. Staiano et al. [63] collected call logs and Bluetooth proximity data from 53 subjects over 8 weeks to build a call network and a Bluetooth proximity network and used network characteristics to predict users' personality traits. A recent study [67] explored using the StudentLife dataset, which comprises Wi-Fi, GPS, Bluetooth, accelerometer, and Piazza usage data to assess personality, finding significant correlations between behavior features Wi-Fi location based behavior features and personality traits. However, much of the aforementioned work used behavioral features focused on capturing 'between-person' variability in behavioral patterns. In this paper, we focus on capturing and assessing within-person variability features from objective sensing behaviors, and show how these features can predict personality traits of smartphone users.

3 WITHIN-PERSON VARIABILITY MEASUREMENTS

In what follows, we describe the passive smartphone sensing measures of behavior that we used to quantify behavioral variability. We first describe the behavioral data collected (social interactions, movement and mobility, daily activity, and situational information), then we introduce the metrics that we used to quantify within-person variability (Standard Deviation, Circadian Rhythm, Regularity Index, Flexible Regularity Index).

3.1 Behaviors Inferred from Passive Smartphone Sensing

In what follows, we present the four aspects of daily behaviors that are captured through passive sensing using smartphones. We extend the StudentLife Android app [1, 40, 68, 70], which was originally used to capture students' behaviors during a term and port it to the Apple iOS platform. The app measures daily social interactions, movement and mobility, daily activities and situational information by continuously collecting audio amplitude, ambient voice, participants' physical activities, lock/unlock events and location coordinates.

Social interaction. We consider inferred *ambient voice labels* as a proxy for social interactions (i.e., being around conversation). We implement a conversation classifier on the phone to infer whether or not a 32ms audio frame is human voice. The voice classifier is implemented using a duty-cycled audio sensor that continuously runs on the smartphone. The voice classifier is the most energy consuming module in the app. To save energy, we set the duty cycle to be 1 minute on and 3 minutes off. It has been show in [68] that using this duty cycle we can achieve a balance between accuracy and resource usage. Further more, to preserve participant privacy, only the labels from the classifier are kept in the dataset and no raw human voices or speech content is recorded. Our classifier uses privacy preserving features [58, 68] and first determines if the frame contains speech and if so a higher level

conversation classifier determines if there are sufficient speech frames to indicate the start and later the end of a conversation. The frequency and duration of conversations are stored on the phone and uploaded for analysis. Note that the conversation classifier does not use speaker identification for privacy reasons and therefore only indicates if the user is in the presence of conversation rather than being an active speaker. Therefore, we consider our conversation inferences a proxy for social engagement. The speech/conversation classifier has been validated in several previous studies [6, 58, 68, 70].

Movement and mobility. The phone provides *location data* that allows us to understand users' movement and mobility patterns. We find a user's significant places (i.e., those places where users spent significant time) during the day and their associated dwell times (when the user arrives and leaves a location) by clustering the sampled coordinates during a day using density-based spatial clustering of applications with noise (DBSCAN) [18]. The DBSCAN algorithm groups the points that are close to each other and computes the center of the cluster. The center of the cluster is considered a significant location.

Daily activity. We consider two kinds of daily activity: *physical activity* and *phone activity* (i.e., phone usage). Personality traits are hypothesized to exert influence on physical activity through a health-behavior model [48, 60, 71]. Our app obtains activity inferences (i.e., stationary, walking, running, cycling, in vehicle) from the Android activity recognition API [24] and iOS Core Motion [32]. We compute the sedentary duration within every hour of the day using the phone's physical activity inferences. Another aspect of daily activity we consider as providing signal is phone usage. An increasing number of researchers have shown that smartphone usage reflects psychological well-being [10, 11, 36, 41, 45]. We compute the number of phone lock/unlock events and phone unlock duration to estimate the phone usage.

Situational information. We use *ambient sound* from the phone as a proxy for contextual information about the environment of users. Previous work [50] shows that participants' personality traits are related to the quotidian manifestations derived from the sampled snippets of ambient sounds of users immediate environment. We periodically collect sound levels to measure the ambient sound environment. For privacy reasons we do not store any raw audio data. Rather, we compute the average sound amplitude over a period of one second so that the audio can not be reconstructed.

3.2 Within-Person Variability Metrics

Standard deviation is one of the simplest and most common approaches to measuring within-person variability. However, there may be other behavioural change metrics that better capture the patterns of change and stability in peoples' everyday lives. These approaches could be more meaningful in better understanding within-person variability and its relationship to personality. To study this, we compute a number of variability measures including: the simple standard deviation, the circadian rhythm [2, 13, 61, 62], the regularity index and the flexible regularity index. We use these measures to assess the within-person variability of each behavior inferred using passive sensing. We first partition a day's data into 24 one-hour periods and process the sensor data in an hourly fashion. For example, consider ambient audio amplitude, we compute the mean audio amplitude for each one-hour period; for voice, we compute the amount of conversation duration measured in one-hour periods; for physical activity, we compute the sedentary duration across each hour period; for the phone usage, we compute the number of phone unlock events registered during each hour period. We do not preprocess location data. Next, we compute the four within-person variability metrics discussed above. Table 1 summarizes the within-person variability features used in this study. In what follows, we describe each metric in detail.

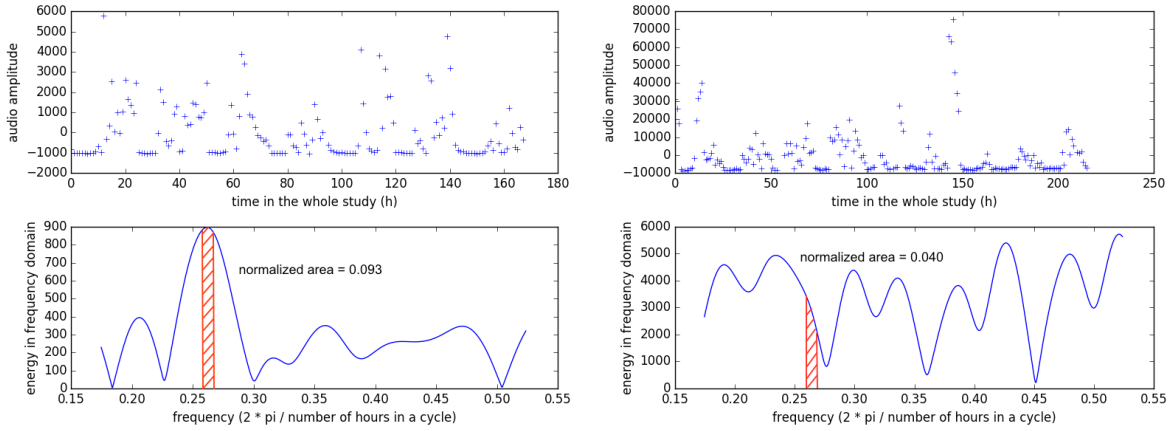
Standard deviations [STD] measure the variance in daily behaviors. We compute the STD over three epochs during a 24-hour period: day time (9am–6pm), evening (6pm-12am), and night (12am-9am) across all days of the week. Because people are likely to have different behavioral patterns during weekdays in comparison to

weekends, we compute the STD for the three epochs (i.e. day, evening, night) for weekdays data only. We do not, however, compute the STD for weekends because of limited amounts of weekend data.

Circadian rhythm [CR][2, 13, 61, 62] measures the strength with which a user follows a 24-hour rhythm in behaviors [61]. Humans have a biological clock that optimizes the physiology and behavior of organisms, hormonal secretion and mood [37]. However, people differ in their circadian rhythms. For example, previous studies have shown individual differences in the morning and evening related to personality [3, 59]. We compute the CR across the study period for hourly behavioral data using spectrum analysis. Specifically, we first use the least-squares spectral analysis [56] to transform the behavioral sensing data (i.e., physical activity, phone activity ambient sound, ambient voice) from the time domain to the frequency domain. We then compute the ratio of energy that fall into the 24 ± 0.5 h period (which corresponds to $2\pi/(24 \pm 0.5) = (0.2565, 0.2674)$) over the total spectrum energy in the 24 ± 12 h period (which corresponds to $2\pi/(24 \pm 12) = (0.1745, 0.5236)$) as the CR:

$$CR = \frac{\int_{2\pi/(24.5)}^{2\pi/(23.5)} psd(x)dx}{\int_{2\pi/(36)}^{2\pi/(12)} psd(x)dx} \tag{1}$$

where $psd(x)$ denotes the power spectral density at frequency bin x. The raw hourly signal is first aligned to zero mean before performing spectral analysis. Fig. 1 shows an example of the computed CR for ambient sound. The plots show the raw hourly ambient sound amplitude and the corresponding power spectrum for two participants selected from the study. The raw data show that participant (a) has a more pronounced 24-hour cycle for ambient sound than (b). As such, the CR value for participant (a) is 0.093 whereas the CR value for participant (b) is 0.040. For the mobility data, we use the steps described in [61] to compute the CR of mobility: we first generate the CR for the latitude and longitude values then combine them through $CR = \log(CR_{lat} + CR_{long})$.



(a) Participant with high circadian rhythm value

(b) Participant with low circadian rhythm value

Fig. 1. Circadian Rhythm analysis using audio amplitude sensing data from two participants.

Regularity index [RI] assesses the difference between the same hours across two different days. We first rescale the behavioral data for each participant to $[-0.5, 0.5]$, where -0.5 corresponds to the minimum value in the origin data and 0.5 corresponds to the maximum value. The product of two rescaled values is positive if the original

values are close and negative if they are not similar. Subsequently, we define the regularity *between* day a and b as:

$$\forall(a,b) \in S, \quad RI_{a,b} = \sum_{t-1}^{T} f(x_t^a) f(x_t^b)/T \tag{2}$$

where S is the set of two-day pairs, a, b are the two days in a two-day pair, T = 24hours, and x_t^a is the rescaled value in hour t of day a.

We compute the mobility RI differently because unlike other behavioral data mobility data is nominal (i.e., a mobility data point represent a location). We therefore compare whether or not a user is at the same place in two days. The mobility RI is formally defined as:

$$RI\_loc_{a,b} = -\sum_{t=1}^{T_{loc}} g(c_t^a, c_t^b)/T_{loc} \tag{3}$$

where a, b represent two different days, t represent the time window in a day (a window lasts 10 minutes), T_{loc} is the available number of overlapped time windows in both days, c_t^a is the significant location id time t in the day a, and $g(m, n)$ indicate whether or not $m = n$. A higher RI score indicates that the user visited similar places around a similar time of day in two given days.

We compute the average and range of the RI values from every possible pair within the following sets: (1) weekdays vs weekends, (2) within weekdays, (3) within all days.

Flexible regularity index [FRI] is an edit distance based (or Levenshtein distance [42]) measure to assess the difference between two days differently. An edit distance quantifies how dissimilar two strings are to one another by counting the minimum number of operations needed to transform one string to the other. Such operations include removing, inserting, or substituting one character in the string. Different operations may have different weights. The edit distance of the behavioral data in two days reveals how similar the behaviors in two days are. A lower edit distance (i.e., lower FRI) between different two days indicates more similar behaviors.

We compute the FRI as follows. First, we transform the behavioral sensing data into strings. Specifically, for behavioral data other than mobility, we label a one-hour chuck as 'a' if the mean sensor reading in this hour is within the bottom 25 percentile of all data from this user; 'c' if the mean is within the top 25 percentile; and 'b' if the mean is between the bottom 25 percentile and top 25 percentile. For mobility, we use the significant location id to generate the mobility string in day. We define the weights for each operation as shown in Table 2.

4 DATA COLLECTION AND PROCESSING

We collected a dataset from 646 students at the University of Texas at Austin (UTA). The participants were enrolled in an online introductory psychology class across two semesters. As part of a course assignment, participants could self-track their lifestyle behaviors in exchange for personalized feedback using a tracking method of their choice: a mobile sensing app, email-based surveys, or a handwritten journal. Here we focus on the data collected from students who elected to use the mobile sensing app. Participants installed the data collection app on their phones and were asked to participate for at least seven days. Participants were able to participate for up to fourteen days. Among the 646 participants, 117 used Android phones and 529 used iPhones. All participants complete the Big Five personality trait questionnaire [34] at the start of the study period, which serves as the ground truth for personality traits in this study.

Fig. 2 shows the overall system and study design. The complete system included the sensing app and cloud and was based on an earlier version of the StudentLife system [68]. We continuously collected behavioral passive sensing data from participants' Android phones and iPhones. The data was then automatically uploaded to our

Table 1. Description of the features computed

Feature category	Features computed	Description of feature	Sensing data
Standard Deviation (STD)	std_night_all	STD on the data at night across all days	Ambient sound Ambient voice Physical activity Phone activity
	std_day_all	STD on the data during day time across all days	
	std_evening_all	STD on the data in the evening across all days	
	std_night_weekday	STD on the data at night across weekdays	
	std_day_weekday	STD on the data during day time across weekdays	
	std_evening_weekday	STD on the data at night across weekdays	
Circadian Rhythm (CR)	circadian_all	Circadian rhythm from data across all days	Ambient sound Voice labels Physical activity Phone activity Location
	circadian_weekday	Circadian rhythm from data across weekends	
Regularity Index (RI)	ri_weekday_vs_weekend_avg	Measures the average hour-by-hour similarity between weekdays and weekends	Ambient sound Ambient voice Physical activity Phone activity Location
	ri_weekday_avg	Measures the average hour-by-hour similarity across weekdays	
	ri_all_avg	Measures the average hour-by-hour similarity across all days	
	ri_weekday_vs_weekend_range	Measures the range (i.e., the difference of the most similar pair and the most distinguished pair) of hour-by-hour similarity between weekdays and weekends	
	ri_weekday_range	Measures the range of hour-by-hour similarity between weekdays	
	ri_all_range	Measures the range of hour-by-hour similarity across all days	
Flexible Regularity Index (FRI)	fri_weekday_vs_weekend_avg	All the definitions are parallel to the RI. The FRI measures the "edit distance" between two days, which allows the hours to be slightly shifted when doing the hour-by-hour comparison with reasonable penalty.	Ambient sound Ambient voice Physical activity Phone activity Location
	fri_weekday_avg		
	fri_all_avg		
	fri_weekday_vs_weekend_range		
	fri_weekday_range		
	fri_all_range		

secure server. The server processed the data, and generated personalized web-pages of feedback reports, which were sent to students via email during the study. Those reports included personalized visualizations of the tracked behaviors as well as class average charts for comparison.

In what follows, we discuss how we processed the data in detail.

Table 2. Cost definition in FRI calculation

Operation	Cost		
	sensing data other than location		location sequence
insertion	1		1
removal	1		1
substitution	if letters are adjacent (e.g., 'a' and 'b')	0.5	1
	others	1.5	

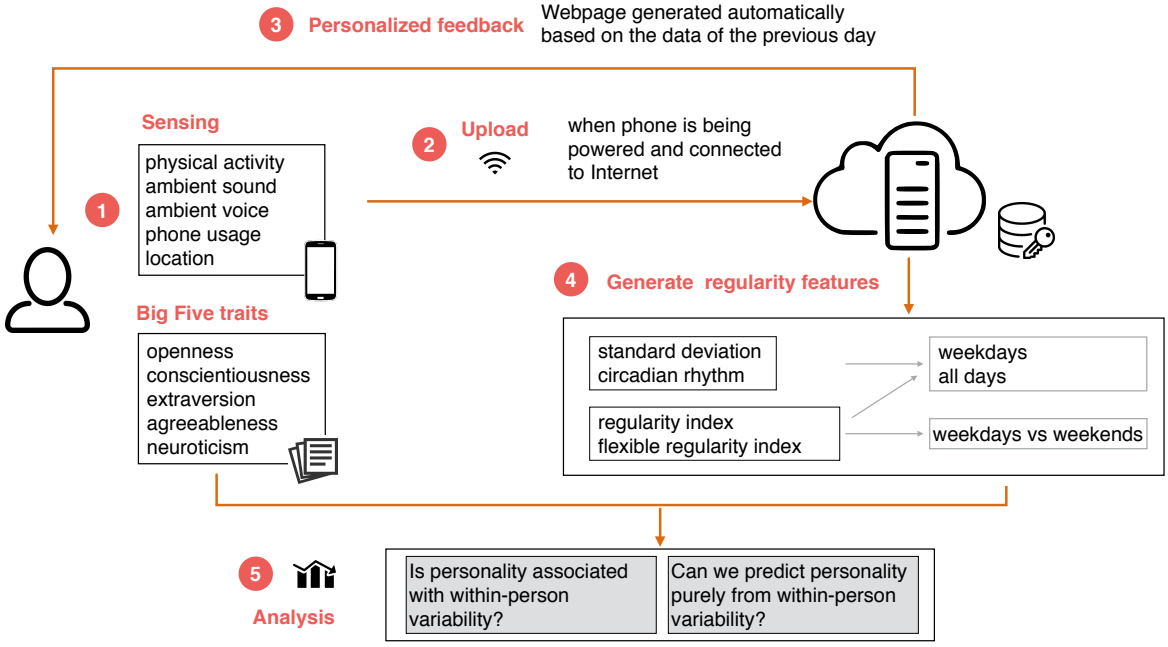


Fig. 2. System and study design.

4.1 Data Inclusion Criteria

Data quality is crucial for analysis. Missing data across a day will adversely affect the accuracy of the within-person variability features. Therefore, we exclude days with less than 19 hours of sensing data. The 19-hour threshold is based on previous studies [68, 70], which balances the need for data quality and quantity. We also exclude participants who have less than 7 days of usable data (more than 19 hours of sensing data). 159 out of the 646 participants satisfy our data inclusion criteria and included in our analysis. Among them, 70 are Android users and 89 are iPhone users. The high data exclusion rate is mainly due to participants running the app for less than 14 days. Other factors include the phone being powered down, turning location off, and stopping the app.

4.2 Big Five Personality Ground truth

We use the self-reported Big Five Inventory (BFI) [34] scores as our personality ground truth. The BFI measures the personality traits: openness, conscientiousness, extraversion, agreeableness and neuroticism. Fig. 3 shows the distributions of the Big Five scores of the included 159 participants. The distributions show that the values for all five traits approximate a normal distribution in this sample. Table 3 shows the mean and standard deviation of the five personality trait scores.

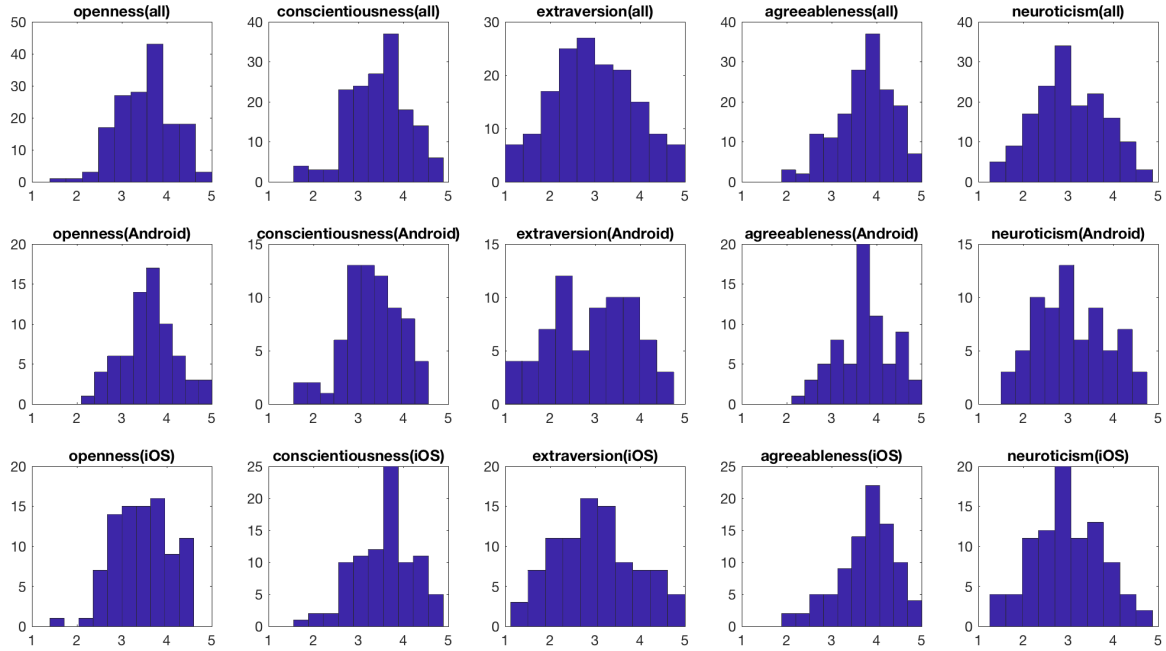


Fig. 3. Histograms of the Big Five scores. The X axis displays the value of the score, which ranges from 1 (lowest) to 5 (highest). The Y axis shows the number of individuals that fall into the specific score bins. The three rows show the distribution of the scores for all participants (first row), the Android (second row) and the iOS (third row) users.

Table 3. Distributions of the personality ground truth

Big Five trait	Mean (std)	Android mean (std)	iOS mean (std)	t-test p value
Openness	3.54 (0.62)	3.61 (0.61)	3.48 (0.62)	0.21
Conscientiousness	3.45 (0.65)	3.31 (0.63)	3.57 (0.65)	**0.01**
Extraversion	2.99 (0.90)	2.97 (0.92)	3.00 (0.88)	0.83
Agreeableness	3.74 (0.63)	3.70 (0.63)	3.76 (0.63)	0.63
Neuroticism	3.03 (0.79)	3.08 (0.79)	3.00 (0.80)	0.50

The mean of the trait scores are close to a score of 3 (the middle of the 1-5 range). The agreeableness score is close to 4, which is the highest mean of the trait scores, followed by openness, conscientiousness, neuroticism

and extraversion. For most of the personality traits, minor differences are of small to negligible effect size. This is inline with a recent study [25] among a large multi-national (N = 1081) and a German-speaking sample (N = 2438). However, in our data Android users seem to be less conscientious than iOS users (t-test $p = 0.01$), which is in contrast to the findings in [25]. The reason could be the population and age: unlike a large multi-national user group with various occupations, our study participants are college students from one university. In addition, our population of users is smaller.

5 ASSESSING PERSONALITY USING WITHIN-PERSON VARIABILITY MEASURES

In what follows, we present the results from association analysis and prediction of personality traits using passive sensing data from smartphones. We first extract all the within-person variability features and report linear correlations between the features and the self-reported personality scores. Then, we create a personality prediction model and analyze its prediction performance using only within-person variability features.

5.1 Association Analysis

We use the bivariate linear mixed model [49] to assess the relationship between the within-person variability features and Big Five personality traits. In our study, the sensing data come from two clusters: Android and iOS. There exist potential differences between two systems (e.g., both Google and Apple have their own physical activity classifiers and audio software development kits). Therefore, the sensed behavioral patterns are not independent. Linear mixed models are an extension of simple linear models to allow both fixed and random effects, and are particularly useful when there is non-independence in the data. Our association results are presented in Table 4. In order to address the multiple comparisons problem, we apply the Benjamini-Hochberg procedure (BH) [8, 9] to control the false discovery rate (FDR) in our exploratory regression analysis. The multiple comparisons problem arise when multiple simultaneous statistical tests are involved in the analysis, which may lead to erroneous discoveries. We present associations with $p < 0.05$ and mark associations that have FDR < 0.1 and FDR < 0.05.

In what follows, we discuss our results as they relate to the personality traits.

Openness. We find that four within-person variability features are positively associated with the openness trait. They are as follows: (1) the deviation in ambient sound on weekday (i.e. Monday-Friday) evenings (6pm-12am); (2) the deviation in physical activity in the evening across all days (i.e. Monday-Sunday); (3) the deviation in physical activity in the evening of weekdays; and finally (4) the range of the regularity index in ambient sound during weekdays. Our results indicate that participants who have various activities during the weekday evenings are more likely to be more open to new experiences. On several weekdays they stay in similar environments at the same hour of each day. However, on some other weekdays they spend time in very different environments (as captured by ambient sound).

We find that two within-person variability features are negatively associated with the openness trait. They are (1) the deviation in ambient voice between 9am - 6pm across all days; and (2) the deviation in ambient voice between 9am - 6pm on weekdays. Assuming ambient voice is a proxy for social interaction, our results indicate that people who have changing patterns in interaction with others during the daytime are likely to be less open to new experiences.

Conscientiousness. We find that four within-person variability features are positively associated with the conscientiousness trait. These are (1) the average of the flexible regularity index in ambient voice on weekdays; (2) the circadian rhythm in phone usage across all days; (3) the circadian rhythm in phone usage on weekdays; and (4) the flexible regularity index in physical activity. Our results indicate that participants who spend weekdays engaged in more regular social interactions (with hours slightly shifted), and who show more regular phone

usage during the day, and have regular patterns of physical activity during weekdays are more likely to have good impulse control and goal-directed behaviors.

We find that five within-person variability features are negatively associated with the conscientiousness trait. These are (1) the deviation of exposure to ambient voice on weekday evenings; (2) the regularity index in locations between weekdays and weekends; (3) the regularity index in locations across all days; (4) the flexible regularity index in locations between weekdays and weekends; and finally (5) the flexible regularity index in locations across all days. This leads us to believe that individuals who have unstable social interactions during weekday evenings, and more overlap between weekdays and weekends in terms of their location routines are likely to be less conscientious.

Extraversion. We find that five within-person variability features are positively associated with the conscientiousness trait. These are (1) the circadian rhythm in ambient sound across all days; (2) the circadian rhythm in phone usage across all days; (3) the circadian rhythm in phone usage on weekdays; (4) the deviation in physical activity in the evening across all days; and finally (5) the deviation in physical activity in the evening on weekdays. This suggests that people who are more extraverted are, interestingly, more likely to follow a 24-hour rhythm with regards to the environments they spend time in (as captured by ambient sound) as well as their smartphone usage. These participants also tend to have various activities during the evening – similar to the patterns observed for openness.

In addition, we find that five within-person variability features are negatively associated with the conscientiousness trait. These are (1) the deviation in exposure to ambient voice during the daytime (9am-6pm) across all days; (2) the deviation in exposure to human voices during the evening (6pm-12am) across all days; (3) the deviation in exposure to human voices during the daytime (9am-6pm) on weekdays; (4) the deviation in exposure to human voices during the evening (6pm-12am) on weekdays; and finally (5) the deviation of phone usage in the evening periods across all days. These results tell us that more extroverted individuals are less likely to socialize in a changing pattern. Rather, they are more likely to maintain a stable pattern of interpersonal communication over the week, particularly in the evenings. Similarly, this also applies to their phone usage during evening periods.

Agreeableness. We find a large number of within-person variability features are positively associated with the agreeableness trait. These are (1) the range of the flexible regularity index in ambient sound; (2) the range of the flexible regularity index in phone usage on weekdays; (3) the circadian rhythm in physical activity; (4) the flexible regular index in physical activity on weekdays; (5) the flexible regularity index in physical activity across all days; (6) the regularity index in physical activity on weekdays; (7) the range of regularity index; and finally (8) the flexible regularity index in locations on weekdays. These results indicate more agreeable participants are more likely to have to follow a 24-hour rhythm in physical activity. They also tend to have more regular physical activity patterns – based on hour-by-hour comparisons between days. They are more likely to change their ambient sound environment and location routines for some weekdays, while they keep them unchanged on other days.

In addition, we find a number of within-person variability features that are negatively associated with the agreeableness trait. These are (1) the deviation in phone usage during the night (12am-9am) on weekdays; (2) the deviation in physical activity during the night across all days; (3) the deviation in physical activity during the night period on weekdays; (4) the range of the flexible regularity index in physical activity on weekdays; and finally (5) the range of the regularity index in physical activity on weekdays within weekdays, between weekdays and weekends, and across all days. This seems to tell us that people who are agreeable show less deviation in the time they use their phone during the night. They maintain high regularity in physical activity during the study, as indicated by the smaller range and higher average of their measured regular index.

Neuroticism. Neuroticism indicates moodiness and emotional instability. In our study, we do not see any within person variability features significantly associated with the neuroticism trait.

To sum up, the association analysis supports our hypothesis that within-person behavior change patterns derived directly from smartphone sensing data are related to self-reported personality traits. Four of the five personality traits were associated with different types of within-person variability features. However, some features associated with multiple personality traits. For example, the higher deviation in physical activity during the evening period was associated with being more open to new experiences and being more extroverted; the higher regularity in physical activity on weekdays was associated with being more conscientious and more agreeable.

Table 4. Big Five traits in related to regularity using mixed-effect model.

Big Five trait	Association	Related within-person regularity features (N=159)
openness	(+)	sound_std_evening_weekday, stationary_std_evening_all, stationary_std_evening_weekday, **sound_ri_weekday_range**
	(-)	voice_std_day_all, voice_std_day_weekday
conscientiousness	(+)	voice_fri_weekday_avg, lock_circadian_all, lock_circadian_weekday, stationary_fri_weekday_avg
	(-)	voice_std_evening_weekday, location_ri_weekday_vs_weekend_avg, location_ri_all_avg, location_fri_weekday_vs_weekend_avg, location_fri_all_avg
extraversion	(+)	sound_circadian_all, **lock_circadian_all***, lock_circadian_weekday, stationary_std_evening_all, stationary_std_evening_weekday
	(-)	voice_std_day_all, **voice_std_evening_all****, voice_std_day_weekday, **voice_std_evening_weekday****, lock_std_evening_all
agreeableness	(+)	sound_fri_all_range, lock_fri_weekday_range, stationary_circadian_all, **stationary_fri_weekday_avg**, stationary_fri_all_avg, stationary_ri_weekday_avg, location_ri_weekday_range, location_fri_weekday_range
	(-)	lock_std_night_weekday, stationary_std_night_all, **stationary_std_night_weekday**, stationary_fri_weekday_range, stationary_ri_weekday_vs_weekend_range, stationary_ri_weekday_range, stationary_ri_all_range
neuroticism	(+)	
	(-)	
$p < 0.05$; **bold** $p < 0.01$, $^{*}FDR < 0.1$, $^{**}FDR < 0.05$		

5.2 Prediction Analysis

We use Gradient Boosted Regression Trees (GBRT) [20, 55] to predict the self-reported Big Five personality scores. GBRT is an ensemble method that trains and combines several weak regression trees to make more accurate and robust predictions. It builds base estimators (i.e., regression trees) sequentially. Each estimator tries to reduce the bias of the previously combined estimators. By doing so, in each stage a new regression tree is trained on the negative gradient of the loss function. GBRT is less sensitive to outliers and robust to overfitting [17]. Another advantage inherited from the tree based model is that it computes feature importance measures, which can be used for feature selection.

We have a total of 96 features (see Table 1) and a relatively small number of training examples (70 Android and 89 iOS). We reduce the feature space dimensionality using the importance vector generated from GBRT. GBRT computes feature importance by averaging the number of times a particular feature is used for splitting a branch

across the ensemble trees. Higher values correspond to higher importance. We select features with a feature importance value higher than the mean importance iteratively. We repeat this process until we get no more than 9 features. Our heuristic of selecting 9 features is based on experiments in which we find we get higher training errors with a lower or higher threshold.

We train and test models separately among the Android and iOS users. We take this strategy on the basis of our observation that a model trained from the mixed dataset gives poorer predictions, even if we normalize the features separately among Android and iOS users. We believe this is due to the differences (i.e., heterogeneities) between Android and iOS devices, i.e. the accuracy of sensed behaviors may be influenced by the sensors and algorithms on different platforms. For example, activities are derived from the Google Activity Recognition API [24] on Android phones and from the iOS Core motion API [32] on iPhones; the values of sound amplitude in the same environment may be distinct between the two platforms; the conversation classifier may have potentially small differences because of the different platforms (e.g., different microphones and audio APIs). While we have designed and implemented our sensing algorithms on both platforms to take known differences between the iOS and Android platforms into account there are still no widely accepted techniques for equalizing sensing data from these different platforms. Scaling the values to standard normal distributions separately on both platforms does not result in an ideal solution - it violates the observation that Android and iOS users may have different behaviors and phone usage patterns [7, 31, 43, 46, 65]. Adding a binary feature *DeviceType* to control for the platform of the device, does not result in better performance when there are no significant differences for most of the personality traits (Table 3). Specifically, in GBRT the tree grows greedily in a top-down fashion using binary splits. For each tree node, the split minimizing the objective is chosen. The *DeviceType* may not be selected at the root of regression trees due to the similar means of ground truth; i.e., the personality trait scores in the two groups. Taking these challenges into account we opted to divide the data into Android and iOS users. We train different models on these two groups.

Usually, with a small sample set leave-one-out cross validation [30] would be the best option to show the performance of the personality trait prediction model. This technique is widely used in estimating the performance of the model from a small human-centered dataset in existing studies [4, 33, 69, 73]. Leave-one-out has low bias, because each fold uses almost the entire dataset as a training set [38]. However, as a result, the estimation is also very specific for this particular dataset. This could results in high variance compared to the same model's performance on new datasets. Therefore, we also use five-fold cross validation. For each personality trait, we use ten times five-fold cross validation and report the average.

We compare our predictive models with two baseline models. The baseline 1 model takes the average of the scores as the predicted value. This is the most basic regression with only an intercept. The baseline 2 model randomly generates a sample from the already known distribution (i.e., the distribution of the personality trait scores) and uses it as the predicted value. We validate our prediction model using the Mean Absolute Error (MAE), the root mean squared error (RMSE), the Pearson correlation and the R-squared value. MAE and RMSE describe the bias of the predictions; the Pearson's r describes how well the predictions are associated with the ground truth; and the R-squared value measures the goodness of fit by indicating how much of the variance our model explains [12].

Table 5 shows the performance of the prediction model purely based on within-person variability features. The models perform better than the two baselines, and capture considerable variance of the original distribution. The predicted personality score is highly correlated with the ground truth. Our model works better in predicting the extraversion and agreeableness traits. For example, for Android users the leave-one-out model for predicting the extraversion trait achieves 0.55 of MAE, which is 0.24 (30%) lower than the average baseline model and 0.5 (48%) lower than the random baseline model. For iOS users the leave-one-out model for predicting the extraversion trait achieves 0.61 of MAE, which is 0.11 (15%) lower than the average baseline model and 0.39 (39%) lower than

the random baseline model. Our predictive model is less effective at predicting neuroticism, which is in line with the fact that we did not find features associated with neuroticism, as discussed in Section 5.1.

Table 5. Prediction performance

System	Big Five trait	Baseline 1	Baseline 2	Leave one out			5-fold cross validation		
		MAE/RMSE	MAE/RMSE	MAE/RMSE	corr	r^2	MAE/RMSE	corr	r^2
Android	Openness	0.47/0.60	0.68/0.85	0.40/0.51	0.54	0.287	0.41/0.53	0.49	0.235
	Conscientiousness	0.48/0.62	0.69/0.88	0.41/0.54	0.53	0.252	0.45/0.57	0.41	0.141
	Extraversion	0.79/0.92	1.05/1.29	0.55/0.67	0.69	0.465	0.65/0.78	0.53	0.268
	Agreeableness	0.49/0.62	0.70/0.88	0.40/0.49	0.61	0.366	0.40/0.51	0.57	0.223
	Neuroticism	0.66/0.79	0.90/1.11	0.56/0.70	0.46	0.203	0.60/0.73	0.38	0.117
iOS	Openness	0.50/0.62	0.69/0.87	0.45/0.59	0.32	0.072	0.49/0.62	0.20	0.021
	Conscientiousness	0.50/0.65	0.72/0.91	0.46/0.59	0.42	0.168	0.46/0.59	0.40	0.143
	Extraversion	0.72/0.88	1.00/1.24	0.61/0.76	0.51	0.254	0.65/0.80	0.42	0.161
	Agreeableness	0.48/0.63	0.69/0.88	0.39/0.50	0.60	0.358	0.40/0.51	0.56	0.311
	Neuroticism	0.64/0.79	0.90/1.12	0.59/0.73	0.40	0.158	0.60/0.77	0.33	0.100

Fig. 4 illustrates the predicted personality trait score (y-axis) and the self-reported ground truth (x-axis) for each participant using the leave-one-out method. The blue line indicates the ideal model where the predicted value is equal to the ground truth. The red points above the blue line are overestimated and the points below it are underestimated. Since our training set is normally distributed, the model gets reinforced in the center area and has bigger absolute errors on the two sides. Even though, we see that the model can still capture the trend and variance of the ground truth distribution. Because traits measured using the Big Five Inventory are designed to be treated as continuous variables, we do not conduct binary classifications. However, the plot indicates that if we did perform a binary classification to distinguish people with higher or lower scores on some traits, we would also achieve good performance results, particularly when predicting the extraversion and agreeableness traits.

6 LIMITATIONS

The current study has a number of limitations that need to be addressed in future research. First, further research is needed to show the data collected from Android and iOS devices can be correctly merged. In our prediction model we take a conservative approach and separate out the Android and iOS groups. We find a model trained from the mixed dataset gives worse prediction, even if we normalize the features separately among Android and iOS users. This could be because of the accuracy of sensed behaviors are influenced by the sensors and algorithms on different platforms, as discussed earlier in the paper. If so, work is needed to mitigate the impairments caused by features collected on different platforms. This is an interesting and important area of research in passive sensing on different devices and bench marking norms between devices. Another possible explanation is that there are real differences between users who select different platforms; that is, there could exist different baseline personality trait expressions across the different user groups - which might in turn be explained by advertising strategies, pricing and the brand personalities of the companies behind the respective operating systems and major phone manufacturers using them. However, further work is needed to explore to what extent more fundamental differences in the operating systems, and how they get used by and interact with the phone users influence these differences, as well as whether the same disparities are observed across samples in other, non-student populations.

Second, further work is needed to explore which sensor-based features are the strongest predictors of personality traits. There might be some other powerful features which can better represent the changes in lifestyle that

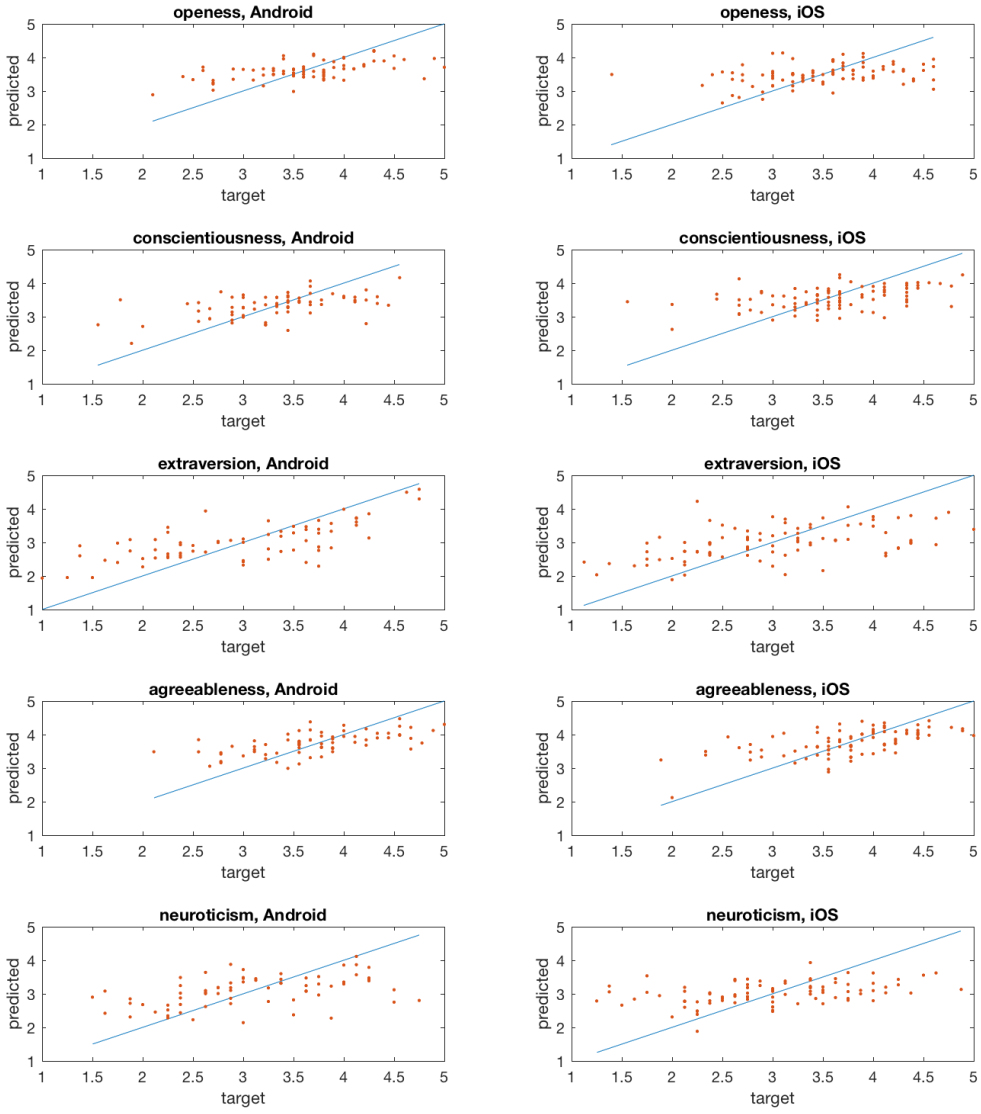


Fig. 4. Predicted values of personality traits and ground truth.

is more predictable in estimating personality other than the metrics we have used: that is, standard deviation,

circadian rhythm, and two regularity indexes. As sensing devices become more powerful and additional sensors become available, the research on passive personality assessment may identify other variability features that are powerful predictors of behavior and personality.

Third, it is unclear how sensing restrictions on the iOS platform influenced our sample. The original sample consisted of 646 students, out of which 117 (18%) were Android users and 529 (82%) reported being iOS users. However, only about a quarter – 159 participants – could be included in the analyses presented in this paper – that is participants that had more than 7 days of over 19 hours of sensing data. This is a conservative inclusion criteria we used in prior studies. Of those, 70 (44%) were Android users and 89 (56%) were iOS users. It is unclear to what extent the comparatively small subset of iOS users with sufficient amounts of sensor data compare to their excluded peers. Typically, Android is more open to continuous passive sensing even though new versions of the Android OS are placing more restrictions. Apple's iOS has been more closed to continuous passive sensing by imposing restrictions on sampling rates and access to sensing data. As a result the more restrictive iOS environment limits data gathering. In our study, some iOS users reported the app was inadvertently terminated requiring us to troubleshoot the issues throughout the data collection process. Despite these challenges, we are confident that the actions we took to mitigate problems (e.g., conservative inclusion criteria, matching data quality collected across platforms, balance of iOS and Android users) has lead to a good quality dataset.

Fourth, further work is required to balance the accuracy and resources usage. We collected an extensive amount of data via passive sensing for capturing as various and accurate behaviors as possible. According to our survey [29] after this study, participants were satisfied with the self-tracking assignment using mobile phones. The average levels of satisfaction were 3.70 (Android) and 3.92 (iOS), respectively (from 1 Very unsatisfied to 5 Excellent). Besides, 61% participants reported they did not feel uncomfortable using the app at all. However, 53% Android and 28% iOS users noticed the draining phone batteries, which indicated the biggest obstacle of allowing the adoption in real life scenarios. As a follow-up, we tested the power consumption by turning on and off each sensing component (i.e, activity detection, accelerometer, in-situ voice classifier, GPS location and scheduled data uploading) on factory-reset Android and iOS phones. Our tests show that the voice classifier and collecting raw accelerometer data are the major causes of energy cost. We decide to decrease the resource usage by stopping collecting the accelerometer data and lowering the duty cycle of voice classifier. The accelerometer data is less useful provided that we have already obtained the activity inferences (i.e., stationary, walking, running, cycling, in vehicle). The new microphone duty cycle is 1 minute (if no conversation detected) up to 3 (if conversations detected) minutes on and 9 minutes off. Based on these adjustments, we have significantly improved the energy efficiency of the sensing system to support long-term studies. The new system is now being used in on-going 4-year study. In the new study, the participant are satisfied with the battery consumption. We believe we can better understand the trade-off between the accuracy and resources usage as the new project progresses.

While our results are statistically significant and encouraging they are limited to students at UT Austin. The length of the study is only 14 days. We acknowledge this is the first step in this area of research and that more is needed to push forward our understanding of the importance of within-person variability. We welcome researchers to use our within-person features in their studies, and encourage others to conduct similar studies at different sites with different populations to examine whether the findings replicate and are generalizable.

7 CONCLUSION AND FUTURE WORK

Personality traits describe people's characteristic patterns of thinking, feeling, and behaving. As such, personality traits describe patterns of variability 'between-persons' - that is, patterns of behavior that distinguish people from one another. Personality states, on the other hand, refer to patterns of variability 'within-persons'. Within-person variability describes fluctuations in how a person's thinking, feeling, and behaving changes over time. This research reports a mobile sensing approach to assess within-person behavior variability, and thus explores

how within-person variability patterns can be used to predict personality traits using features derived from mobile sensing. Past research has shown within-person behavior variability to be linked to various psychological characteristics. However, much of the past research relies on a person's capacity to accurately recall their daily experiences retrospectively. Although some researchers have utilized information from smartphones in personality prediction, most work focuses on between-person variability and usually only includes Android participants. Other approaches that use texts on social networks for personality prediction require access to extensive information about people's online social networks. We designed and implemented a cross-platform mobile sensing study to capture the within-person variability among college students in their social interactions, mobility and movement, daily activities and situational/environmental information. We demonstrate how within-person variability patterns measured by smartphone sensing are related to and thus can be used to predict self-reported personality traits. Our results show that our proposed personality prediction model based on within-person variability features provides good estimation of personality traits, particularly for extraversion and agreeableness. This is, to the best of our knowledge, the first scaled study investigating how within-person variability is predictive of personality traits. It complements and extends existing methods, providing researchers with an additional measure that assesses large groups of participants with minimal burden.

This work represents an important first step toward passive personality assessment. There is a need for the community interested in personality prediction to take the next step and conduct a large scale, longitudinal study with a diverse cohort (e.g., including students, working adults, the elderly). As a contribution, our system can be easily deployed to collect the necessary mobile sensing data for behavioral tracking in new personality-related projects. It has some potential applications. First, a hybrid approach that combines mobile sensing data with the content available on social networking such as Facebook and Twitter would likely improve predictive performance. Second, the sensing system offers a practical alternative for passive personality assessment, allowing assessment of psychological characteristics in large-scale applications when questionnaires are impractical. The goal of such an approach is to achieve personality assessment without any human intervention. Such an assessment technique could be widely used in recommendation systems, recruiting procedure, for target filtering and for many human-centered applications. Finally, using mobile sensing for measuring within-person variability in behavioral patterns is not limited to personality. This method can be adapted for use in other research areas, such as those focused on prediction of psychological well-being, mental health, or even workplace performance.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation (NSF) Award BCS-1520288.

REFERENCES

[1] Saeed Abdullah, Mark Matthews, Ellen Frank, Gavin Doherty, Geri Gay, and Tanzeem Choudhury. 2016. Automatic detection of social rhythms in bipolar disorder. *Journal of the American Medical Informatics Association* 23, 3 (2016), 538–543.

[2] Saeed Abdullah, Mark Matthews, Elizabeth L Murnane, Geri Gay, and Tanzeem Choudhury. 2014. Towards circadian computing: early to bed and early to rise makes some of us unhealthy and sleep deprived. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 673–684.

[3] Juan Manuel Antúnez, José Francisco Navarro, and Ana Adan. 2014. Morningness–eveningness and personality characteristics of young healthy adults. *Personality and Individual Differences* 68 (2014), 136–142.

[4] Olivier Augereau, Kai Kunze, Hiroki Fujiyoshi, and Koichi Kise. 2016. Estimation of english skill with a mobile eye tracker. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 1777–1781.

[5] Min S Hane Aung, Faisal Alquaddoomi, Cheng-Kang Hsieh, Mashfiqui Rabbi, Longqi Yang, John P Pollak, Deborah Estrin, and Tanzeem Choudhury. 2016. Leveraging multi-modal sensing for mobile health: a case review in chronic pain. *IEEE Journal of Selected Topics in Signal Processing* 10, 5 (2016), 962–974.

[6] Dror Ben-Zeev, Rachel Brian, Rui Wang, Weichen Wang, Andrew T Campbell, Min SH Aung, Michael Merrill, Vincent WS Tseng, Tanzeem Choudhury, Marta Hauser, et al. 2017. CrossCheck: Integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse. *Psychiatric Rehabilitation Journal* 40, 3 (2017), 266.

[7] Zinaida Benenson, Freya Gassmann, and Lena Reinfelder. 2013. Android and iOS users' differences concerning security and privacy. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. ACM, 817–822.

[8] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* (1995), 289–300.

[9] Yoav Benjamini and Daniel Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* (2001), 1165–1188.

[10] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Sandy Pentland. 2014. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 477–486.

[11] Sarah Butt and James G Phillips. 2008. Personality and self reported mobile phone use. *Computers in Human Behavior* 24, 2 (2008), 346–360.

[12] A Colin Cameron and Frank AG Windmeijer. 1996. R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics* 14, 2 (1996), 209–220.

[13] Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 1293–1304.

[14] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. 2013. Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing* 17, 3 (2013), 433–450.

[15] Yves-Alexandre de Montjoye, Jordi Quoidbach, Florent Robic, and Alex Sandy Pentland. 2013. Predicting personality using novel mobile phone-based metrics. In *International conference on social computing, behavioral-cultural modeling, and prediction*. Springer, 48–55.

[16] Rodrigo de Oliveira, Alexandros Karatzoglou, Pedro Concejero Cerezo, Ana Armenta Lopez de Vicuña, and Nuria Oliver. 2011. Towards a psychographic user model from mobile phone usage. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2191–2196.

[17] Jane Elith, John R Leathwick, and Trevor Hastie. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology* 77, 4 (2008), 802–813.

[18] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *Kdd*, Vol. 96. 226–231.

[19] William Fleeson and Patrick Gallagher. 2009. The implications of Big Five standing for the distribution of trait manifestation in behavior: fifteen experience-sampling studies and a meta-analysis.

[20] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[21] Asma Ghandeharioun, Szymon Fedor, Lisa Sangermano, Dawn Ionescu, Jonathan Alpert, Chelsea Dale, David Sontag, and Rosalind Picard. 2017. Objective Assessment of Depressive Symptoms with Machine Learning and Wearable Sensors Data. In *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, Texas*.

[22] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, 149–156.

[23] Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011. Predicting personality with social media. In *CHI'11 extended abstracts on human factors in computing systems*. ACM, 253–262.

[24] Google Activity Recognition Api. 2017. Google Activity Recognition Api. https://developers.google.com/android/reference/com/google/android/gms/location/ActivityRecognitionClient.

[25] Friedrich M Götz, Stefan Stieger, and Ulf-Dietrich Reips. 2017. Users of the main smartphone operating systems (iOS, Android) differ only little in personality. *PloS one* 12, 5 (2017), e0176921.

[26] Gabriella M Harari, Samuel D Gosling, Rui Wang, Fanglin Chen, Zhenyu Chen, and Andrew T Campbell. 2017. Patterns of behavior change in students over an academic term: A preliminary study of activity and sociability behaviors using smartphone sensing methods. *Computers in Human Behavior* 67 (2017), 129–138.

[27] Gabriella M Harari, Sandrine R Müller, Min SH Aung, and Peter J Rentfrow. 2017. Smartphone sensing methods for studying behavior in everyday life. *Current Opinion in Behavioral Sciences* 18 (2017), 83–90.

[28] Gabriella M Harari, Sandrine R Müller, Varun Mishra, Rui Wang, Andrew T Campbell, Peter J Rentfrow, and Samuel D Gosling. 2017. An Evaluation of Students' Interest in and Compliance With Self-Tracking Methods: Recommendations for Incentives Based on Three Smartphone Sensing Studies. *Social Psychological and Personality Science* 8, 5 (2017), 479–492.

[29] Gabriella M Harari, Weichen Wang, Sandrine R Müller, Rui Wang, and Andrew T Campbell. 2017. Participants' compliance and experiences with self-tracking using a smartphone sensing app. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. ACM, 57–60.

[30] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. Unsupervised learning. In *The elements of statistical learning*. Springer, 485–585.

[31] Todd Hixon. 2014. What Kind Of Person Prefers An iPhone? https://www.forbes.com/sites/toddhixon/2014/04/10/what-kind-of-person-prefers-an-iphone/#2a55c41fd1b0
[32] iOS Core Motion. 2017. iOS Core Motion. https://developer.apple.com/documentation/coremotion.
[33] Xinlong Jiang, Yiqiang Chen, Junfa Liu, Gillian R Hayes, Lisha Hu, and Jianfei Shen. 2016. AIR: recognizing activity through IR-based distance sensing on feet. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 97–100.
[34] Oliver P John and Sanjay Srivastava. 1999. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research* 2, 1999 (1999), 102–138.
[35] Martina K Kanning, Ulrich W Ebner-Priemer, and Wolfgang Michael Schlicht. 2013. How to investigate within-subject associations between physical activity and momentary affective states in everyday life: a position statement based on a literature overview. *Frontiers in psychology* 4 (2013).
[36] Jung-Hyun Kim. 2017. Smartphone-mediated communication vs. face-to-face interaction: Two routes to social support and problematic use of smartphone. *Computers in Human Behavior* 67 (2017), 282–291.
[37] Willard L Koukkari and Robert B Sothern. 2007. *Introducing biological rhythms: A primer on the temporal organization of life, with implications for health, society, reproduction, and the natural environment*. Springer Science & Business Media.
[38] Peter A Lachenbruch and M Ray Mickey. 1968. Estimation of error rates in discriminant analysis. *Technometrics* 10, 1 (1968), 1–11.
[39] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. 2010. A survey of mobile phone sensing. *IEEE Communications magazine* 48, 9 (2010).
[40] Nicholas D Lane, Mashfiqui Mohammod, Mu Lin, Xiaochao Yang, Hong Lu, Shahid Ali, Afsaneh Doryab, Ethan Berke, Tanzeem Choudhury, and Andrew Campbell. 2011. Bewell: A smartphone application to monitor, model and promote wellbeing. In *5th international ICST conference on pervasive computing technologies for healthcare*. 23–26.
[41] Yu-Kang Lee, Chun-Tuan Chang, You Lin, and Zhao-Hong Cheng. 2014. The dark side of smartphone usage: Psychological traits, compulsive behavior and technostress. *Computers in Human Behavior* 31 (2014), 373–383.
[42] Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10. 707–710.
[43] Yao Liu, Fei Li, Lei Guo, Bo Shen, and Songqing Chen. 2013. A comparative study of android and iOS for accessing internet streaming services. In *International Conference on Passive and Active Network Measurement*. Springer, 104–114.
[44] Yong Liu, Jayant Venkatanathan, Jorge Goncalves, Evangelos Karapanos, and Vassilis Kostakos. 2014. Modeling what friendship patterns on Facebook reveal about personality and social capital. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 3 (2014), 17.
[45] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T Chittaranjan, Andrew T Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. 2012. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 351–360.
[46] Chris Matyszczyk. 2011. Study: Android users sad hicks, iPhone users rich girls. https://www.cnet.com/news/study-android-users-sad-hicks-iphone-users-rich-girls/
[47] Bryan P Mccormick, Georgia Frey, Chien-Tsung Lee, Sanghee Chun, Jim Sibthorp, Tomislav Gajic, Branka Stamatovic-Gajic, and Milena Maksimovich. 2008. Predicting transitory mood from physical activity level among people with severe mental illness in two cultures. *International Journal of Social Psychiatry* 54, 6 (2008), 527–538.
[48] Robert R McCrae and Paul T Costa. 1995. Trait explanations in personality psychology. *European Journal of Personality* 9, 4 (1995), 231–252.
[49] Charles E McCulloch and John M Neuhaus. 2001. *Generalized linear mixed models*. Wiley Online Library.
[50] Matthias R Mehl, Samuel D Gosling, and James W Pennebaker. 2006. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology* 90, 5 (2006), 862.
[51] Abhinav Mehrotra and Mirco Musolesi. 2017. Designing Effective Movement Digital Biomarkers for Unobtrusive Emotional State Mobile Monitoring. (2017).
[52] Timothy H Monk, Ellen Frank, Jaime M Potts, and David J Kupfer. 2002. A simple way to measure daily lifestyle regularity. *Journal of sleep research* 11, 3 (2002), 183–190.
[53] Timothy K Monk, Joseph F Flaherty, Ellen Frank, Kathleen Hoskinson, and David J Kupfer. 1990. The Social Rhythm Metric: An instrument to quantify the daily rhythms of life. *Journal of Nervous and Mental Disease* (1990).
[54] Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology* 108, 6 (2015), 934.
[55] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, Oct (2011), 2825–2830.
[56] William H Press. 2007. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press.

[57] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 2011. Our twitter profiles, our selves: Predicting personality with twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, 180–185.
[58] Mashfiqui Rabbi, Shahid Ali, Tanzeem Choudhury, and Ethan Berke. 2011. Passive and in-situ assessment of mental and physical well-being using mobile sensors. In *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 385–394.
[59] Christoph Randler, Verena Petra Baumann, and Mehmet Barış Horzum. 2014. Morningness–eveningness, Big Five and the BIS/BAS inventory. *Personality and Individual Differences* 66 (2014), 64–67.
[60] Ryan E Rhodes. 2006. The built-in environment: The role of personality and physical activity. *Exercise and Sport Sciences Reviews* 34, 2 (2006), 83–88.
[61] Sohrab Saeb, Emily G Lattie, Stephen M Schueller, Konrad P Kording, and David C Mohr. 2016. The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ* 4 (2016), e2537.
[62] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research* 17, 7 (2015).
[63] Jacopo Staiano, Bruno Lepri, Nadav Aharony, Fabio Pianesi, Nicu Sebe, and Alex Pentland. 2012. Friends don't lie: inferring personality traits from social network structure. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. ACM, 321–330.
[64] Timothy J Trull and Ulrich Ebner-Priemer. 2013. Ambulatory assessment. *Annual review of clinical psychology* 9 (2013), 151–176.
[65] Harveen Kaur Ubhi, Daniel Kotz, Susan Michie, Onno CP van Schayck, and Robert West. 2017. A comparison of the characteristics of iOS and Android users of a smoking cessation app. *Translational behavioral medicine* 7, 2 (2017), 166–171.
[66] Jayant Venkatanathan, Evangelos Karapanos, Vassilis Kostakos, and Jorge Gonçalves. 2012. Network, personality and social capital. In *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, 326–329.
[67] Hui Wang and Stacy Marsella. 2017. Assessing personality through objective behavioral sensing. In *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*. IEEE, 131–137.
[68] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 3–14.
[69] Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T Campbell. 2015. SmartGPA: how smartphones can assess and predict academic performance of college students. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 295–306.
[70] Rui Wang, Weichen Wang, Min SH Aung, Dror Ben-Zeev, Rachel Brian, Andrew T Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A Scherer, et al. 2017. Predicting Symptom Trajectories of Schizophrenia using Mobile Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 110.
[71] Deborah J Wiebe and Timothy W Smith. 1997. Personality and health: Progress and problems in psychosomatics. (1997).
[72] Wu Youyou, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences* 112, 4 (2015), 1036–1040.
[73] Yang Zhang, Robert Xiao, and Chris Harrison. 2016. Advancing hand gesture recognition with high resolution electrical impedance tomography. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 843–850.

Received November 2017; revised May 2018; accepted September 2018