Person-Centered Predictions of Psychological Constructs with Social Media Contextualized by Multimodal Sensing

KOUSTUV SAHA*, Georgia Institute of Technology, USA
TED GROVER*, University of California, Irvine, USA
STEPHEN M. MATTINGLY, University of Notre Dame, USA
VEDANT DAS SWAIN, Georgia Institute of Technology, USA
PRANSHU GUPTA, Georgia Institute of Technology, USA
GONZALO J. MARTINEZ, University of Notre Dame, USA
PABLO ROBLES-GRANDA, University of Notre Dame, USA
GLORIA MARK, University of California, Irvine, USA
AARON STRIEGEL, University of Notre Dame, USA
MUNMUN DE CHOUDHURY, Georgia Institute of Technology, USA

Personalized predictions have shown promises in various disciplines but they are fundamentally constrained in their ability to generalize across individuals. These models are often trained on limited datasets which do not represent the fluidity of human functioning. In contrast, generalized models capture normative behaviors between individuals but lack precision in predicting individual outcomes. This paper aims to balance the tradeoff between one-for-each and one-for-all models by clustering individuals on mutable behaviors and conducting cluster-specific predictions of psychological constructs in a multimodal sensing dataset of 754 individuals. Specifically, we situate our modeling on social media that has exhibited capability in inferring psychosocial attributes. We hypothesize that complementing social media data with offline sensor data can help to personalize and improve predictions. We cluster individuals on physical behaviors captured via Bluetooth, wearables, and smartphone sensors. We build *contextualized* models predicting psychological constructs trained on each cluster's social media data and compare their performance against *generalized* models trained on all individuals' data. The comparison reveals no difference in predicting affect and a decline in predicting cognitive ability, but an improvement in predicting personality, anxiety, and sleep quality. We construe that our approach improves predicting psychological constructs sharing theoretical associations with physical behavior. We also find how social media language associates with offline behavioral contextualization. Our work bears implications in understanding the nuanced strengths and weaknesses of personalized predictions, and how the effectiveness may vary by multiple factors. This work reveals the importance of taking a critical stance on evaluating the effectiveness before investing efforts in personalization.

*These authors contributed equally to this work.

Authors' addresses: Koustuv Saha, Georgia Institute of Technology, Atlanta, GA, USA, koustuv.saha@gatech.edu; Ted Grover, University of California, Irvine, Irvine, CA, USA, grovere@uci.edu; Stephen M. Mattingly, University of Notre Dame, South Bend, IN, USA, smattin1@nd.edu; Vedant Das swain, Georgia Institute of Technology, Atlanta, GA, USA, vedantswain@gatech.edu; Pranshu Gupta, Georgia Institute of Technology, Atlanta, GA, USA, prangupt@gatech.edu; Gonzalo J. Martinez, University of Notre Dame, South Bend, IN, USA, gmarti11@nd.edu; Pablo Robles-Granda, University of Notre Dame, South Bend, IN, USA, problesg@nd.edu; Gloria Mark, University of California, Irvine, Irvine, CA, USA, gmark@uci.edu; Aaron Striegel, University of Notre Dame, South Bend, IN, USA, striegel@nd.edu; Munmun De Choudhury, Georgia Institute of Technology, Atlanta, GA, USA, munmund@gatech.edu.

2474-9567/2021/0-ART0 $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

CCS Concepts: • **Human-centered computing** → *Empirical studies in ubiquitous and mobile computing*; *Empirical studies in collaborative and social computing*; *Social media*; • **Applied computing** → *Psychology*.

Additional Key Words and Phrases: social media; multimodal sensing; person-centered; personalization; machine learning; clustering; personality traits; affect; cognitive ability; sleep; language

ACM Reference Format:
Koustuv Saha, Ted Grover, Stephen M. Mattingly, Vedant Das swain, Pranshu Gupta, Gonzalo J. Martinez, Pablo Robles-Granda, Gloria Mark, Aaron Striegel, and Munmun De Choudhury. 2021. Person-Centered Predictions of Psychological Constructs with Social Media Contextualized by Multimodal Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 0, 0, Article 0 (2021), 32 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In the past few years, research has developed several passive sensing approaches to improve our understanding of human behavior both longitudinally and scalably. Simultaneously, research has utilized ubiquitous social media platforms as a "passive sensor" [106] and an unobtrusive source of behavioral data, which is self-recorded and self-initiated by individuals in naturalistic settings. Because this data contains language and social interactions, it is a unique form of *verbal* and *social* sensor, unlike several physical sensing modalities. A large body of research reveals the potential of inferring psychological constructs with social media [29, 46, 112].

However, social media data may not include the valuable contextual information that drives posting and presentation behaviors. For instance, even within the same emotional state, individual Facebook posting behavior varies [73]. This interaction of various factors underscores the idea of the Social Ecological Model [15] in which psychological constructs are embedded in a complex interplay between individual, social, and environmental factors. In fact, even posting (or not posting) can be dictated by external factors that vary for every person. Therefore, social media based sensing is unique in its sensitivity to factors driving an individual's self- initiation, motivation, and presentation. This between-person variability in data may impact predictions of an individual's underlying psychology, routines, and other personal attributes. Incorporating additional offline context that captures factors affecting online behavior could boost the ability of social media to predict individual outcomes.

Personalizing, where models are tuned and optimized for each individual [102] can overcome between-subject variability. Indeed, personalized modeling methods are gaining attention in the disciplines of social science, psychology, and health, including precision medicine and digital phenotyping [54, 56, 63, 82, 87]. Person-centered methods can glean a more comprehensive understanding of an individual, and in some cases explain their outcomes better than variable-centered or generalized methods (i.e. focusing on global variables that are measured through the same means for everyone in a target population) [70, 103, 135]. A variety of personalized predictions have also been conducted in computing, particularly in content recommendation and data mining [115]. Drawing on such approaches, we can consider predictions by building individual-level models. However, such approaches would be impeded due to the temporal sparsity of social media data (because individuals post on social media only at discrete intervals). Alternatively, we can consider stratifying individuals on demographic attributes such as age, race, and gender. However, these attributes are not only privacy-intrusive but are also static and exclusionary. Demographic attribute-based stratified modeling has been identified by the Fairness, Accountability, Transparency literature to reinforce stereotypes and existing societal biases, and even exacerbate them [53, 58, 96]. Additionally, these approaches may not have sufficient data for a particular demographic or marginalized group. Using dynamic features shared more broadly can be a better alternative.

This work avoids demographic based and personalized modeling shortcomings by embracing multimodal sensing in capturing behavior and context, in the form of "small data" about a person [37]. In particular, we propose a person-centered approach similar to computational phenotyping (in which machine learning identifies relevant predictors of an outcome shared across individuals), that leverages passively collected dynamic attributes spanning phone use, physical activities, mobility, and work behaviors. Data is collected from Bluetooth beacons, smartphones, and wearables. We then obtain clusters (or groups) of individuals who demonstrate

similar combinations of multidimensional offline behaviors. Clustering individuals is theoretically motivated in that people's offline behaviors drive online (or social media posting) behaviors and vice-versa. Clustering also serves to capture both within-individual heterogeneity and between-individual homogeneity. This approach is a middle-ground between "one-for-each" and "one-for-all" models, thereby aiming to balance the drawbacks of extremely personalized (one or few individuals per model) which may not generalize or consolidate findings across individuals, and extremely generalized models, which face difficulty in generating precise predictions per individual and can also be impeded by variability in individual data quality and completeness. Our work draws motivation from clustering based approaches previously adopted in digital phenotyping research with electronic health records to identify comorbidity of symptoms [33] and to compare and summarize clinical models [45].

We hypothesize that contextualizing on offline and naturalistic behaviors can provide a degree of personalization and improve predicting psychological constructs with social media. Combining multiple sensing modalities in this manner can allow us to leverage complementary strengths of different sensing techniques. Additionally, this approach can provide a theoretical lens of understanding the interaction between offline and online behaviors that is useful in both research and in practice. This paper, therefore, targets the below *research aims*:

Aim 1: To predict psychological constructs with social media in a person-centered approach of contextualizing people's offline physical behaviors.
Aim 2: To evaluate how contextualized predictions compare against generalized prediction models.
Aim 3: To examine how social media language associates with offline behavioral contextualization.

This paper uses data from the Tesserae project [75], a year-long multisensor study of 757 individuals, where 572 provided social media (Facebook) data. Consented participants provided self-reported measures of psychological constructs of *cognitive ability, personality traits, affect*, and *wellbeing*, which serve as ground-truth in this study. We use this data to achieve the *aims* above, through three-fold contributions:

First, our work contributes an approach of building *contextualized* person-centered models that predict psychological constructs from naturalistic passive data describing a multitude of contextual factors. We build *contextualized* models trained on each cluster's social media data and compare the performance against *generalized* models trained on the entire social media dataset of all participants, as is typically done.

Second, our work provides insights about the relative performance of predicting psychological constructs with generalized and contextualized models. Our evaluations reveal that contextualized predictions show a significant increase in predicting anxiety, sleep, and personality traits. However, we find no significant difference in predicting affect, whereas a significant decrease in predicting cognitive ability.

Third, we critically discuss the tradeoff between personalization and statistical power, and the importance of evaluating the costs and benefits of personalizations as implications in research and practice. We construe that the utility of contextualizing on offline behavior for social media based predictions relies on the strength of the theoretical associations between a construct of interest and offline manifestations of the construct. Additionally, personalized models are not only costly but may also be impacted by the limitations associated with smaller training data sizes compared to generalized models. Theoretically, our work can be useful in behavioral modeling in emergent fields like human-centered machine learning, as well as to generate hypotheses for future investigations that leverage the relationship between passively sensed behavior and psychological constructs.

2 BACKGROUND AND RELATED WORK

2.1 Assessing Psychological Constructs

Traditionally, psychological constructs such as personality traits, affect, anxiety, and mood have been captured using interviews and survey instruments [22]. These survey instruments are psychometrically validated and act as "gold standard" measurements for capturing quantifiable construct scores. However, these approaches are highly reliant on a respondent's retrospective recall and subjective assessment, can be subject to bias, and also can be costly and burdensome to distribute frequently and at scale [42, 43].

To mitigate some of the confounds of static long-form survey instruments, ecological momentary assessments (EMA) (also termed as experience sampling) have been adopted as a methodology over the past few decades [114, 122]. In this technique, participants are prompted to respond to survey items in-the-moment and within their natural context. EMAs have many advantages over traditional research designs for characterizing complex psychological processes [122]. With the advances in mobile active-sensing technologies (e.g. smartphones), EMAs can be now be conducted at scale. Accordingly, EMAs are now extensively used in ubiquitous computing research. For example, the photographic affect meter [90], an EMA that records in-the-moment affective states, has been incorporated as a mobile EMA and used in multiple multimodal sensing studies [10, 17, 93, 130]. Although better than static survey instruments on many fronts, active sensing comes with limitations of scale, access, and affordance [114]. The EMA methods often disseminated through prompts induce a response burden on participants through disruptions [123]. This requires a balance between the construct validity of distributing short survey items responses with participant compliance [17, 44].

Consequently, unobtrusive and low burden passively sensed data has emerged to complement active sensing data. The ubiquity and widespread use of smartphones and wearables allow researchers to capture longitudinal and dense human behavior at scale [130, 131]. Recent work by Stachl et al. on predicting Big 5 personality traits using passive sensing data from smartphones with a large participant sample (n = 642). However, such data collection can only be done prospectively, or only during the participation period of studies. To obtain historical or before-study data of individuals, researchers have recently started to leverage social media as a passive sensing modality, which enables unobtrusive data collection of longitudinal and historical data of individuals that is self-recorded in their naturalistic settings [106].

Motivated by the above literature, we combine the complementary strengths of multimodal sensing to understand psychological constructs. Our work builds prediction models specific to clusters of individuals based on offline behaviors. Such an approach provides theoretical benefits of understanding how different offline behaviors are associated with online (social media) behaviors.

2.2 Social Media as a Passive Sensing Modality

A rich body of research has demonstrated that social media technologies provide several benefits as a passive sensing modality [23, 29, 76, 106]. Social media is low-cost, large-scale, non-intrusive to collect, and can comprehensively reveal naturalistic patterns of mood, behavior, cognition, psychological states and social milieu [46, 66, 113]. Prior work has leveraged social media data at scale to quantitatively identify mental health attributes such as mood, stress, suicidal ideation, depressive and psychotic symptoms [21, 29, 36, 106]. Culotta studied that social media data can be useful in predicting county-level health and wellness metrics [23].

In this regard, social media language based research builds on the vast success of psycholinguistics research, notably by Pennebaker and colleagues [19, 88, 124], who have studied how language reveals psychological markers in a variety of states and contexts. The same group of authors also developed the Linguistic Inquiry and Wordcount (LIWC) — a lexicon of linguistic markers grounded with psychometric validity [124], which has also been found to work well on short texts and social media data [23, 29, 36, 106]. Prior work has also harnessed social media for analyzing personality traits and their relationship to psychosocial wellbeing, through machine learning and linguistic analysis [48, 91, 92, 109, 112]. Schwartz et al. revealed that social media language predicts individual wellbeing [113], and Kosinski et al. predicted a range of sensitive personal attributes including sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender by simply using "likes" on Facebook [62]. The authors used social media data to build predictive models of ground truth psychological construct scores obtained from psychometrically validated survey instruments administered to participants. In the context of predicting Big Five personality traits [117], which has been investigated extensively in this domain, social media data driven predictions have yielded some of the most precise automatic predictions compared to other data modalities [86].

However, each stream has limitations. In the case of social media, it suffers from data sparsity, and not everyone is *equally active* on it [106, 137]. Therefore, the variability in the use of social media across individuals may impact the predictive capabilities of models built on "all" individuals' datasets — e.g., some features may have high-variance, some features' effects may be washed out, and some features may be downplayed by other features, although these could bear significant signals for certain cohorts of individuals. This paper addresses the above gap by proposing a methodology to adapt predictions on social media data per cohorts of "similar" individuals on the basis of their offline and physical behavior (captured using in-the-wild sensing technologies). Our work is motivated by the recent success of machine learning approaches in the ubiquitous computing and multimodal sensing studies to infer latent human behavioral attributes [30, 95, 99, 136]. Further, if certain offline behaviors mirror online behaviors [14, 40, 130], then social media data when contextualized with offline behaviors should improve predictions of the constructs that share strong relationships with offline behaviors. We examine this by looking across a range of psychological constructs of cognitive ability, personality traits, affect, and wellbeing.

2.3 Need for Person-Centered Predictions of Psychological Constructs

In recent years, researchers in psychology and clinical assessment have stressed a greater need to better understand the individual variability and context surrounding psychological constructs and how they are measured [63, 135]. Hekler et al. elaborated how new technologies, methods, and research frameworks from social and behavioral sciences can play a key role in precision health [54]. The Social Ecological Model posits that psychological constructs are deeply embedded in the complex interplay between an individual, their relationships, their context, and their behaviors and experiences [15]. Research into factors that predict constructs by focusing on both general global variable focused analyses and more person-centric analyses [56] found that the predictors from both analysis frameworks can sometimes be quite different, each contributing novel insights into the behavioral predictors of the construct in the social context in which they are measured. Laursen et al. found that global variable measurements of social support in various relationships influenced different dimensions of interpersonal competence independently, however person-centered analyses found that social-support measurements in relationships have an aggregate effect on all dimensions of inter-personal competence [63]. Howard and Hoffman further noted that variable-centered, person-centered, and person-specific (fully individualized) approaches in social science are not necessarily competitive, rather, should be considered complementary approaches in methodological, statistical, and theoretical differences [56].

In the area of diagnosing and detecting psychopathology, recent research has suggested that general psychiatric evaluations may often do a poor job at describing an individual patient's psychopathology, especially when an individual presents many heterogeneous issues and experiences of symptoms [135], and that more personalized (i.e. idiographic) statistical models have shown promise in modeling a specific person's psychopathology and generating best-fit interventions [41, 135]. Computational phenotyping, a set of parameters, derived from neural and behavioral data, which characterize an individual's cognitive mechanisms, has been shown to improve our understanding of personality, development, fluid intelligence, and psychopathology [87]. Recent advancements gained of fine-grained temporal resolution with passive data collection, and automated statistical procedures have enabled exploration into the utility of personalized models in these domains [24, 26, 51, 64, 65, 97, 102].

In the area of predicting psychological constructs from multimodal sensing, the traditional approach has been to build generalized (or variable-centered) models that combine all behavioral features derived from sensing modalities into a single model predicting outcome variables (e.g. [1, 10, 131]. However, the promise of more personalized models encourages exploring this approach in multimodal sensing. For instance, significant research has focused on building models to predict depression from individuals' social media behavior [13, 29, 67]. On clinical psychometric surveys to measure depressive symptoms, e.g. the Center for Epidemiological Studies Depression Scale (CES-D) [94], certain items are relatively and intentionally vague like "I felt sad". However, the behavioral context that contributes to a subjective rating of sadness likely varies between individuals — for some,

a lack of interpersonal interaction may be the strongest conscious or unconscious factor contributing towards sadness, while for others a lack of mobility and sleep may be stronger contributing factors.

While generalized models may lack precision in predicting individual outcomes, highly personalized models may not consolidate theoretical findings across individual models, and suffer from variability in data quality across individuals. Person-centric assessment has therefore been cited as a matter of degree that is dependent on the specific research problem [135]. In our work, we take a middle ground approach between a purely generalized approach and a purely personalized approach, that aims to capture between-individual homogeneity and within-individual heterogeneity of information. We use offline sensors to group individuals into a few behavioral clusters, and then build cluster-specific prediction models of a wide range of psychological constructs using social media data. To our knowledge, our work is the first to evaluate the utility of this approach with data obtained from a large-scale and longitudinal multimodal sensing study.

3 STUDY AND DATA

We deploy social media as a passive sensing modality of psychological constructs in a large-scale multisensor study, Tesserae [75]. This study recruited 757 participants (of which 754 completed enrollment) who are information workers belonging to diverse geographical areas, job roles, and organizations in the U.S. The participants were requested to remain in the study for either up to a year or through April 2019. They were enrolled from January 2018 through July 2018. They either received a series of staggered stipends up to $750 or they participated in a set of weekly lottery drawings (multiples of $250 drawings) depending on their employer restrictions.

Privacy and Ethics. The Tesserae project was approved by the Institutional Review Board at the researchers' institutions. Given the sensitivity of the data, participant privacy was a key concern. The participants were provided with informed-consent documents describing the sensing streams, and specifics of what data each device captured and how it would be stored. The participants needed to consent to each sensing stream individually, and could also clarify concerns and opt out of any stream. The data was de-identified and stored in secured databases and servers which were physically located in the researcher institutions with limited access privileges.

3.1 Self-Reported Data

The enrollment process consisted of an initial survey questionnaire related to demographics (age, gender, education, type of occupation, role in the company, and income), and survey questionnaires of self-reported psychological constructs including: 1) *Cognitive Ability*, as assessed by the Shipley scales of Abstraction (fluid intelligence) and vocabulary (crystallized intelligence) [116], 2) *Personality Traits*, the big-five personality traits as assessed by the Big Five Inventory (BFI-2) scale [117, 125], and 3) *Wellbeing*, the general positive and negative affect levels as assessed through the Positive And Negative Affect (PANAS-X) scale [133], the anxiety level as measured via State Trait Anxiety Inventory (STAI-Trait scale) [118], and the quality of sleep as measured via the Pittsburg Sleep Quality Index (PSQI) scale [31]. Table 1 summarizes the distribution of the self-reported data within our dataset, where we find a reasonable distribution within demographics and psychological traits among our participants.

Fig. 1 presents Pearson's r between the psychological constructs (Cognitive Ability, Personality Trait, and Affect and Wellbeing variables) and regression (R^2) results for the demographic and job-related characteristics as independent variables, and the psychological constructs as dependent variables. These correlations between psychological constructs, mirror prior literature. The observed positive correlation ($r=0.79$) between Neuroticism and Anxiety and Negative Affect is consistent with past research showing positive associations between elevated Neuroticism and mood and anxiety disorders [83]. The strong positive correlation ($r=0.67$) between Anxiety and Negative Affect is consistent with past research showing a strong co-morbidity between depressed mood and elevated anxiety [84]. Extraversion and Positive Affect ($r=0.54$) have also been found to be strongly positively correlated [68, 69]. Other past research has also found a negative association ($r=-0.51$) between Positive Affect and Anxiety [11], as well as a negative association ($r=-0.41$) between Conscientiousness and Anxiety [39]. All

Table 1. Descriptive statistics of self-reported demographics and psychological constructs of participants.

Covariates	Value Type	Values / Distribution
Demographic Characteristics		
Gender	Categorical	Male | Female
Age	Continuous	Range (20:68), Mean = 34.90, Std. = 9.74
Education Level	Ordinal	5 values [HS., College, Grad., Master's, Doctoral]
Job-Related Characteristics		
Income	Ordinal	7 values [<$25K, $25-50K, ... , >150K]
Tenure	Ordinal	10 values [<1 Y, 1Y, 2Y, ... 8Y, >8Y]
Supervisory Role	Boolean	Non-Supervisor | Supervisor
Cognitive Ability (Shipley scale)		
Fluid (Abstraction)	Continuous	Range (0:24), Mean = 16.98, Std. = 2.84
Crystallized (Vocabulary)	Continuous	Range (0:40), Mean = 33.15, Std. = 4.11
Personality Trait (BFI scale)		
Openness	Continuous	Range (1.17:5), Mean = 3.82, Std. = 0.61
Conscientiousness	Continuous	Range (1.42:5), Mean = 3.88, Std. = 0.66
Extraversion	Continuous	Range (1.58:5), Mean = 3.42, Std. = 0.69
Agreeableness	Continuous	Range (2.08:5), Mean = 3.89, Std. = 0.56
Neuroticism	Continuous	Range (1:4.92), Mean = 2.46, Std. = 0.79
Affect and Wellbeing		
Pos. Affect	Continuous	Range (13:50), Mean = 34.53, Std. = 6.05
Neg. Affect	Continuous	Range (10:43), Mean = 17.52, Std. = 5.35
Anxiety	Continuous	Range (20:72), Mean = 38.13, Std. = 9.49
Sleep Quality	Continuous	Range (0:19), Mean = 6.65, Std. = 2.59

other inter-construct correlations are moderate at $|r|<0.40$. Next, looking at the association between demographic and job related variables and the psychological constructs, we observe only modest associations, with the strongest association being between Income bracket and the Shipley Crystallized Vocabulary scale ($R^2=0.05$). When all demographic and job-related variables are included in a regression model predicting the psychological constructs, we still observe only modest predictive performance for all psychological constructs (all $R^2<0.12$).

3.2 Passive Sensing Data for Offline/Physical Activity

The study used three physical sensors, which are briefly described below.

Bluetooth Beacons. Participants were provided with two static and two portable Bluetooth beacons (Gimbal [7]). Static beacons were to be placed at their work and home locations, and the portable beacons were to be carried either in their backpacks or their wallets. Combined, these beacons tracked participant presence at home and/or work location, and also help to assess their commute and time at their work desk.

Wearable. Participants were provided with a fitness band based smartwatch (Garmin Vivosmart [6]), which they wore throughout the day. The wearable continually tracked and recorded health measures, such as heart rate variability, stress, and physical activity in the form of sleep, footsteps, and calories lost.

Smartphone Application. A smartphone application [79, 130] was installed on Participant smartphones (Android and iPhones) . This application tracked phone use, lock or unlock behavior, call durations, and also leveraged their smartphone-based mobile sensors to track their location (mobility), and physical activity.

3.3 Social Media Data

The Tesserae project asked permission from the consented participants to authorize Facebook and Linkedin data, and optionally for Twitter, Instagram, Gmail, and Calendar, *unless they opted out, or did not have an account.*

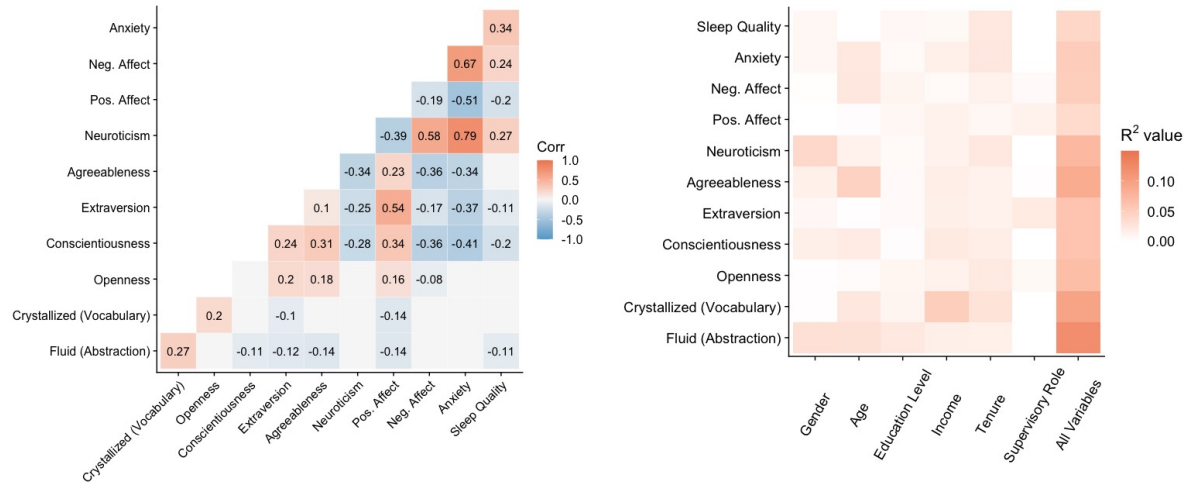

Fig. 1. **Left**: Pearson's r between psychological constructs, non-significant correlations (p>0.05) are left as blank. **Right**: Regression (R^2) results for the demographic and job-related variables as independent variables, and the psychological constructs as dependent variables (right sub-figure). The "All Variables" column provides regression results for the psychological constructs with all demographic and job-related variables included in the model.

Participants authorized access to their social media data through an Open Authentication (OAuth) based data collection infrastructure that was developed in [105]. OAuth protocol is an open standard for access delegation, which is commonly used as a way for internet users to log in and grant third party access to their information, without sharing passwords. Compared to other data collection strategies,(e.g. downloading and sharing of social media archives, or scraping through webpage crawlers or smartphone applications) the OAuth protocol provides a more privacy-preserving, streamlined, and convenient means of data collection at scale, over a secured channel without the transfer of any personal credentials.

The data collection application used the official APIs from the platforms (e.g., Facebook Graph API for Facebook data). Given that Facebook is the most popular social media platform [49], and that its longitudinal nature has facilitated several social media studies of understanding individual differences [28, 73, 106], it suits our particular problem setting. Facebook is also the most prevalent social media stream in our dataset, with 572 participants authenticating their Facebook data, among which 32 participants have no entries in their Facebook data — this paper uses the remaining 540 participants' Facebook data for measuring psychological constructs.

4 FEATURE ENGINEERING

We derive machine-usable features from our raw multimodal sensing data. We draw on prior work to derive features that have shown theoretical relevance in measuring psychological constructs [130, 131]. This section explains our features: first those derived from physical sensors, followed by those derived from social media. Out of the 757 participants' data, we set aside a random sample of 6.7% (50) participants' data as the held-out dataset for final evaluation purposes. We conduct feature engineering, and build (train and validate) our models within the remaining 93.3% (704) participants' data.

4.1 Deriving Features from Physical Sensor Dataset

From the passive sensor data streams, we derive a variety of features that are related to participants' activity, sleep, and other physical behaviors. We summarize the features below.

Step Count. The Garmin wearable collects fitness-related measures such as the daily step count of participants [6].

Physical Activity. The smartphone app installed on participant smartphones used the Google Activity Recognition API [5] to identify physical activity at regular intervals. For each individual, we obtain durations of (1) *high* and (2) *moderate* intensity activities using the Metabolic Equivalent of Task metric from the wearable data [128].

Mobility. The smartphone application continually recorded the GPS coordinates of the individuals. We derive the number of locations and the distance traveled between each location based on a 15-minute pooling window. We use this data to also derive (1) information on the total distance traveled each day, (2) the number of distinct locations visited, and the (3) maximum and (4) average distance from home traveled by an individual each day.

Phone Use Activity. The installed smartphone application recorded the activity of the smartphone locks or unlocks that the individuals made. We derive the (1) number of phone locks and unlocks, and (2) the average duration of time between phone locks and unlocks each day.

Desk Activity. The Bluetooth beacons in conjunction with the smartphone application captured the presence of individuals (whether at work or home locations). We derive a number of daily features about activity patterns at work and home each day, including (1) time at work, (2) minutes at desk, (3) mean desk session duration, (4) median desk session duration, (5) percent of time at work spent at desk, (6) and the percent of time of the 24 hour day spent at work. We also compute *break session* information, i.e. the intervals between the participant's desk beacon being out of range and the desk beacon appearing within range. Specifically, we compute daily counts of break sessions at three different interval measures: (7) number of 5-minute breaks, (8) number of 15-minute breaks, and (9) number of 30 minute breaks.

Sleep. The wearable sensed the sleep activity of the individuals [6]. Wearables can accurately detect sleep [60, 138], and we improved this measurement by further accounting for phone use and wearable-derived bed times, wake times, and sleep duration drawing on Martinez et al. [74]. In addition to collecting daily measures of (1) bed time, (2) wake time, and (3) sleep duration using this method, we also derive duration measures directly from the wearable for (4) light sleep, (5) deep sleep, and (6) Rapid Eye Movement (REM) sleep [72]).

To compute physical sensor features for clustering the participants into different behavioral contextualizations, we calculate the mean (μ) and standard deviation (σ) of each daily measure described above, for each individual. In addition to mean and sd features, we also compute features characterizing the *regularity* of each measure using Recurrence Quantification Analysis (RQA) [134] RQA estimates the number and duration of occurrences of a dynamical system presented through a phase space trajectory [134]. Regularity measures derived from multimodal sensing data have shown valuable utility in predicting psychological traits [131].

In particular, we obtain features using RQA for the *recurrence rate*, which represents the probability that a specific state will occur, and can be interpreted as the repetitiveness of the elements in a given sequence (i.e. the repetitiveness of values across the days for which data was collected). RQA is computed using three parameters: (1) the delay parameter τ, which is the delay unit by which the series is lagged, the dimension embedding D, which is the number of embedding dimensions for phase reconstruction, i.e. the lag intervals, and the radius R, which is the threshold cut-off constant used to determine if two points are recurrent or not. For each daily sensor measures series, we use the method recommended by Wallot [129] to compute the optimal parameters for each series, we computing the optimal parameters for each individual, and then using the mean value from this distribution of parameters to apply to the sensor measure stream. Among the RQA features, we could not attain useful features for mean and maximum average distance from home, as these RQA features show almost no variability across the participants, and hence, we discard them from our final feature set. In total, from mean, standard deviation, and RQA aggregation methods, we obtain 76 behavioral features for all participants.

4.2 Deriving Features from Social Media Dataset

Longitudinal Social media data of individuals is self-recorded in naturalistic settings. This data also enables us to obtain historical behavior of participants, i.e., from before study participation. Drawing on prior work [27, 29, 106, 107, 112], we obtain a variety of features from the Facebook data of the participants, and summarize them below.

Psycholinguistic Attributes. A number of prior work in the space of social media and psychological wellbeing [29, 112] have used psycholinguistic attributes in building predictive models. On the Facebook posts of the individuals, we use the well-validated Linguistic Inquiry and Word Count (LIWC) lexicon [124] to extract a variety of psycholinguistic categories (50 in total). These categories consist of words related to 1) *affect* (categories: anger, anxiety, negative and positive affect, sadness, swear); 2) *cognition* (categories: causation, inhibition, cognitive mechanics, discrepancies, tentativeness); 3) *perception* (categories: feel, hear, insight, see); 4) *interpersonal focus* (categories: first person singular, second person plural, third person plural, indefinite pronoun); 5) *temporal references* (categories: future tense, past tense, present tense); 6) lexical density and awareness (categories: adverbs, verbs, article, exclusive, inclusive, negation, preposition, quantifier); 7) *personal and social concerns* (categories: bio, body, death, health, sexual, achievement, home, money, religion, family, friends, humans, social).

Open Vocabulary *n*-grams. Open-vocabulary based approaches can infer psychological constructs of individuals [112]. We obtain the top 5000 n-gram ($n = 1, 2, 3$) from our dataset as features.

Sentiment. An important dimension in the language expressed on social media is the tone or *sentiment* of a social media post, which has also been used to understand psychological constructs and shifts in mood of individuals [46, 106]. We use the Stanford CoreNLP library's deep learning based sentiment analysis tool [71] to identify the major sentiment of a post among positive, negative, and neutral sentiment labels.

Latent Lexico-Semantics (Word Embeddings). Word embeddings are vector representations of language in latent semantic dimensions, enabling us to capture the lexico-semantics of language on social media. Prior work reveals that word embeddings can improve several natural language analysis and classification problems [25, 89, 109]. We use pre-trained word embeddings (GloVe [89]) on an internet corpus of 6B tokens in 50-dimensions to characterize the social media posts of our participants in a 50-dimensional feature space.

Social Capital. Social capital is an important aspect and contributor in shaping our lives and behaviors [121]. Drawing on prior work [28], we obtain features quantifying social capital of the individuals based on social media interactions and engagement. We use regular expression based pattern matching to identify individuals' updates relating to 1) check-ins to places (or locations visited), 2) posts of status updates, 3) upload of media (photo or video), 4) spend time (or an occasion) with other people (or friends), 5) change in relationship status and 6) use of apps (such as games or quizzes on Facebook). For each of these social attributes, we compute the number of updates, frequency of updates, and the number of likes and comments received in them.

In total, we obtain 5,127 derived features corresponding to each participant on their social media data.

5 AIM 1: CONTEXTUALIZING AND PREDICTING PSYCHOLOGICAL CONSTRUCTS

Our work focuses on predicting psychological constructs with social media data. Social media use and expressiveness may not only vary significantly across individuals, but also they are driven by offline factors. Therefore, we hypothesize that contextualizing on offline behaviors may make models better adapted to the social media signals predictive of psychological constructs per individual. As briefly introduced before, we take a middle-ground approach between fully individualized and fully generalized prediction models, which aims to capture between-individual homogeneity and within-individual heterogeneity. Intuitively speaking, given the sparsity of social media data, for an individual, whose social media data of certain behaviors or moments is "missing" (or lack of within-individual heterogeneity in data), we could fill these gaps by capturing the data from other similar individuals (between-individual homogeneity); here the similarity is captured via offline behavioral clustering.

To do so, we first cluster individuals on the basis of offline physical behaviors (e.g. sleep, work, phone use, physical activity) captured from sensors on Bluetooth beacons, wearables, and smartphones. Then, we build cluster-specific prediction models of psychological constructs, where each cluster-specific model uses the social media data of participants only within the corresponding cluster. Figure 2 schematically summarizes our contextualized prediction approach in comparison to generalized prediction approach with social media data.

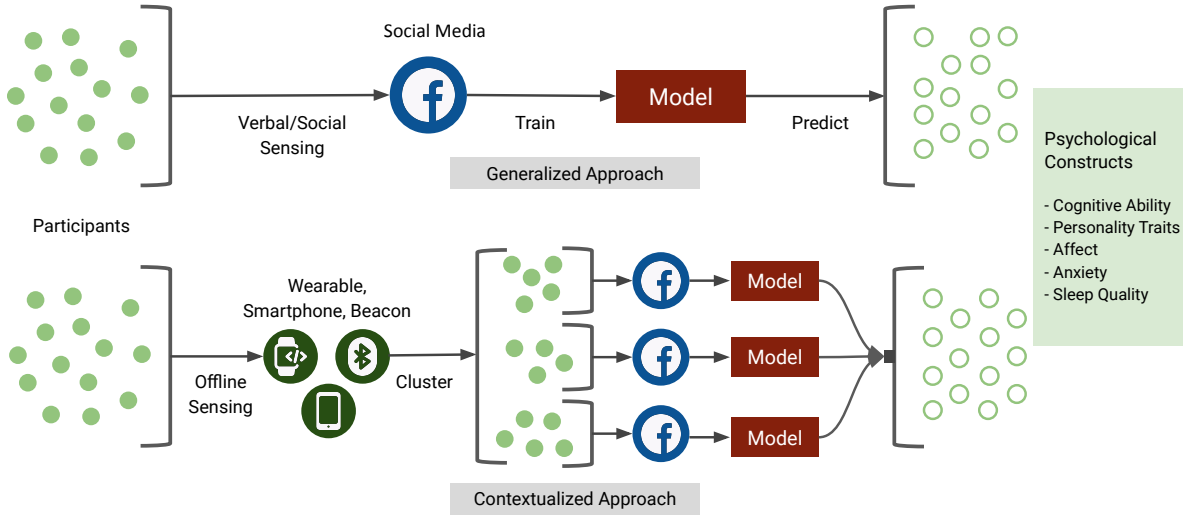

Fig. 2. A Schematic diagram comparing generalized prediction with social media data and our person-centered approach by contextualizing offline behaviors.

5.1 Clustering Individuals on Physical Sensor Behavior

To categorize our participants into different clusters based on their behavioral features, we first perform data imputation and feature selection to reduce missing values and feature redundancy. Feature selections and feature transformations are important preliminary steps for any machine learning problem to overcome problems of multi-collinearity, co-variance, etc. among the features — issues that can potentially affect downstream prediction problems [32]. In particular, because our features are derived from multimodal data streams, there is a high likelihood many of the features are already related, are redundant, and/or show extremely high variance and lack predictive power [95]. For example, the activity and stress-related features as captured by our wearable, are both intuitively and theoretically correlated [6].

We start with 76 physical sensor features for 704 participants in the training and validation data as explained in the previous section. We first impute any existing missingness (as is common in any longitudinal and large-scale data collection) in the data by using mean imputation per feature. Next, we step-wise remove multi-collinear features by calculating the variance inflation factor (VIF) of features against each other [24, 78]. We eliminate correlated features that show a VIF higher than 5, reducing our feature set from 76 down to 46 features.

5.1.1 Building Clustering Models. On the above finalized training and validation dataset, we apply four clustering algorithms to obtain the optimal arrangement: K-means, partitioning around medoids (PAM) [127], and two versions of hierarchical clustering. While both hierarchical clustering methods use Wards method for agglomeration between clusters [132], the first method ($hclust_1$) uses Ward's approach on the two observations and/or clusters which were recently merged when updating the distance matrix, while the other method ($hclust_2$) uses Ward's approach on all observations in the merged clusters, i.e. using less shortcuts when updating the distance matrix.

For each clustering algorithm we test cluster arrangements ranging from 2 to 8 clusters, using the mean Silhouette score [101], Dunn index [34], and the connectivity of the clusters (i.e. the degree of connection within the clusters, measured by k-nearest neighbors) [52] to determine the most optimal clustering arrangement. Table 2 presents the results of the clustering tests by varying parameters. We find that the most optimal clustering

Table 2. Comparing goodness of fit metrics for clustering methods and parameters (Number of clusters, Silhouette score, Dunn index, and Connectivity). The **green** highlighted row is the finally used clustering approach.

# Clusters	Sil. Score	Dunn Idx.	Connectivity
K-Means			
2	0.11	0.09	381.27
3	0.04	0.09	431.09
4	0.05	0.09	463.40
5	0.06	0.10	446.15
6	0.04	0.08	576.84
7	0.05	0.08	609.36
8	0.05	0.07	618.97
Partitioning Around Medoids (PAM)			
2	0.03	0.09	223.37
3	0.02	0.09	454.82
4	0.03	0.09	473.11
5	0.02	0.09	521.16
6	0.01	0.09	584.95
7	0.02	0.09	585.91
8	0.01	0.09	599.11

# Clusters	Sil. Score	Dunn Idx.	Connectivity
Hierarchical 1 (distance on recent two observations)			
2	0.23	0.13	214.92
3	0.02	0.10	406.09
4	0.02	0.10	443.02
5	0.03	0.10	453.82
6	-0.03	0.05	584.50
7	-0.03	0.05	626.90
8	-0.02	0.05	632.54
Hierarchical 2 (distance on all observations)			
2	0.27	0.14	149.10
3	0.26	0.14	151.66
4	0.05	0.10	350.83
5	0.03	0.10	454.58
6	0.04	0.10	455.95
7	0.04	0.10	459.50
8	0.04	0.10	462.36

arrangement in terms of mean Silhouette score and Dunn index is $hclust_2$ with 2 clusters, however, a close second is $hclust_2$ with 3 clusters, where the connectivity score is slightly higher but the mean silhouette score is slightly lower (0.26 vs. 0.27). As our primary research goal is to investigate the utility of building separate social media based models based on *contextualized* behavioral information about individuals, rather than a rigorous evaluation the most optimal clustering arrangement of our dataset, we consider that having more individualization in behavioral categories might provide a better evaluation of our theoretical approach. We therefore decide to proceed with our analysis and model building using the $hclust_2$ clustering algorithm with 3 distinct clusters.

5.1.2 Describing Clusters on Physical Behaviors. Applying the *hclust2* clustering algorithm with three distinct clusters to our training and validation subset, we find Cluster C_1 to have the majority of the participants (N=601, 85%), while Cluster C_2 has N=76 (11%) participants, and Cluster C_3 has N=27 (4%) participants. To better understand how the clusters differed among the behavioral features we generated, we apply the Kruskal-Wallis H-test to each of the 46 behavioral features with the responding variable as the behavioral feature, and the independent variable as the cluster membership category. We use the Kruskal-Wallis H-test as our cluster sizes are very different in size, and therefore we cannot likely assume a normal distribution of the feature values within each cluster. Table 3 reports the top 20 behavioral features with significant H-statistic values from the tests.

We find many regularity (RQA-based) features in the top 20 features. Regularity in minutes at desk per day, desk session duration, REM sleep duration, and number of phone unlocks are strong explanatory features to distinguish the three clusters. To investigate more specifically how the top features differ across each cluster, we transform the values for these features into z-scores within our entire participant set — Table 3 also provides the mean z-scores. z-score transformations are not sensitive to absolute values and measure the raw value in terms of standard deviations above or below the mean. We observe differences in C_3 compared to C_1 and C_2 for a number of the features, primarily with respect to work behaviors. Participants in C_3 had much higher daily regularity in minutes at their work desk per day (mean z=3.46) than those in C_1 (mean z=−0.17) or C_2 (mean z=0.12), but also on average spent a much lower percentage of their workday at their desk (mean z=−1.55) compared to those in C_1 (mean z=0.08) or C_2 (mean z=−0.03). We also observe distinct differences in C_2 compared to C_1 and C_3 with respect to sleep patterns. Participants in C_2 had more regularity in nightly seconds of REM sleep (mean z=1.44)

Table 3. Mean z-scores per cluster for top 20 significant features as per Kruskal Wallis H-test used for clustering. Statistical significance reported after Bonferroni correction (*** $p < .001$, ** $.001 < p < .01$, * $.01 < p < .05$).

Features	C_1	C_2	C_3	H-stat.
Phone Use				
Regularity of Number of Phone Unlocks per day	-0.15	1.09	0.33	63.12***
Regularity of Duration Spent with Phone Unlocked per day	-0.15	1.12	0.22	38.38***
Mean Number of Phone Unlocks per day	0.06	-0.38	-0.19	20.83**
Work Behaviors				
Regularity of Minutes at Desk per day	-0.17	0.12	3.46	79.72***
Regularity of Mean Desk Session Duration	-0.14	0.10	2.88	74.34***
Regularity of Median Desk Session Duration	-0.14	0.02	2.95	60.73***
Regularity of Percent Time Spent at Work per day	-0.14	0.03	3.07	70.84***
Mean Percent of Time at Work Spent at Desk	0.08	-0.03	-1.55	47.62***
Stdev. of Time at Work Spent at Desk	0.07	-0.11	-1.24	20.00**
Sleep				
Regularity of Total REM Sleep Duration per night	-0.18	1.44	-0.10	76.27***
Mean of Total REM Sleep Duration per night	0.13	-0.91	-0.32	54.48***
Stdev. of Total REM Sleep Duration per night	0.14	-0.98	-0.24	37.68***
Regularity of Total Deep Sleep Duration per night	-0.17	1.26	0.13	32.52***
Regularity of Nightly Bed Time	-0.03	0.38	-0.33	32.03***
Mean of Total Light Sleep Duration per night	0.11	-0.89	0.02	27.60***
Stdev. of Nightly Bed time	-0.07	0.23	0.88	18.07**
Physical Activities				
Regularity of Steps Count per day	-0.11	0.85	-0.01	51.58***
Regularity of Total High/Strenuous Activity Duration per day	-0.14	1.08	-0.06	46.92***
Regularity of Total Activity Duration per day	-0.14	1.13	-0.03	30.70***
Mobility				
Mean of Total Distance Travelled per Day	0.05	-0.33	-0.41	14.55*

compared to C_1 (mean z=−0.18) or C_3 (mean z=−0.10), but also had a lower average in nightly seconds of REM sleep (mean z=−0.91) compared to C_1 (mean z=0.13) or C_3 (mean z=−0.32).

5.1.3 Examining Clusters On Demographic Composition. We also examine the demographic composition of each cluster. Creating separate participant clusters based on behavioral features might be similar to directly clustering on demographic information. For instance, older adults are known to be more sedentary [12], and factors like age and gender have been shown to explain daily smartphone usage [4]. As our goal concerns the utility of building person-centred models based on passively sensed behavioral data, rather than static demographic information, we strive to have our clusters to have heterogeneous demographic compositions.

To test the heterogeneity of the demographic composition across clusters, we perform χ^2 tests between the clusters and the categorical demographic variables (Gender, Education Level, Income, Tenure, and Supervisory role), and a one-way ANOVA test between the clusters and age (the only numeric demographic variable). The tests reveal no significant association between the clusters and Age ($\chi^2(2)=0.91$, $p=0.63$), Gender ($\chi^2_{Pearson}(2)=3.06$, $p=0.22$), Income ($\chi^2_{Pearson}(12)=10.99$, $p=0.53$), Supervisory Role ($\chi^2_{Pearson}(2)=3.72$, $p=0.16$), and Tenure ($\chi^2_{Pearson}(18)=19.97$, $p=0.34$). While we see a weak significant association between the clusters and Education ($\chi^2_{Pearson}(8)=17.25$, $p=0.03$), the effect size is very small ($\hat{V}_{Cramer}=0.08$). We observe slight compositional differences in Education in C_3, such that there are proportionately more participants with High school as the highest level of education (7%), compared to C_1 (1%) and C_2 (0%). C_3 also has proportionately less participants with a College degree as the highest level of education (44%), compared to C_1 (55%) and C_2 (54%). However, these significant demographic differences are relatively negligible and only occur for education. Therefore, we conclude that our clusters are much more strongly separated by the passive behavioral features than by demographic information.

5.2 Predicting Psychological Constructs with Social Media

We use the features described in Section 4 to predict self-reported psychological constructs (Table 1). For each psychological construct, we build two kinds of models: 1) **generalized models** which are built on the entire dataset of all participants, 2) **contextualized models** which are separately built per behaviorally contextualized clusters. Here, the generalized prediction models emulate typical practices of predicting behavioral attributes with social media data, whereas the clustered models are more person-centered driven by incorporating people's behavioral features (passively inferred via physical sensors).

We use k-fold cross-validation ($k = 5$) for parameter tuning and evaluation with *pooled accuracy technique* on the training and validation subset of our data, i.e., for each model, we first divide the dataset into five equal segments, then iteratively train models on four of the segments to predict on the held-out fifth segment, and finally collate all the predicted values together and compare their collated against actual values using Pearson's correlation (r) and Symmetric Mean Absolute Percentage Error (SMAPE). We adopt several prediction algorithms spanning across linear regression (with and without L_1, L_2 regularization), gradient boosted random forest (GBR), support vector regressor (SVR), and multilayer perceptron (MLP).

We also transform our social media feature set using Principal Component Analysis (PCA) with a singular value decomposition solver, selecting the number of components on the basis of explained variance [55][1]. However, prediction models using PCA transformed features show no improvement over those using raw features (no-PCA transformation), likely because language and n-gram features are inherently sparse and contain predictive information despite the variance and sparsity. The remaining paper concerns analyses with raw features, which serves an additional advantage of feature interpretation and model explanation with respect to contextualization.

To verify that our training models do not overfit, we also apply the cross-validated and trained models to the held-out unseen subset of our data (N=50) to test performance on unseen data (introduced at the beginning of Section 4). We derive the 46 physical sensor features for our held-out data, and apply the same trained $hclust_1$ model to obtain cluster labels for the held-out data. We again evaluate the relative performance of the generalized and contextualized models on the held-out data.

6 AIM 2: EVALUATING PERFORMANCE OF CONTEXTUALIZED AND GENERALIZED MODELS

On the best performing algorithm for generalized and contextualized prediction models from the above, we compare the performance metrics of these predictions for the cross-validated evaluation with the training data (Table 4, detailed metrics in Appendix, Table A1 and A2), and for the held-out data (Table 5). To measure statistical significance in prediction differences, we conduct t-tests using the dependent overlapping correlation method, which controls for comparing against a common variable of interest (here, each psychological construct) [35]. We observe that the efficacy of contextualizing on offline behavior for social media based predictions can be explained by the theoretical associations between the construct and its offline manifestations. We now discuss the performance of the models, and ask: When does contextualization help? Are there any cases where contextualization does not improve over generalized models?

6.1 Cognitive Ability

We use the Shipley scales to obtain ground-truth measures of two kinds of cognitive ability. The Shipley (Abstraction) scale measures fluid cognitive ability, which is how one thinks logically, reasons, and problem solves in novel situations. The Shipley (Vocabulary) scale measures crystallized cognitive ability (Section 3) [116], or an individual's grasp of general and cultural knowledge including verbal communication [16]. These two abilities mutually interact and combine to form overall individual cognitive ability [57].

We refer to Table 4 and Table 5 to compare the best predictions as per generalized and clustered models in the cross-validated evaluation and held-out data respectively. In the case of abstraction, we find that there is no

[1]The Appendix Tables A3 and A4 show the performance of modeling with PCA-transformed features. Qualitatively, these predictions show similar comparison directions between generalized and contextualized models as observed in models with non-transformed features.

Table 4. Cross-validated Evaluation: Comparing the accuracy metrics of best performing generalized and contextualized prediction models. Statistical significance is computed using t-test as per dependent overlapping correlations [35] on predictions by generalized and contextualized models for each construct. For significant rows, **pink** bars indicate a **decrease in performance** in contextualized models compared to generalized models and **green** bars indicate an **increase in performance** (*** p<.001, ** .001<p<.01, * .01<p<.05).

Construct	Generalized		Contextualized		Comparison		
	r	SMAPE	r	SMAPE	Δr %	ΔSMAPE %	t-stat.
Cognitive Ability							
Shipley (Abstraction)	0.25	6.81	0.23	6.88	-8.00	1.03	-1.73^{-}
Shipley (Vocabulary)	0.29	4.13	0.21	4.25	-27.59	2.91	-4.82***
Personality Traits							
Openness	0.25	6.89	0.29	6.08	14.81	-11.76	1.94*
Conscientiousness	0.13	7.29	0.19	7.08	46.15	-11.76	2.80**
Extraversion	0.17	8.54	0.21	8.46	23.53	-0.94	1.70*
Agreeableness	0.17	5.84	0.19	5.89	11.76	0.86	0.88^{-}
Neuroticism	0.12	13.56	0.18	13.09	50.00	-3.47	2.50*
Affect and Wellbeing							
Pos. Affect	0.13	7.10	0.14	6.90	7.69	-2.82	0.56^{-}
Neg. Affect	0.11	10.90	0.13	10.89	18.18	-0.09	-1.13^{-}
Anxiety (STAI)	0.12	9.66	0.21	8.51	75.00	-11.90	5.61***
Sleep Quality (PSQI)	0.15	16.02	0.25	10.59	66.67	-33.90	5.07***

significant difference in the performances of generalized and clustered models. However, in the case of vocabulary, we find a statistically significant difference (t=-4.61), where the generalized model performs 27.6% better in r and 2.41% better in SMAPE in the cross-validated evaluation. We observe similar prediction results in the held-out data, where the generalized model performs 29.41% better in r and 8.41% better in SMAPE. However, the difference in the held-out data is not quite significant (t=-0.99, p=0.08), likely due to the smaller sample size.

The above suggests that clustering individuals on physical and offline behaviors does not add any new information in predicting cognitive ability. We construe that more heterogeneity of individuals in training sample and larger size of data are in fact stronger factors in predicting cognitive ability, likely because language is known to be a correlate of cognitive ability, more strongly in the case of crystallized cognitive ability (vocabulary) [111].

6.2 Personality Traits

Personality traits are considered to be robust and parsimonious correlates of a variety of individual outcomes, characteristics, behavior, and abilities [61]. We find that in both the cross-validated evaluation and the held-out data, contextualized predictions perform significantly better for Openness, Extraversion, and Neuroticism. Contextualized predictions of Conscientiousness perform significantly better in cross-validated evaluation, (t=2.80) without a significant difference in performance in the held-out set (t=1.26, p=0.21). Conversely, contextualized prediction for Agreeableness does not perform significantly better in the cross-validated evaluation (t=0.88, p=0.38), but prediction is significantly better in the held-out set (t=3.70). Despite these inconsistencies in statistical significance, there is a general trend towards increased performance in contextualized predictions for Conscientiousness and Agreeableness. The inconsistencies in statistical significance for improved Agreeableness predictions with contextualized models partially aligns with prior meta-analysis of physical activities and personality traits that found Agreeableness to show the least significant relationship with physical activities [98]. We also note that our participant pool is drawn from information workers, and certain activities such as work behaviors and phone use have been found to be direct correlates of traits like Neuroticism [24], so capturing such information during clustering (Table 3) may have contributed to the effectiveness of person-centered models. We construe that driving contextualized models through physical behavior based clusters likely allows to capture distinct linguistic features per cluster that are better predictive of personality traits.

Table 5. Held-out data: Comparing the accuracy metrics of best performing generalized and contextualized prediction models. Statistical significance is computed using t-test as per dependent overlapping correlations [35] on predictions by generalized and contextualized models for each construct. For significant rows, **pink** bars indicate a **decrease in performance** in contextualized models compared to generalized models and **green** bars indicate an **increase in performance** (*** p<.001, ** .001<p<.01, * .01<p<.05).

Construct	Generalized		Contextualized		Comparison		
	r	SMAPE	r	SMAPE	Δr %	ΔSMAPE %	t-stat.
Cognitive Ability							
Shipley (Abstraction)	0.36	4.52	0.33	5.03	-8.33	11.28	-0.36^{-}
Shipley (Vocabulary)	0.34	4.65	0.24	5.04	-29.41	8.40	-0.99^{-}
Personality Traits							
Openness	0.51	5.00	0.79	4.23	66.67	-15.40	4.00***
Conscientiousness	0.19	6.87	0.21	6.06	10.53	-11.79	1.26^{-}
Extraversion	0.19	7.71	0.32	6.88	68.42	-10.77	2.57**
Agreeableness	0.38	6.51	0.62	5.81	63.16	-10.75	3.77***
Neuroticism	0.12	12.86	0.63	11.11	425	-13.61	7.95***
Affect and Wellbeing							
Pos. Affect	0.30	9.20	0.60	8.54	100	-7.17	2.57***
Neg. Affect	0.20	11.33	0.42	10.87	110	-4.06	2.25***
Anxiety (STAI)	0.14	11.57	0.33	8.78	135.71	-24.11	1.14*
Sleep Quality (PSQI)	0.16	16.49	0.41	12.33	156.25	-25.23	2.27*

6.3 Affect and Wellbeing

Contextualized models reveal no benefit for predicting positive or negative affect in the cross validation data set, though a significant benefit is found in the held out data.. This inconsistency suggests that there is inconclusive evidence whether clustering individuals on their physical activities or offline behaviors contributes new information in the prediction of affect. It is likely that offline behaviors might not provide enough new information for predicting more moderate constructs of day-to-day affect like those measured with the PANAS scale.

However, for anxiety and sleep quality (PSQI), we see large and significant improvements for the contextualized predictions over generalized models in both cross-validated and held-out evaluations. We conjecture that these improvements are due to strong correlations between physical behaviors and sleep quality and anxiety. For instance, the duration of deep sleep is known to have a significant effect on reported sleep quality [18], while improved physical activity has a long-established relationship in reducing subjective anxiety [9]. Although extreme affective disorders like depression (which is often comorbid with anxiety) [47] are known to have relationships with offline behaviors like sleep and mobile phone use [126], offline behaviors might not provide enough new information for predicting more moderate constructs of day-to-day affect like those measured with the PANAS scale. We note that our clustering approach includes features obtained using wearables and smartphones, both of which capture behaviors correlated with sleep (e.g., accelerometer data, phone use, etc.), likely helping the contextualized predictions of sleep quality.

6.4 Robustness of Contextualized Person-Centered Approach

We test for empirical robustness of our contextualized person-centered approach against typical approaches of using physical sensor features for prediction (M_s models), as well as using all (physical sensor and social media) features together, i.e., a multisensor feature fused model (M_{ms} models). We conduct similar rigorous model and parameter tunings as above, and obtain the best models of predicting the constructs. Appendix Table A5 and Table A6 present the detailed prediction results for M_s and M_{ms} models respectively, and Fig. 3 presents a summary view of prediction performance comparison across various modeling approaches. For the M_s models, we see the prediction performance is considerably worse than contextualized models for all psychological constructs, with the strongest prediction from any M_s model is for Extraversion (r=0.17, SMAPE=8.41). We find

that the best M_{ms} models of cognitive ability perform as similar as generalized social media models, and better than contextualized models for both abstraction (r=0.25, SMAPE=6.66) and vocabulary (r=0.28, SMAPE=4.10). For personality traits, we find that M_{ms} performs similarly or worse than contextualized models in openness (r=0.26, SMAPE=6.40), conscientiousness (r=0.17, SMAPE=7.86), extraversion (r=0.17, SMAPE=8.35), agreeableness (r=0.17, SMAPE=5.91), and neuroticism (r=0.11, SMAPE=12.93). For affect and wellbeing, M_{ms} performs similarly in positive (r=0.13, SMAPE=6.77) and negative (r=0.13, SMAPE=11.18) affect, and significantly worse in anxiety (r=0.13, SMAPE=16.07) and sleep quality (r=0.21, SMAPE=14.02). Together, this suggests that our approach of person-centered contextualization not only mines signals in the multimodal sensing data better, but also likely filters out noisy information from user groups whose behavior is too far away from the average group. Further, the person-centered contextualization approach provides additional theoretical interpretation and explanation which we elaborate further in the following sections.

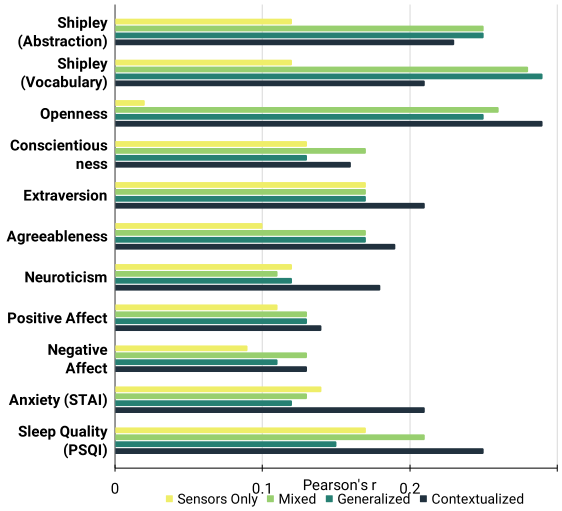

Fig. 3. Comparing performance (r) various modeling approaches for predicting psychological constructs.

In addition, we also target rejecting the null hypothesis that any prediction improvement by our contextualization approach is by chance or any random cluster-label assignment. Drawing on permutation test approaches [3, 104], we permute (randomize) the cluster label of all individuals, and repeat our entire pipeline predicting of psychological constructs. We run 1,000 such permutations, and we find that the probability (p-value) of improvement by a random-cluster assignment over contextualized approaches is almost zero across all the measures (p=0.002 for abstraction, p=0.001 for positive affect are the only non-zero probabilities). This rejects the null hypothesis and provides additional statistical significance and credibility to our person-centered approach of contextualization using offline behavioral clustering.

7 AIM 3: OFFLINE CONTEXTUALIZATION AND SOCIAL MEDIA LANGUAGE

The sensitivity of social media data to people's unique characteristics and variable social media use motivates us to study person-centered contextualized predictions. We have already proposed and validated an approach to contextualize social media predictions of psychological constructs by clustering people on offline behaviors (Aim 1 and 2). Now, we are interested in interpreting the same clusters in terms of people's social media use. Our third research aim targets understanding how the social media language varies by the clusters. We investigate if clustering individuals on *offline* behaviors leads to clusters of individuals who also have different online behaviors.

To understand the language differences better, we first interpret the composition of clusters on psychological constructs which can help us validate the theoretical foundation of building person-centered models on contextualized offline behaviors. Then, for each cluster, we obtain salient language use and interpret that with respect to cluster composition in offline behaviors, psychological constructs, and the literature.

7.1 Interpreting Cluster Composition on Psychological Constructs

We examine the between-cluster differences in psychological constructs (or our outcome measures). Figure 4 shows z-transformed representation of the mean composition of each cluster per construct. At a mean-aggregated level, C_1 shows the greatest average conscientiousness (μ=3.89, σ=0.66) and agreeableness (μ=3.90, σ=0.57). C_2

shows greatest average positive affect (μ=35, σ=5.89), negative affect (μ=17.49, σ=4.97), and anxiety (μ=38.76, σ=10.49). C_3 shows greatest average cognitive ability in both abstraction (μ=17.41, σ=2.61) and vocabulary (μ=33.41, σ=3.84), openness (μ=3.9, σ=0.41), neuroticism (μ=2.49, σ=0.76), and self-reported sleep (μ=7.03, σ=2.92), while showing low affective traits.

For all measures, Kruskal-Wallis H-tests across clusters show no statistical significance, which could mean that each cluster is already composed of heterogeneous psychological traits. This within-cluster variation in psychological constructs suggests that clustering on offline (dynamic) behaviors does not necessarily translate to clustering individuals with only "similar" psychological constructs. For each cluster and psychological construct, we compute the coefficient of variation (cv) [38], expressed as a percentage in the ratio of standard deviation to mean of a distribution, and quantifies the amount of variability with respect to the mean of the distribution — higher values indicate higher variability. We find that cluster-wise cv is on an average 7% lower compared to the entire (non-clustered) data's cv across all the measures. Together, this suggests that clustering finds a compromise in both preserving and reducing the variability in training data compared to the entire data. Methodologically, the within-cluster heterogeneity in outcomes plausibly helps the data within-clusters to be neither too biased (if only predicting homogeneous or skewed distribution of labels), nor too varianced (predicting high variance distribution of label, e.g., the entire data), thus helping the predictive performance of psychological attributes, as noted in Section 6.

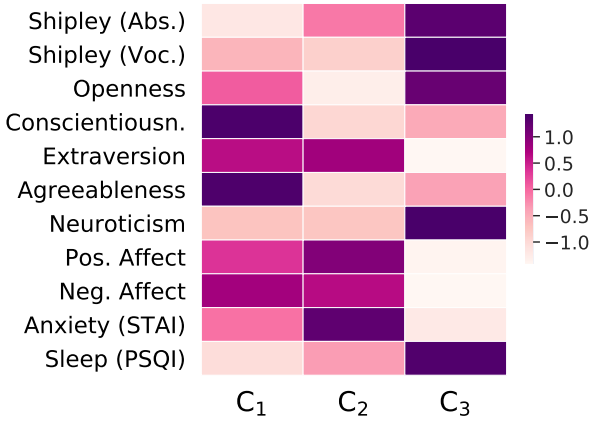

Fig. 4. Heatmap representing the clusters on mean psychological constructs. Values are z-transformed per measure.

7.2 Examining Social Media Language Differences Across Clusters

Next, we investigate how offline contextualization leads to groups of individuals with different social media use, and what are the likely theoretical interpretation of such groupings in terms of understanding psychological constructs. We base this examination on the differential psycholinguistic usage of individuals. We measure the statistical difference in language using Kruskal-Wallis H-test. Table 6 characterizes the clusters on psycholinguistic use categories which show significant differences as per Kruskal-Wallis H test. The average standard deviations across the categories are 0.12 for C_1, 0.16 for C_2, and 0.16 for C_3. We discuss this below.

Although **Cluster C_1** is the largest (468 individuals), it shows the lowest standard deviation (0.12), suggesting that this cluster is psycholinguistically least heterogeneous. C_1 shows the lowest use of all psycholinguistic attributes, suggesting that these individuals are typically less expressive on social media. An alternate explanation would be C_1 represents the more common language use on social media — language containing lesser presence of non-content words (articles, prepositions, etc.) that is known to be associated with less complex language [124]. Moreover, the low mean in psycholinguistic attributes could also be associated with high mean conscientiousness of C_1 (Figure 4), aligning with prior findings on social media language and psychological traits [112]. Similarly,examining C_1's offline behaviors (Table 3) shows these individuals typically score low on the regularity based features (work behaviors, sleep, and physical activities), suggesting that the relative heterogeneity of offline behaviors, does not translate to online behavior and social media expressiveness. Physically, C_1 travels the most (high mean total distance travelled per day), which suggests interesting associations between their offline mobility and online posting behavior, as has also been noted in prior work [59].

Table 6. Comparing psycholinguistic attributes across clusters. Statistical significance reported after Bonferroni correction (*** $p < .001$, ** $.001 < p < .01$, * $.01 < p < .05$).

Category	C_1	C_2	C_3	H-stat.
Affect				
Anger	0.04	0.41	0.29	46.48***
Negative Affect	0.15	0.26	0.39	28.46***
Positive Affect	0.32	0.62	0.55	105.52***
Sadness	0.14	0.21	0.29	18.42***
Cognition				
Causation	0.19	0.42	0.25	53.01***
Certainty	0.11	0.54	0.19	81.71***
Cognitive Mech.	0.51	0.69	0.55	46.03***
Inhibition	0.27	0.41	0.20	22.11***
Discrepancies	0.15	0.48	0.27	81.84***
Tentativeness	0.24	0.59	0.30	65.10***
Perception				
Feel	0.14	0.40	0.46	79.17***
Hear	0.11	0.41	0.31	64.18***
Insight	0.22	0.57	0.41	74.12***
Percept	0.20	0.55	0.57	94.32***
See	0.10	0.34	0.59	90.77***
Interpersonal Focus				
1st P. Singular	0.20	0.26	0.43	23.74***
1st P. Plural	0.05	0.37	0.42	78.10***
2nd P.	0.12	0.39	0.27	77.64***
Indef. P.	0.35	0.40	0.54	19.92***
Lexical Density and Awareness				
Adverbs	0.30	0.52	0.54	86.57***
Article	0.28	0.64	0.49	110.01***
Verbs	0.42	0.59	0.61	64.05***
Aux. Verbs	0.25	0.57	0.54	98.30***
Conjunction	0.27	0.63	0.61	107.54***
Exclusive	0.33	0.52	0.45	39.65***
Inclusive	0.30	0.67	0.58	104.74***
Preposition	0.37	0.58	0.77	111.5***
Negation	0.09	0.21	0.19	51.07***
Quantifier	0.16	0.52	0.17	66.07***
Relative	0.22	0.71	0.56	127.55***
Temporal References				
Future Tense	0.10	0.31	0.58	100.78***
Past Tense	0.18	0.34	0.37	73.22***
Present Tense	0.34	0.54	0.66	79.73***
Personal and Social Concerns				
Achievement	0.24	0.52	0.45	82.62***
Bio	0.08	0.31	0.54	98.16***
Body	0.02	0.29	0.38	78.54***
Family	0.09	0.13	0.22	28.11***
Friends	0.05	0.19	0.24	69.43***
Health	0.14	0.30	0.28	33.43***
Home	0.05	0.14	0.14	69.07***
Humans	0.08	0.19	0.49	81.18***
Money	0.06	0.46	0.30	83.66***
Religion	0.08	0.20	0.19	14.02***
Social	0.33	0.60	0.46	81.78***
Work	0.09	0.31	0.32	107.08***

Cluster C_2 shows the greatest use of *anger* and *positive affect*, and all cognitive attributes, suggesting these individuals have an average high emotion on social media language. Interestingly, the same cluster's composition showed high average affect and anxiety traits (Figure 4), suggesting that individuals with higher affect traits are likely to be more expressive with affect and emotional language on social media. These individuals have a high use of function words such as *articles, auxiliary verbs, conjunction, inclusive, exclusive, quantifier, and relative*. Function words are strong linguistic markers of understanding psychological processes [88]. These individuals also show a high use of *achievement* and *money* related language, which may be associated expressiveness about their career and self-actualization. We see plausible connections between these psycholinguistic trends and the behavioral sensing features which best separated the clusters (see Table 3), e.g., the greatest use of *health* words may be associated with high physical activities shown by them [2]. Additionally, C_2 has an average low duration of REM sleep per night, more regular daily exercise patterns, and more regularity in time spent with their phone unlocked per day, possibly because participants might be more driven for achievement in social, work, or athletic goals, sometimes at the expense of their sleep quality [110].

Cluster C_3 shows the greatest use of pronouns, with pronoun use associated with narrative language and interpersonal discourse on social media [88]. For instance, the greater use of first person singular pronouns (e.g. "I", "me") suggests narrating personal experiences and self-reflection, and that of first person plurals (e.g. "We", "Us") indicates narrating experiences as collective identities [19]. These individuals also score high on the use of language related to social concerns and relationships such as *family, friends, home, and humans*. We again see plausible connections between these psycholinguistic trends and the behavioral sensing features which best separate the clusters. C_3 has on average more regularity in the percent of time they spent at work and desk each day. Regularity in work and at desk suggest these participants might prefer a more regular daily work schedule to balance a need for a more consistent family or social life outside of work.

The above cluster-wise decomposition on social media language reveals how people's offline behaviors can help us group individuals who are also separated in social media language. Further, our analyses in Aim 1 and 2 also reveals how this approach helps improving predictions of psychological constructs, particularly those that bear strong associations with physical behavior (e.g., sleep quality). While it is intuitive that offline dynamic

behaviors indeed drive online behavior, there is a paucity of theoretical evidence [2]. Our work sheds light on this important aspect and opens up opportunities for future explorations of understanding human behavior.

8 DISCUSSION

We adopted machine learning and statistical modeling approaches to contextualize social media predictions of psychological constructs. We first clustered individuals on physical activities as captured by passive sensors and then built cluster-specific prediction models of cognitive ability, personality traits, affect, and wellbeing constructs. We evaluated the effectiveness of person-centered predictions against generalized predictions per construct to find that the effectiveness varies by construct, suggesting that personalization is only better than generalization in specific circumstances. In fact, clusters based on dynamic offline behaviors were not only heterogeneous in static traits (Section 5.1.3 and 7.1), but were also separated on social media language use (Aim 3). Our study is grounded on the Social Ecological Model that individual behaviors are influenced by factors related to an individual and their context [15]. We found that our person-centered (contextualized) models showed better overall predictive performance compared to generalized models using sensor data alone (M_s), generalized models using both sensor and social media together (M_{ms}), as well as when contextualization was done on random cluster-label assignment instead of sensor data. Although the overall predictive performance of contextualized models is somewhat modest in terms of r, our predictions are relatively close to the theoretical upper bound on the correlations between real-world behaviors and psychological traits of $r \sim 0.3$ to 0.4 [77], and to the predictions of previous past work predicting personality traits from passive mobile phone data [120]. Beyond just an evaluation of predictive performance, our work also provides insights applicable to studies grounded in similar theoretical settings where there is a need to focus on comprehensive social ecological signals, and an opportunity to infer behavioral and psychological attributes of individuals.

8.1 Theoretical and Methodological Implications

8.1.1 Beyond Traditional Forms of Personalization. Recent research in applied computing has highlighted the value of personalizations via "one-size may not fit all" arguments, as all individuals are not the same and have different experiences [102, 135]. This has also motivated various ubiquitous computing research in personalizing interactive, informatic, and intervention systems [20, 26, 64, 65]. While person-centered analyses have been studied in other disciplines such as social science, psychology, and health [56, 63, 135], we note that such analyses remain under-explored in computational assessments despite the abundance of data. A close application is personalized content recommendation systems [115]. These personalizations have typically relied on a single modality of data (e.g., historical content browsing), along with demographic and static information. Relying on isolated modalities are limited by several blindspots that challenge the comprehensibility of the models, such as the varying data quality across individuals. Importantly, collecting demographic information not only removes user anonymity and threatens data privacy, which is a growing public concern in social media use, but also can promote bias, exclusion, and stereotypes [58]. The implications of our work situate that with recent research revealing that behavioral predictions can sway away from demographic- and static data, by only accounting for short-term and behavioral data [24, 108].

Further, based on our examination of the association between demographic information and our target constructs (see Fig. 1), it is not readily evident that demographic information would provide more accurate predictions. In contrast, our work leverages naturalistic behavior collected via multimodal sensing to guide person-centered analyses. In comparison to stratifications on demographics and static traits, or other forms of strata assumptions, passive sensing allows us to cluster individuals on physical behaviors, which is robust and dynamic. The efficacy of person-centered models is plausibly explained by the notion that sensing streams both independently, as well as in conjunction can predict the constructs in consideration [50, 99, 120, 130, 131]. Our current work applies new ways of thinking about person-centered approaches in human-centric, context-aware, and social sensing and applications requiring personalized attributes.

8.1.2 Complementary Prediction Approaches. Our work contributes to the body of literature studying the complementary advantages of variable-centered and person-centered approaches in various social science and psychological constructs [63]. Person-centered approaches allow investigating individual attributes with precision and personalized context-adaptation. We construe that improved predictions are due to personalized training datasets by stratifying individuals on their lifestyle and offline behaviors, rather than relatively less helpful demographic information. On the other hand, no difference in performance could be either due to better statistical power of larger training data or due to no added signal in personalized training data. Our observations support Howard and Hoffman's study that determined no single approach whether generalized or person-centered can be considered to be the "best", and it depends on the particular problem of interest and research setting (in our case contextualizing predictions on offline behaviors) [56].

8.1.3 Tradeoff between Statistical Power and Personalization. social media data is sensitive to people's self-presentation, context, and other factors, thereby making it harder for generalized prediction models that target behaviors of average populations. Moreover, given that social media data is characteristically sparse (both within and between individuals), it may not be ideal for fully-individualized models. Our work overcomes these challenges by using data from offline behaviorally similar individuals, thereby increasing the training data compared to complete personalization while preserving the personalization aspect. However, the training data size in contextualized models is *still smaller* than that of generalized models. Therefore, we posit that personalized research requires a consideration of the tradeoff between statistical power against the personalization component of predictions. This tradeoff would likely arise in any kind of personalized predictions, and potentially builds on the classical "bias-variance tradeoff" in machine learning predictions [8] — over-personalized models can be too biased, whereas over-generalized models can suffer from variance in the dataset.

8.1.4 Generating New Hypotheses. Our work allows generating hypotheses on the relationship between human behavior, psychological constructs, and personalized predictions. These hypotheses can guide us to explore newer questions on what factors make some attributes personalizable, and how between-individual homogeneity and within-individual heterogeneity of information can serve as either a noise or a signal in such predictions. Example hypotheses guided by our findings are, 1) social media sufficiently predicts attributes related to cognitive ability, 2) physical behaviors may not be as effective as social media in predicting affect, 3) social media data needs to be complemented with offline data to accurately predict a physical measure such as sleep quality, and so on.

8.2 Implications for Researchers and Practitioners

Our work demonstrates the efficacy of person-centered predictions of psychological constructs (cognitive ability, personality, affect, and wellbeing) by using complementary ubiquitous technologies. In doing so, we take a critical stance to reflect upon conducting personalized predictions in practice.

8.2.1 Tradeoffs between Generalized and Personalized Models. Contextualizing social media predictions using passively sensed offline behavior allows us to go beyond the more common, user-profiling like approaches on demographic information based on one's age, race, and gender, which are not only less-robust, but also could lead to misleading findings or "stereotypy" about particular demographic groups. On the other hand, we understand that building personalized prediction models with dynamic and mutable behaviors (such as ours) demands additional overhead, including and not limited to obtaining an individual's multiple modalities of data which is both longitudinal and dense.

We provide insights on what kind of measures may or may not be personalizable. Personalized prediction is considered a useful approach for improving user experience and understanding human behavior in a variety of problem settings, however, personalization is costly in terms of statistical power and effort. One way to navigate through that is to conduct pilot studies with a small number of participants' data and groundtruth measures, before investing and implementing these approaches at scale. Such approaches can identify an appropriate granularity of an effective personalization, and sensors are most likely to improve predictions of a particular

construct. Indeed, more data *does not* necessarily lead to better predictions, as we observed from larger data available to generate generalized predictions. For instance, researchers interested in social media and cognitive ability could use our study as evidence that social media data is sufficient, and predictions would not be improved by additional effort spent collecting passive sensor data and clustering on it.

8.2.2 Considerations for When to Personalize. When considering person-centered predictions, one should consider the need and means. In our case, social media data is sensitive to an individual's choice of social media use and expressiveness. We were theoretically motivated in that social media activity is a function of people's offline behaviors and therefore clustered individuals on these behaviors for our analyses. A similar analog of data sensitivity and variability in quality in case of other sensors could be compliance (e.g., use and non-use of a wearable), and attributes that may drive compliance may be accounted for to cluster individuals.

In parallel, the theoretical motivation of personalizing with respect to a construct is also important to consider. For instance, in precision medicine or when dealing with health constructs, it could be critical to conduct personalized assessments, as many health conditions have clinical heterogeneity, driven by varying experiences and traits of people. In such cases, generalized models capturing average target behavior may not provide actionable insights or useful information to build Just In Time Adaptive Interventions (JITAIs) to address such health conditions [65, 119]. Similarly, in the workplace context, when predicting workplace outcomes, it might be useful to incorporate apriori information about the type and hours of work for contextualization.

The efficacy of personalized predictions is also plausibly dependent on the universality and applicability of a psychological construct on given population. For instance, given that sleep is a universal activity across individuals irrespective of their mutable and immutable characteristics, personalizing sleep quality predictions using physical activity turned out to be significantly effective. However, in cases of less-universal or cohort-specific metrics, such as academic success in grade school, it may make more sense to use demographic or other apriori groupings, e.g. grade level. In addition, person-centered approaches may be uniquely valuable in the cases of rare and less-prevalent attributes, such as understanding phobia, anxiety or psychotic disorders, where globalized datasets may be imbalanced and negatively-skewed due to rarity of the condition, significantly impacting the predictions.

Our work has an implication regarding *the reasonable use of sensors in accordance with the construct of interest.* As an example, contextualization improved predictions of sleep quality but impaired predictions of cognitive ability. One likely reason that person-centered models predicted sleep quality significantly better was that our multimodal sensing pool included wearable-based sleep sensor and physical activity sensors. In contrast, theoretically, none of our physical sensors bear a strong relationship with cognitive abilities, therefore increasing noise and impairing predictions using social media features which are inherently strong predictors of cognitive ability. An interesting question for future research could establish if sensors that capture speech and communication patterns and social interactions could improve predicting cognitive ability better than social media based predictions alone.

8.2.3 Ease of Interpretation and Domain Adaptations. Person-centered analytical approaches can represent individuals on their characteristics rather than defining them simply as a collection of variables [24, 56, 135]. These approaches can be more readily interpretable relative to all-inclusive features, where certain features may obscure others. Rather, our approach allows us to cross-introspect features, e.g., how certain offline behaviors are associated with social media language within and across clusters. This kind of explanation and interpretation may be immensely valuable in healthcare and precision medicine, which is defined as *"an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person"* [80]. Such approaches would allow simultaneous inspection at in-conjunction and isolated behaviors. For example, a particular combination of linguistic markers and disrupted physical movements may be useful for early detection of certain mental health symptoms and accordingly guide tailored intervention. Moreover, person-centered approaches can improve human-centered algorithms to be more considerate of the individual in focus by incorporating their circumstances and context.

In light of our findings, We reiterate Howard and Hoffman's point to researchers of human-centered machine learning that personalized or generalized approaches are not necessarily competitive, but are complementary in terms of methodological, statistical, and theoretical advantages [56]. While one approach may be 'better' outcome-wise, it is worth considering theoretical and practical objectives in understanding the relationships between known and unknown attributes, and in making interventions of addressing a condition. Researchers can benefit from theory-driven computational frameworks that incorporate the predictive capabilities of both precision and generalizability, as well as the explainability in feature interpretation and actionable insights.

8.3 Ethical and Privacy Implications

Our work has several ethical implications. We caution against our work being perceived or misused as a methodology to facilitate surveillance or profiling users on their offline behaviors [85]. We advocate balancing the costs and benefits of such systems with an emphasis on privacy-preservation. As Pandit and Lewis describe, "the use of personal data is a double-edged sword that on one side provides benefits through personalisation and user profiling, while the other raises several ethical and moral implications that impede technological progress" [85]. While physical sensor data may be better anonymized and secured compared to demographic information, it is still critical to ensure that the data is simultaneously useful and privacy-preserving.

Our work clusters individuals on their physical activity data as collected via passive sensors — while these clusters can be characterized on different variables, they may not necessarily translate to mappable individual characteristics, and there is no particular means to simply label them as "desirable" or "undesirable". We caution against such misinterpretation and misuse because these can bear consequences. For instance, a possible misuse in workplace contexts could be clustering employees on their behaviors such as work-times and routines, then characterizing them on their productivity, proactivity, and pro-socialness, followed by rewarding and penalizing employees on such characterizations. Any such empirical analyses require careful and in-depth supplemental ethical analyses before enacting any inference and decision-related outcomes.

Finally, with the ubiquity of digital data, our approach of personalizing predictions could be adopted in various contexts, including ones without awareness or consent of individuals. This forms a part of larger discussions on ethical and responsible use of data which require forthcoming discussions among ubiquitous computing researchers, ethicists, and practitioners to understand and respect the individual perspectives on such use of data, which can start from those who choose to participate (or not participate) in multimodal sensing studies [100].

8.4 Limitations and Future Directions

Our work has limitations, some of which also open up opportunities for future research. Our findings are limited to studying trait-based constructs. Future work should examine the effectiveness in state-based measures. Our predictions are based on a single social media platform (Facebook), and we only study those who are on Facebook and chose to participate in the study, likely introducing self-selection bias. With the ubiquity and prevalence of several social media and online interaction platforms, incorporating other streams of data can augment predicting individual attributes. The participant pool of our study is drawn from information workers, which is vulnerable to biases within the dataset, such that certain measures are plausibly easier to predict. For instance, social media use varies across demographics and is skewed towards young adults [49], and the way social media use is distributed within the participants may not be representative of the population or social media landscape [81]. The generalizability of our observations remains to be explored if our approach would yield similar findings in other populations. We, therefore, caution against making sweeping generalizable claims.

We note that our work may be argued to not be "true personalization" — as we do not build one-for-each prediction models. While we elaborated on the motivation and advantages of considering a middle-ground between one-for-each and one-for-all models, we understand the value of conducting individualized models in certain contexts such as precision medicine. Again, one of our clusters was significantly larger than other clusters. Building on our approach and using hierarchical clustering would enable researchers to tune models for more

precise clusters. However, the standard empirical evaluations of clustering may not fully capture the "goodness of clusters". We suggest other evaluations of clusters, such as domain expertise. This is especially important in cases of unique or understudied populations and/or theoretical constructs.

The present work leveraged a specific range of passive sensors and assessed constructs. It remains a ripe future direction to investigate how different sensing streams can improve clustering individuals on physical behaviors. Similarly, other data modalities capturing external and localized effects on human behavior, such as weather data or community-level mood can be incorporated for more comprehensive predictions [64]. Importantly, while we chose to exclude demographic attributes in clustering individuals (with the objective of building cross-demographic representations), certain problems may require including such attributes, e.g., in the context of precision medicine and prescribing drugs, it might be critical to include age, gender, and underlying conditions. Depending on the problem setting, other modalities of data (including surveys and EMAs) may also be used to build person-centered models. Taken together, future work can study the efficacy, robustness, and generalizability of person-centered models from a variety of aspects — population, data modalities, and the problem of interest.

9 CONCLUSION

This paper applied a person-centered approach to predict psychological constructs with social media by leveraging multimodal sensing. We conducted our study on a longitudinal dataset of 754 participants, and we situated our modeling on social media data to predict cognitive ability, personality traits, affect, and wellbeing constructs. We first clustered individuals on their mutable behaviors such as physical activity, commute, phone use, sleep, and work behaviors collected via tangible passive sensors such as smartphones, wearables, and Bluetooth beacons. Then, for each cluster, we built *contextualized* prediction models that used cluster-attuned social media data as features and psychological constructs as dependent variables. We compared their performance against *generalized* models trained on the entire social media data of all participants. We found no significant difference in affect, however, the *generalized* models performed significantly better in predicting cognitive ability, whereas *contextualized* models performed significantly better in predicting personality traits, anxiety, and sleep quality. We conjectured that contextualized models were effective in predicting constructs that theoretically shared strong relationships with physical behaviors. In contrast, predictions of constructs sharing weaker relationships were likely impacted due to the minimized training data in contextualized models than generalized models. We discussed the implications of our work regarding the nuances of personalized predictions, and how their effectiveness may vary by construct, problem setting, and other factors.

REFERENCES

[1] Firoj Alam and Giuseppe Riccardi. 2014. Predicting personality traits using multimodal information. In *Proceedings of the 2014 ACM multi media on workshop on computational personality recognition*. 15–18.

[2] Tim Althoff, Pranav Jindal, and Jure Leskovec. 2017. Online actions with offline impact: How online social networks influence online and offline user behavior. In *Proceedings of the tenth ACM international conference on web search and data mining*. 537–546.

[3] Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. 2008. Influence and correlation in social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 7–15.

[4] Ionut Andone, Konrad Błaszkiewicz, Mark Eibes, Boris Trendafilov, Christian Montag, and Alexander Markowetz. 2016. How age and gender affect smartphone usage. In *Proc. ACM international joint conference on pervasive and ubiquitous computing: adjunct*.

[5] Activity Recognition API. 2018. https://developers.google.com/location-context/activity-recognition/. Accessed: 2018-11-01.

[6] Garmin Health API. 2018. http://developer.garmin.com/health-api/overview/. Accessed: 2018-11-01.

[7] Manager REST API. 2018. https://docs.gimbal.com/rest.html. Accessed: 2018-11-01.

[8] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences* 116, 32 (2019), 15849–15854.

[9] Stuart JH Biddle, Ken Fox, and Steve Boutcher. 2003. *Physical activity and psychological well-being*. Routledge.

[10] Mehrab Bin Morshed, Koustuv Saha, Richard Li, Sidney K. D'Mello, Munmun De Choudhury, Gregory D. Abowd, and Thomas Plötz. 2019. Prediction of Mood Instability with Passive Sensing. *PACM IMWUT* (2019).

[11] Timothy A Brown, Bruce F Chorpita, and David H Barlow. 1998. Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. *Journal of abnormal psychology* 107, 2

(1998), 179.
[12] Aron S Buchman, Robert S Wilson, Lei Yu, Bryan D James, Patricia A Boyle, and David A Bennett. 2014. Total daily activity declines more rapidly with increasing age in older adults. *Archives of gerontology and geriatrics* 58, 1 (2014), 74–79.
[13] Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. 1293–1304.
[14] Hancheng Cao, Zhilong Chen, Fengli Xu, Yong Li, and Vassilis Kostakos. 2018. Revisitation in urban space vs. online: A comparison across POIs, websites, and smartphone apps. *PACM IMWUT* (2018).
[15] Ralph Catalano. 1979. *Health, behavior and the community: An ecological perspective*. Pergamon Press New York.
[16] Raymond B Cattell. 1963. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of educational psychology* 54, 1 (1963), 1.
[17] Larry Chan, Vedant Das Swain, Christina Kelley, Kaya de Barbaro, Gregory D Abowd, and Lauren Wilcox. 2018. Students' Experiences with Ecological Momentary Assessment Tools to Report on Emotional Well-being. *IMWUT* (2018).
[18] Chih-Kuang Chen, Yu-Cheng Pei, Ning-Hung Chen, Li-Ting Huang, Shih-Wei Chou, Katie P Wu, Pei-Chih Ko, Alice MK Wong, and Chih-Kuan Wu. 2014. Sedative music facilitates deep sleep in young adults. *Journal of Alternative and Complementary Medicine* (2014).
[19] Cindy Chung and James W Pennebaker. 2007. The psychological functions of function words. *Social communication* (2007), 343–359.
[20] Chia-Fang Chung, Qiaosi Wang, Jessica Schroeder, Allison Cole, Jasmine Zia, James Fogarty, and Sean A Munson. 2019. Identifying and planning for individualized change: Patient-provider collaboration using lightweight food diaries in healthy eating and irritable bowel syndrome. *PACM IMWUT* (2019).
[21] Glen Coppersmith, Craig Harman, and Mark Dredze. 2014. Measuring post traumatic stress disorder in Twitter. In *International AAAI Conference on Web and Social Media*.
[22] Rex William Cowdry, David L Gardner, et al. 1991. Mood variability: A study of four groups. *The American Journal of Psychiatry* (1991).
[23] Aron Culotta. 2014. Estimating county health statistics with Twitter. In *Proc. CHI*. 1335–1344.
[24] Vedant Das Swain et al. 2019. A Multisensor Person-Centered Approach to Understand the Role of Daily Activities in Job Performance with Organizational Personas. *PACM IMWUT* (2019).
[25] Vedant Das Swain, Koustuv Saha, Manikanta D Reddy, Hemang Rajvanshy, Gregory D Abowd, and Munmun De Choudhury. 2020. Modeling Organizational Culture with Workplace Experiences Shared on Glassdoor. In *CHI*. ACM.
[26] Nediyana Daskalova, Bongshin Lee, Jeff Huang, Chester Ni, and Jessica Lundin. 2018. Investigating the effectiveness of cohort-based sleep recommendations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–19.
[27] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 3267–3276.
[28] Munmun De Choudhury, Scott Counts, Eric Horvitz, and Aaron Hoff. 2014. Characterizing and Predicting Postpartum Depression from Facebook Data. In *Proc. CSCW*.
[29] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *ICWSM*.
[30] Mariella Dimiccoli, Juan Marín, and Edison Thomaz. 2018. Mitigating Bystander Privacy Concerns in Egocentric Activity Recognition with Deep Learning and Intentional Image Degradation. *IMWUT* (2018).
[31] Yuriko Doi, Masumi Minowa, Makoto Uchiyama, Masako Okawa, Keiko Kim, Kayo Shibui, and Yuichi Kamei. 2000. Psychometric assessment of subjective sleep quality using the Japanese version of the Pittsburgh Sleep Quality Index (PSQI-J) in psychiatric disordered and control subjects. *Psychiatry research* 97, 2-3 (2000), 165–172.
[32] Carsten F Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J Leitão, et al. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36, 1 (2013), 27–46.
[33] Finale Doshi-Velez, Yaorong Ge, and Isaac Kohane. 2014. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics* 133, 1 (2014), e54–e63.
[34] Joseph C Dunn. 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. (1973).
[35] Olive Jean Dunn and Virginia Clark. 1971. Comparison of tests of the equality of dependent correlation coefficients. *J. Amer. Statist. Assoc.* 66, 336 (1971), 904–908.
[36] Sindhu Kiranmai Ernala, Asra F Rizvi, Michael L Birnbaum, John M Kane, and Munmun De Choudhury. 2017. Linguistic markers indicating therapeutic outcomes of social media disclosures of schizophrenia. *PACM Human-Computer Interaction* CSCW (2017).
[37] Deborah Estrin. 2014. Small data, where n= me. *Commun. ACM* 57, 4 (2014), 32–34.
[38] Brian Everitt and Anders Skrondal. 2002. *The Cambridge dictionary of statistics*. Vol. 106. Cambridge University Press Cambridge.
[39] Jiayin Fan et al. 2020. Relationships between Five-Factor Personality Model and Anxiety: The Effect of Conscientiousness on Anxiety. *Open Journal of Social Sciences* 8, 08 (2020), 462.
[40] Yali Fan, Zhen Tu, Yong Li, Xiang Chen, Hui Gao, Lin Zhang, Li Su, and Depeng Jin. 2019. Personalized Context-aware Collaborative Online Activity Prediction. *PACM IMWUT* (2019).
[41] Katya C Fernandez, Aaron J Fisher, and Cyrus Chi. 2017. Development and initial implementation of the Dynamic Assessment Treatment Algorithm (DATA). *PloS one* 12, 6 (2017), e0178806.
[42] Robert J Fisher. 1993. Social desirability bias and the validity of indirect questioning. *Journal of consumer research* 20, 2 (1993), 303–315.

[43] Barbara L Fredrickson. 2000. Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions. *Cognition & Emotion* 14, 4 (2000), 577–606.

[44] Jon Froehlich, Mike Y Chen, Sunny Consolvo, Beverly Harrison, and James A Landay. 2007. MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones. In *Proceedings of the 5th international conference on Mobile systems, applications and services*. ACM, 57–70.

[45] Kirstine Rosenbeck Gøeg, Ronald Cornet, and Stig Kjær Andersen. 2015. Clustering clinical models from local electronic health records based on semantic similarity. *Journal of biomedical informatics* 54 (2015), 294–304.

[46] Scott A Golder and Michael W Macy. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333, 6051 (2011), 1878–1881.

[47] Fernando Gomez. 2016. A Guide to the Depression, Anxiety and Stress Scale (DASS 21).

[48] Samuel D Gosling, Adam A Augustine, Simine Vazire, Nicholas Holtzman, and Sam Gaddis. 2011. Manifestations of personality in online social networks: Self-reported Facebook-related behaviors and observable profile information. *Cyberpsychology, Behavior, and Social Networking* 14, 9 (2011), 483–488.

[49] Shannon Greenwood, Andrew Perrin, and Maeve Duggan. 2016. Demographics of Social Media Users in 2016. http://www.pewinternet.org/2016/11/11/social-media-update-2016/. Accessed: 2017-02-12.

[50] Ted Grover and Gloria Mark. 2017. Digital footprints: Predicting personality from temporal patterns of technology use. 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*. 41–44.

[51] Xiaonan Guo, Jian Liu, Cong Shi, Hongbo Liu, Yingying Chen, and Mooi Choo Chuah. 2018. Device-free personalized fitness assistant using WiFi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–23.

[52] Julia Handl, Joshua Knowles, and Douglas B Kell. 2005. Computational cluster validation in post-genomic data analysis. *Bioinformatics* 21, 15 (2005), 3201–3212.

[53] Hoda Heidari, Vedant Nanda, and Krishna Gummadi. 2019. On the Long-term Impact of Algorithmic Decision Policies: Effort Unfairness and Feature Segregation through Social Learning. In *International Conference on Machine Learning*. 2692–2701.

[54] Eric Hekler, Jasmin A Tiro, Christine M Hunter, and Camille Nebeker. 2020. Precision health: The role of the social and behavioral sciences in advancing the vision. *Annals of Behavioral Medicine* (2020).

[55] Robin K Henson and J Kyle Roberts. 2006. Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educ. Psychol. Meas.* (2006).

[56] Matt C Howard and Michael E Hoffman. 2018. Variable-centered, person-centered, and person-specific approaches: where theory meets the method. *Organizational Research Methods* 21, 4 (2018), 846–876.

[57] Wan Nurul Izza Wan Husina, Angeli Santosa, Hazel Melanie Ramosa, and Mohamad Sahari Nordinb. 2013. Crystallized intelligence or fluid intelligence factor? *Procedia-Social and Behavioral Sciences* 97 (2013), 214–223.

[58] Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 49–58.

[59] Raja Jurdak, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth. 2015. Understanding human mobility from Twitter. *PloS one* 10, 7 (2015), e0131469.

[60] Seung-Gul Kang, Jae Myeong Kang, Kwang-Pil Ko, Seon-Cheol Park, Sara Mariani, and Jia Weng. 2017. Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. *Journal of psychosomatic research* 97 (2017), 38–44.

[61] Meera Komarraju, Steven J Karau, Ronald R Schmeck, and Alen Avdic. 2011. The Big Five personality traits, learning styles, and academic achievement. *Personality and individual differences* 51, 4 (2011), 472–477.

[62] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences* (2013).

[63] Brett Laursen, Wyndol Furman, and Karen S Mooney. 2006. Predicting interpersonal competence and self-worth from adolescent relationships and relationship networks: Variable-centered and person-centered perspectives. *Merrill-Palmer Quarterly (1982-)* (2006).

[64] Boning Li and Akane Sano. 2020. Extraction and Interpretation of Deep Autoencoder-based Temporal Features from Wearables for Forecasting Personalized Mood, Health, and Stress. *PACM IMWUT* (2020).

[65] Peng Liao, Kristjan Greenewald, Predrag Klasnja, and Susan Murphy. 2020. Personalized HeartSteps: A Reinforcement Learning Algorithm for Optimizing Physical Activity. *PACM IMWUT* (2020).

[66] Jason Liu, Elissa R Weitzman, and Rumi Chunara. 2017. Assessing behavioral stages from social media data. In *CSCW*.

[67] Jin Lu, Chao Shang, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jayesh Kamath, Athanasios Bamis, Alexander Russell, Bing Wang, and Jinbo Bi. 2018. Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. *PACM IMWUT* (2018).

[68] Richard E Lucas, Ed Diener, Alexander Grob, Eunkook M Suh, and Liang Shao. 2000. Cross-cultural evidence for the fundamental features of extraversion. *Journal of personality and social psychology* 79, 3 (2000), 452.

[69] Richard E Lucas and Frank Fujita. 2000. Factors influencing the relation between extraversion and pleasant affect. *Journal of personality and social psychology* 79, 6 (2000), 1039.

[70] David Magnusson. 1998. *The logic and implications of a person-oriented approach.* Sage Publications, Inc.

[71] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 55–60.

[72] Pierre Maquet, Jean-Marie Péters, Joël Aerts, Guy Delfiore, Christian Degueldre, André Luxen, and Georges Franck. 1996. Functional neuroanatomy of human rapid-eye-movement sleep and dreaming. *Nature* 383, 6596 (1996), 163–166.

[73] Gloria Mark, Yiran Wang, Melissa Niiya, and Stephanie Reich. 2016. Sleep Debt in Student Life: Online Attention Focus, Facebook, and Mood. In *Proc. 2016 CHI Conference on Human Factors in Computing Systems*. 5517–5528.

[74] Gonzalo J Martinez, Stephen M Mattingly, Jessica Young, Louis Faust, Anind K Dey, Andrew T Campbell, Munmun De Choudhury, Shayan Mirjafari, Subigya Nepal, Pablo Robles-Granda, et al. 2020. Improved Sleep Detection Through The Fusion of Phone Agent and Wearable Data Streams. In *WristSense 2020*.

[75] Stephen M Mattingly et al. 2019. The Tesserae Project: Large-Scale, Longitudinal, In Situ, Multimodal Sensing of Information Workers. In *CHI Ext. Abstracts*.

[76] Abhinav Mehrotra, Veljko Pejovic, and Mirco Musolesi. 2014. SenSocial: a middleware for integrating online social networks and mobile sensing data streams. In *Proceedings of the 15th International Middleware Conference*. ACM, 205–216.

[77] Gregory J Meyer, Stephen E Finn, Lorraine D Eyde, Gary G Kay, Kevin L Moreland, Robert R Dies, Elena J Eisman, Tom W Kubiszyn, and Geoffrey M Reed. 2001. Psychological testing and psychological assessment: A review of evidence and issues. *American psychologist* 56, 2 (2001), 128.

[78] Jeremy Miles. 2014. Tolerance and variance inflation factor. *Wiley StatsRef: Statistics Reference Online* (2014).

[79] Shayan Mirjafari, Kizito Masaba, Ted Grover, Weichen Wang, Pino Audia, Andrew T. Campbell, Nitesh V. Chawla, Vedant Das Swain, Munmun De Choudhury, Anind K. Dey, and et al. 2019. Differentiating Higher and Lower Job Performers in the Workplace Using Mobile Sensing. *Proc. ACM IMWUT* (2019).

[80] NIH. 2020. https://ghr.nlm.nih.gov/primer/precisionmedicine/definition. Accessed: 2020-08-07.

[81] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.

[82] Jukka-Pekka Onnela and Scott L Rauch. 2016. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology* 41, 7 (2016), 1691.

[83] Johan Ormel, A Bastiaansen, Harriëtte Riese, Elisabeth H Bos, Michelle Servaas, Mark Ellenbogen, Judith GM Rosmalen, and André Aleman. 2013. The biological and psychological basis of neuroticism: current status and future directions. *Neuroscience & Biobehavioral Reviews* 37, 1 (2013), 59–72.

[84] Augustine Osman, Jane L Wong, Courtney L Bagge, Stacey Freedenthal, Peter M Gutierrez, and Gregorio Lozano. 2012. The depression anxiety stress Scales—21 (DASS-21): further examination of dimensions, scale reliability, and correlates. *Journal of clinical psychology* 68, 12 (2012), 1322–1338.

[85] Harshvardhan J Pandit and Dave Lewis. 2018. Ease and ethics of user profiling in black mirror. In *Companion Proceedings of the The Web Conference 2018*. 1577–1583.

[86] 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* (2015).

[87] Edward H Patzelt, Catherine A Hartley, and Samuel J Gershman. 2018. Computational phenotyping: using models to understand individual differences in personality, development, and mental illness. *Personality Neuroscience* 1 (2018).

[88] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54, 1 (2003), 547–577.

[89] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[90] John P Pollak, Phil Adams, and Geri Gay. 2011. PAM: a photographic affect meter for frequent, in situ measurement of affect. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 725–734.

[91] Lin Qiu, Han Lin, Jonathan Ramsay, and Fang Yang. 2012. You are what you tweet: Personality expression and perception on Twitter. *Journal of research in personality* 46, 6 (2012), 710–718.

[92] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 2011. Our twitter profiles, our selves: Predicting personality with twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE, 180–185.

[93] Mashfiqui Rabbi, Min Hane Aung, Mi Zhang, and Tanzeem Choudhury. 2015. MyBehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. In *Proc. Ubicomp*.

[94] LS Radloff. 1977. Center for epidemiological studies – Depression scale. *Actualisation de la base de données BeST & Inclusion de nouvelles échelles dans la base de données existante BeST* (1977).

[95] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. 2018. Multimodal deep learning for activity and context recognition. *IMWUT* (2018).

[96] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 469–481.

[97] Shriti Raj, Joyce M Lee, Ashley Garrity, and Mark W Newman. 2019. Clinical data in context: towards Sensemaking tools for interpreting personal health data. *PACM IMWUT* (2019).

[98] Ryan E Rhodes and NEI Smith. 2006. Personality correlates of physical activity: a review and meta-analysis. *British journal of sports medicine* 40, 12 (2006), 958–965.

[99] Pablo Robles-Granda, Suwen Lin, Xian Wu, Sidney D'Mello, Gonzalo J Martinez, Koustuv Saha, Kari Nies, Gloria Mark, Andrew T Campbell, Munmun De Choudhury, et al. 2020. Jointly Predicting Job Performance, Personality, Cognitive Ability, Affect, and Well-Being. *arXiv preprint arXiv:2006.08364* (2020).

[100] John Rooksby, Alistair Morrison, and Dave Murray-Rust. 2019. Student perspectives on digital phenotyping: The acceptability of using smartphone data to assess mental health. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[101] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.

[102] Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Björn Schuller, and Rosalind W Picard. 2018. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics* 3, 19 (2018).

[103] Reza Safdari, Elham Maserat, Hamid Asadzadeh Aghdaei, et al. 2017. Person centered prediction of survival in population based screening program by an intelligent clinical decision support system. *Gastroenterology and Hepatology from bed to Bench* (2017).

[104] Koustuv Saha et al. 2019. Imputing Missing Social Media Data Stream in Multisensor Studies of Human Behavior. In *Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII 2019)*.

[105] Koustuv Saha et al. 2019. Social Media as a Passive Sensor in Longitudinal Studies of Human Behavior and Wellbeing. In *CHI Ext. Abstracts*. ACM.

[106] Koustuv Saha, Larry Chan, Kaya De Barbaro, Gregory D Abowd, and Munmun De Choudhury. 2017. Inferring mood instability on social media by leveraging ecological momentary assessments. *PACM IMWUT* (2017).

[107] Koustuv Saha and Munmun De Choudhury. 2017. Modeling stress with social media around incidents of gun violence on college campuses. *PACM Human-Computer Interaction* CSCW (2017).

[108] Koustuv Saha, Yozen Liu, Nicholas Vincent, Farhan Asif Chowdhury, Leonardo Neves, Neil Shah, and Maarten W Bos. 2021. AdverTiming Matters: Examining User Ad Consumption for Effective Ad Allocations on Social Media. In *Proc. CHI*.

[109] Koustuv Saha, Ingmar Weber, and Munmun De Choudhury. 2018. A Social Media Based Examination of the Effects of Counseling Recommendations After Student Deaths on College Campuses. In *ICWSM*.

[110] Muhammad Zubair Satti, Tayyab Mumtaz Khan, et al. 2019. Association of physical activity and sleep quality with academic performance among fourth-year MBBS students of Rawalpindi Medical University. *Cureus* 11, 7 (2019).

[111] Stefan Schipolowski, Oliver Wilhelm, and Ulrich Schroeders. 2014. On the nature of crystallized intelligence: The relationship between verbal ability and factual knowledge. *Intelligence* 46 (2014), 156–168.

[112] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8, 9 (2013), e73791.

[113] H Andrew Schwartz, Maarten Sap, Margaret L Kern, Johannes C Eichstaedt, Adam Kapelner, Megha Agrawal, Eduardo Blanco, Lukasz Dziurzynski, Gregory Park, David Stillwell, et al. 2016. Predicting individual well-being through the language of social media. In *Pac Symp Biocomput*, Vol. 21. 516–527.

[114] Christie Napa Scollon, Chu-Kim Prieto, and Ed Diener. 2009. Experience sampling: promises and pitfalls, strength and weaknesses. In *Assessing well-being*. Springer, 157–180.

[115] Andriy Shepitsen, Jonathan Gemmell, Bamshad Mobasher, and Robin Burke. 2008. Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM conference on Recommender systems*. 259–266.

[116] Walter C Shipley. 2009. *Shipley-2: manual*. WPS.

[117] Christopher J Soto and Oliver P John. 2017. The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology* 113, 1 (2017), 117.

[118] Charles D Spielberger, Fernando Gonzalez-Reigosa, Angel Martinez-Urrutia, Luiz FS Natalicio, and Diana S Natalicio. 2017. The state-trait anxiety inventory. *Revista Interamerican Journal of Psychology* (2017).

[119] Donna Spruijt-Metz and Wendy Nilsen. 2014. Dynamic models of behavior for just-in-time adaptive interventions. *IEEE Pervasive Computing* 13, 3 (2014), 13–17.

[120] Clemens Stachl, Quay Au, Ramona Schoedel, Samuel D Gosling, Gabriella M Harari, Daniel Buschek, Sarah Theres Völkel, Tobias Schuwerk, Michelle Oldemeier, Theresa Ullmann, et al. 2020. Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences* 117, 30 (2020), 17680–17687.

[121] Charles Steinfield, Nicole B Ellison, and Cliff Lampe. 2008. Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *JADP* (2008).

[122] Arthur A Stone, Joseph E Schwartz, John M Neale, Saul Shiffman, Christine A Marco, Mary Hickcox, Jean Paty, Laura S Porter, and Laura J Cruise. 1998. A comparison of coping assessed by ecological momentary assessment and retrospective recall. *Journal of personality and social psychology* 74, 6 (1998), 1670.

[123] Hyewon Suh, Nina Shahriaree, Eric B Hekler, and Julie A Kientz. 2016. Developing and Validating the User Burden Scale: A Tool for Assessing User Burden in Computing Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*.
[124] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.
[125] Robert P Tett, Douglas N Jackson, and Mitchell Rothstein. 1991. Personality measures as predictors of job performance: A meta-analytic review. *Personnel psychology* 44, 4 (1991), 703–742.
[126] Sara Thomée, Annika Härenstam, and Mats Hagberg. 2011. Mobile phone use and stress, sleep disturbances, and symptoms of depression among young adults-a prospective cohort study. *BMC public health* 11, 1 (2011), 66.
[127] Mark Van der Laan, Katherine Pollard, and Jennifer Bryan. 2003. A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation* 73, 8 (2003), 575–584.
[128] Hans Van Remoortel, Carlos Augusto Camillo, Daniel Langer, Miek Hornikx, Heleen Demeyer, Chris Burtin, Marc Decramer, Rik Gosselink, Wim Janssens, and Thierry Troosters. 2013. Moderate intense physical activity depends on selected Metabolic Equivalent of Task (MET) cut-off and type of data analysis. *PLoS One* 8, 12 (2013), e84365.
[129] Sebastian Wallot. 2017. Recurrence quantification analysis of processes and products of discourse: A tutorial in R. *Discourse Processes* (2017).
[130] 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*. 3–14.
[131] 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. *PACM IMWUT* (2018).
[132] Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* (1963).
[133] David Watson and Lee Anna Clark. 1999. The PANAS-X: Manual for the positive and negative affect schedule-expanded form. (1999).
[134] Charles L Webber Jr and Joseph P Zbilut. 2005. Recurrence quantification analysis of nonlinear dynamical systems. *Tutorials in contemporary nonlinear methods for the behavioral sciences* 94, 2005 (2005), 26–94.
[135] Aidan GC Wright and William C Woods. 2020. Personalized models of psychopathology. *Annual review of clinical psychology* 16 (2020).
[136] Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K Villalba, Janine M Dutcher, Michael J Tumminia, Tim Althoff, Sheldon Cohen, Kasey G Creswell, J David Creswell, et al. 2019. Leveraging Routine Behavior and Contextually-Filtered Features for Depression Detection among College Students. *PACM IMWUT* (2019).
[137] Daniel Yue Zhang, Rungang Han, Dong Wang, and Chao Huang. 2016. On robust truth discovery in sparse social media sensing. In *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 1076–1081.
[138] Melanie Zinkhan, Klaus Berger, Sabrina Hense, Maren Nagel, Anne Obst, Beate Koch, Thomas Penzel, Ingo Fietze, Wolfgang Ahrens, Peter Young, et al. 2014. Agreement of different methods for assessing sleep characteristics: a comparison of two actigraphs, wrist and hip placement, and self-report with polysomnography. *Sleep medicine* 15, 9 (2014), 1107–1114.

A APPENDIX

Table A1. **Generalized Models**: Predicting psychological constructs with social media using the entire data of all participants. Prediction algorithms used include Ridge, Elastic Net (ElNet), Support Vector Regressor (SVR), XGBoost (XGB), Gradient Boosted Random Forest (GBR), and Multilayer Perceptron Regressor (MLP). Reported accuracy numbers are Symmetric Mean Absolute Percentage Error (SMAPE) and Pearson's correlation coefficient (r), which are pooled in k-fold cross-validation (k=5). The bold-faced number in each row indicate the best performing model for that construct.

Construct	**Algorithm**											
	Ridge		**ElNet**		**SVR**		**XGB**		**GBR**		**MLP**	
	r	SMAPE	r	SMAPE	r	SMAPE	r	SMAPE	r	SMAPE	r	SMAPE
Cognitive Ability												
Shipley (Abstraction)	**0.25**	**6.81**	-0.19	6.78	0.22	6.75	0.18	6.85	0.23	6.78	0.09	17.75
Shipley (Vocabulary)	**0.29**	**4.13**	-0.14	4.33	0.24	4.14	0.18	4.28	0.22	4.24	0.14	10.04
Personality Traits												
Openness	**0.25**	**6.89**	-0.13	6.65	0.1	6.60	0.15	6.68	0.15	6.71	0.12	13.04
Conscientiousness	**0.13**	**7.29**	-0.14	7.07	0.08	7.02	0.04	7.34	0.06	7.28	0.04	11.35
Extraversion	0.13	8.93	-0.14	8.69	-0.06	8.69	0.17	8.61	**0.17**	**8.54**	0.14	12.54
Agreeableness	**0.17**	**5.84**	-0.15	5.78	-0.04	5.76	0.18	5.89	0.16	6.09	0.12	11.9
Neuroticism	**0.12**	**13.56**	-0.17	13.17	-0.14	13.11	0.05	13.59	0.06	13.37	0.09	13.87
Affect and Wellbeing												
Pos. Affect	0.07	7.27	-0.07	6.88	0.11	6.83	**0.13**	**7.10**	0.13	6.92	0.07	12.81
Neg. Affect	**0.11**	**10.90**	-0.17	10.89	-0.11	10.89	-0.05	11.51	-0.04	11.66	-0.0	20.08
Anxiety (STAI)	**0.12**	**9.66**	-0.14	9.66	-0.1	9.54	-0.06	10.2	-0.02	9.97	0.07	14.85
Sleep Quality (PSQI)	**0.15**	**16.02**	-0.14	15.52	-0.12	15.15	0.17	15.17	0.16	15.28	0.08	23.01

Table A2. **Contextualized Models**: Predicting psychological constructs with social media separately for each behaviorally contextualized clusters. Prediction algorithms used include Ridge, Elastic Net (ElNet), Support Vector Regressor (SVR), XGBoost (XGB), Gradient Boosted Random Forest (GBR), and Multilayer Perceptron Regressor (MLP). Reported accuracy numbers are Symmetric Mean Absolute Percentage Error (SMAPE) and Pearson's correlation coefficient (r), which are pooled in k-fold cross-validation (k=5). The bold-faced number in each row indicate the best performing model for that construct.

Construct	**Algorithm**											
	Ridge		**ElNet**		**SVR**		**XGB**		**GBR**		**MLP**	
	r	SMAPE	r	SMAPE	r	SMAPE	r	SMAPE	r	SMAPE	r	SMAPE
Cognitive Ability												
Shipley (Abstraction)	**0.23**	**6.88**	-0.17	6.82	-0.09	6.8	0.22	6.77	0.19	6.84	0.09	19.56
Shipley (Vocabulary)	**0.21**	**4.25**	0.03	4.33	0.03	4.16	0.21	4.25	0.28	4.11	0.13	15.82
Personality Traits												
Openness	**0.29**	**6.08**	0.01	6.66	0.07	6.6	0.15	6.75	0.13	6.81	0.08	22.3
Conscientiousness	**0.16**	**7.08**	-0.1	7.1	-0.0	7.04	0.06	7.25	0.08	7.24	-0.02	22.64
Extraversion	**0.21**	**8.46**	-0.07	8.72	-0.06	8.71	0.16	8.66	0.16	8.62	0.1	22.66
Agreeableness	**0.19**	**5.89**	-0.14	5.8	-0.06	5.78	0.1	6.14	0.13	6.0	0.02	21.74
Neuroticism	0.14	13.22	-0.15	13.22	-0.1	13.19	0.12	13.37	**0.18**	**13.09**	0.01	20.72
Affect and Wellbeing												
Pos. Affect	**0.14**	**6.90**	-0.01	6.88	0.04	6.82	0.07	7.31	0.04	7.35	0.06	15.25
Neg. Affect	**0.13**	**10.89**	0.01	10.87	0.0	10.8	0.03	11.26	0.01	11.36	0.01	22.23
Anxiety (STAI)	**0.21**	**8.51**	-0.04	9.68	-0.13	9.55	0.01	10.11	0.06	9.81	0.07	16.81
Sleep Quality (PSQI)	0.21	10.06	-0.05	15.49	-0.01	15.12	0.20	11.43	**0.25**	**10.59**	0.07	27.12

Table A3. **Generalized Models with PCA**: Predicting psychological constructs with social media using the entire data of all participants, after applying PCA-transformed features. Prediction algorithms used include Ridge, Elastic Net (ElNet), Support Vector Regressor (SVR), XGBoost (XGB), Gradient Boosted Random Forest (GBR), and Multilayer Perceptron Regressor (MLP). Reported accuracy numbers are Symmetric Mean Absolute Percentage Error (SMAPE) and Pearson's correlation coefficient (r), which are pooled in k-fold cross-validation (k=5). The bold-faced number in each row indicate the best performing model for that construct.

Construct	**Algorithm**											
	Ridge		**ElNet**		**SVR**		**XGB**		**GBR**		**MLP**	
	r	SMAPE	r	SMAPE	r	SMAPE	r	SMAPE	r	SMAPE	r	SMAPE
Cognitive Ability												
Shipley (Abstraction)	0.10	11.95	-0.19	6.78	**0.22**	**6.68**	0.21	6.86	0.2	6.95	-0.06	22.44
Shipley (Vocabulary)	0.11	7.57	-0.09	4.32	**0.23**	**4.10**	0.22	4.4	0.22	4.42	0.12	29.34
Personality Traits												
Openness	0.12	10.89	-0.13	6.65	**0.23**	**6.40**	0.16	6.68	0.16	6.64	0.14	15.87
Conscientiousness	0.03	11.25	-0.14	7.07	**0.13**	**6.97**	0.12	7.13	0.11	7.19	0.08	16.22
Extraversion	0.11	13.43	-0.14	8.69	**0.20**	**8.47**	0.17	8.73	0.19	8.66	0.1	16.25
Agreeableness	0.11	10.1	-0.15	5.78	**0.20**	**5.66**	0.11	6.01	0.12	5.98	0.11	14.02
Neuroticism	0.02	22.47	-0.17	13.17	-0.02	13.29	0.07	13.54	0.05	13.48	**0.09**	**21.8**
Affect and Wellbeing												
Pos. Affect	0.05	11.38	-0.07	6.88	**0.13**	**6.79**	0.09	7.15	0.03	7.24	0.03	29.89
Neg. Affect	0.07	17.54	-0.17	10.89	-0.07	10.82	**0.10**	**11.29**	0.10	11.38	0.12	22.91
Anxiety (STAI)	**0.07**	**15.41**	-0.12	9.66	-0.0	9.51	0.05	9.93	0.0	10.06	0.10	30.92
Sleep Quality (PSQI)	0.04	26.74	-0.14	15.52	0.10	15.04	0.09	15.94	**0.12**	**15.68**	0.12	24.60

Table A4. **Contextualized Models with PCA**: Predicting psychological constructs with social media separately for each behaviorally contextualized clusters, after applying PCA-transformed features. Prediction algorithms used include Ridge, Elastic Net (ElNet), Support Vector Regressor (SVR), XGBoost (XGB), Gradient Boosted Random Forest (GBR), and Multilayer Perceptron Regressor (MLP). Reported accuracy numbers are Symmetric Mean Absolute Percentage Error (SMAPE) and Pearson's correlation coefficient (r), which are pooled in k-fold cross-validation (k=5). The bold-faced number in each row indicate the best performing model for that construct.

Construct	**Algorithm**											
	Ridge		**ElNet**		**SVR**		**XGB**		**GBR**		**MLP**	
	r	SMAPE	r	SMAPE	r	SMAPE	r	SMAPE	r	SMAPE	r	SMAPE
Cognitive Ability												
Shipley (Abstraction)	0.10	12.78	-0.05	6.91	0.14	6.75	**0.21**	**6.97**	0.16	7.11	-0.08	40.14
Shipley (Vocabulary)	0.03	8.74	0.09	4.31	0.14	4.14	0.17	4.49	**0.21**	**4.46**	0.06	51.14
Personality Traits												
Openness	0.14	12.16	-0.0	6.68	**0.24**	**6.45**	0.09	7.02	0.16	6.96	0.05	26.83
Conscientiousness	0.08	12.34	-0.06	7.10	**0.14**	**6.94**	0.05	7.53	0.03	7.6	0.06	27.91
Extraversion	0.06	14.02	0.02	8.69	0.16	8.63	**0.21**	**8.69**	0.24	8.54	0.11	26.64
Agreeableness	0.08	10.79	-0.15	5.8	**0.13**	**5.75**	0.05	6.23	0.06	6.24	0.09	27.29
Neuroticism	0.04	23.19	-0.12	13.31	0.05	13.17	0.04	13.66	0.04	13.73	**0.14**	**18.31**
Affect and Wellbeing												
Pos. Affect	0.08	11.95	0.06	6.88	0.08	6.81	0.14	7.22	**0.16**	**7.20**	0.01	49.12
Neg. Affect	**0.09**	**20.94**	-0.05	11.05	0.01	10.81	0.09	11.35	-0.0	11.69	0.02	38.29
Anxiety (STAI)	0.06	16.05	-0.0	9.77	-0.01	9.5	**0.10**	**9.88**	0.15	9.68	0.09	52.22
Sleep Quality (PSQI)	0.05	28.73	0.05	15.53	**0.18**	**11.00**	0.05	16.23	0.05	16.33	0.06	34.74

Table A5. **Physical Sensor based Models**: Predicting psychological constructs with only physical sensor based features. Prediction algorithms used include Ridge, Elastic Net (ElNet), Support Vector Regressor (SVR), XGBoost (XGB), Gradient Boosted Random Forest (GBR), and Multilayer Perceptron Regressor (MLP). Reported accuracy numbers are Symmetric Mean Absolute Percentage Error (SMAPE) and Pearson's correlation coefficient (r), which are pooled in k-fold cross-validation (k=5). The bold-faced number in each row indicate the best performing model for that construct.

Construct	**Algorithm**											
	Ridge		**ElNet**		**SVR**		**XGB**		**GBR**		**MLP**	
	r	SMAPE	r	SMAPE	r	SMAPE	r	SMAPE	r	SMAPE	r	SMAPE
Cognitive Ability												
Shipley (Abstraction)	**0.12**	**6.81**	-0.19	6.78	0.03	6.75	0.11	7.17	0.06	7.31	-0.06	8.01
Shipley (Vocabulary)	0.10	4.33	-0.14	4.33	**0.12**	**4.13**	-0.0	4.76	0.03	4.77	-0.02	5.62
Personality Traits												
Openness	0.00	6.87	-0.13	6.65	-0.05	6.74	**0.02**	**7.02**	0.01	7.01	-0.02	8.05
Conscientiousness	0.12	7.89	-0.14	7.07	**0.13**	**7.86**	0.12	7.29	0.12	7.21	-0.01	8.21
Extraversion	**0.17**	**8.41**	-0.14	8.69	0.13	8.43	0.12	8.69	0.16	8.69	0.12	9.3
Agreeableness	0.08	5.94	-0.15	5.78	**0.10**	**5.73**	0.03	6.09	0.0	6.12	-0.03	7.13
Neuroticism	**0.12**	**12.94**	-0.17	13.17	0.13	12.93	0.13	13.13	0.12	13.22	0.08	14.34
Affect and Wellbeing												
Pos. Affect	**0.11**	**7.90**	-0.07	6.88	0.11	7.78	0.14	7.03	0.10	7.06	0.08	7.10
Neg. Affect	**0.09**	**11.88**	-0.17	10.89	-0.01	10.81	0.06	11.42	0.08	11.47	0.08	11.26
Anxiety (STAI)	0.11	9.45	-0.14	9.66	0.04	9.50	0.09	9.98	**0.14**	**9.78**	0.09	9.92
Sleep Quality (PSQI)	**0.17**	**16.28**	-0.14	15.52	0.13	15.96	0.15	16.63	0.14	16.71	0.06	16.94

Table A6. **Mixed-effects Models**: Predicting psychological constructs with both physical activity and social media features together. Prediction algorithms used include Ridge, Elastic Net (ElNet), Support Vector Regressor (SVR), XGBoost (XGB), Gradient Boosted Random Forest (GBR), and Multilayer Perceptron Regressor (MLP). Reported accuracy numbers are Symmetric Mean Absolute Percentage Error (SMAPE) and Pearson's correlation coefficient (r), which are pooled in k-fold cross-validation (k=5). The bold-faced number in each row indicate the best performing model for that construct.

Construct	**Algorithm**											
	Ridge		**ElNet**		**SVR**		**XGB**		**GBR**		**MLP**	
	r	SMAPE	r	SMAPE	r	SMAPE	r	SMAPE	r	SMAPE	r	SMAPE
Cognitive Ability												
Shipley (Abstraction)	0.15	11.96	-0.19	6.78	**0.25**	**6.65**	0.16	6.96	0.18	6.95	-0.02	22.2
Shipley (Vocabulary)	0.14	8.01	-0.11	4.33	0.25	4.10	0.26	4.36	**0.28**	**4.33**	0.12	29.37
Personality Traits												
Openness	0.13	11.14	-0.13	6.65	**0.26**	**6.40**	0.16	6.65	0.17	6.63	0.16	15.13
Conscientiousness	0.09	10.93	-0.14	7.07	**0.17**	**7.86**	0.17	8.03	0.17	7.99	0.07	15.15
Extraversion	0.03	13.98	-0.14	8.69	0.17	8.37	**0.17**	**8.35**	0.17	8.57	0.09	17.52
Agreeableness	0.07	9.88	-0.15	5.78	**0.17**	**5.91**	0.16	5.92	0.11	5.94	0.14	14.37
Neuroticism	0.10	21.18	-0.17	13.17	**0.11**	**12.93**	0.07	13.45	0.08	13.49	0.13	19.59
Affect and Wellbeing												
Pos. Affect	0.07	11.19	-0.0	6.87	**0.13**	**6.77**	0.13	6.95	0.13	6.85	0.07	29.51
Neg. Affect	0.13	17.82	-0.17	10.89	-0.05	10.83	**0.13**	**11.18**	0.11	11.23	0.12	22.95
Anxiety (STAI)	**0.13**	**16.07**	-0.08	9.65	0.02	9.50	-0.0	10.15	0.07	9.91	0.13	30.71
Sleep Quality (PSQI)	0.13	27.64	-0.14	15.52	0.20	14.9	0.19	15.15	0.19	15.28	**0.21**	**23.22**