Observer Effect in Social Media Use

Koustuv Saha (koustuvsaha@microsoft.com)
Microsoft Research https://orcid.org/0000-0002-8872-2934
Pranshu Gupta
Georgia Institute of Technology
Gloria Mark
University of California, Irvine
Emre Kıcıman
Microsoft Research https://orcid.org/0000-0001-5429-468X
Munmun De Choudhury
Georgia Institute of Technology https://orcid.org/0000-0002-8939-264X

Article

Keywords: social media, observer effect, hawthorne effect, human behavior, machine learning, language, self-presentation

Posted Date: January 19th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2492994/v1

Additional Declarations: There is **NO** Competing Interest.

Observer Effect in Social Media Use

Koustuv Saha[1*], Pranshu Gupta[2], Gloria Mark[3], Emre Kıcıman[4], and Munmun De Choudhury[2]

[1]Microsoft Research, Microsoft, Montréal, Québec, Canada
[2]School of Interactive Computing, College of Computing, Georgia Institute of Technology, Atlanta, Georgia, USA
[3]Department of Informatics, Donald Bren School of Information and Computer Sciences, University of California, Irvine, California, USA
[4]Microsoft Research, Microsoft, Redmond, Washington, USA

[*]Corresponding Author:
Koustuv Saha
Microsoft Research Lab - Montréal
6795 Rue Marconi, Suite 400
Montréal, Québec, H2S 3J9
Canada
koustuvsaha@microsoft.com

ABSTRACT

Research has revealed the potential of social media as a source of large-scale, verbal, and naturalistic data for human behavior both in real-time and longitudinally. However, the in-practice utility of social media to assess and support wellbeing will only be realized when we account for extraneous factors. One such factor that might confound our ability to make inferences is the phenomenon of the "observer effect"—that individuals may deviate from their otherwise typical social media use because of the awareness of being monitored. This paper conducts a causal study to measure the observer effect in longitudinal social media use. We operationalized the observer effect in two dimensions of social media (Facebook) use—behavioral and linguistic changes. Participants consented to Facebook data collection over an average retrospective period of 82 months and an average prospective period of 5 months around the enrollment date to our study. We measured how they deviated from their expected social media use after enrollment. We obtained expected use by extrapolating from historical use using time-series (ARIMA) forecasting. We find that the deviation in social media use varies across individuals based on their psychological traits. Individuals with high cognitive ability and low neuroticism immediately decreased posting after enrollment, and those with high openness significantly increased posting. Linguistically, most individuals decreased the use of first-person pronouns, reflecting lowered sharing of intimate and self-attentional content. While some increased posting about public-facing events, others increased posting about family and social gatherings. We validate the observed changes with respect to psychological traits drawing from psychology and behavioral science theories, such as self-monitoring, public self-consciousness, and self-presentation. The findings provide recommendations to correct observer effects in social media data-driven assessments of human behavior.

Keywords: social media; observer effect; hawthorne effect; human behavior; machine learning; language; self-presentation

1 Introduction

The past decade has witnessed burgeoning research that has employed unobtrusively gathered social media data to infer a variety of behavioral and psychological attributes and states of individuals [1;2]. Harnessing rapid advancements in machine learning, Facebook data, for instance, can allow us to identify an individual's personality traits [3], or assess if they are at risk of forthcoming mental illness [4]. Research claims a lot of promise in these pursuits — algorithms developed with social media data can support designing wellbeing interventions [5], assisting decision-making [6;7], and providing actionable insights that have been difficult to gather through conventional social science methods that use self-reported information alone [8;9].

We note that most of the above research relies on retrospectively collected social media data — data that was created by subjects when they were unaware of the possibility of it being used for algorithmic inferences. For social media data-driven algorithms to be usable and useful in the real world, these algorithms would have to go beyond showcasing feasibility on retrospective data, to functioning accurately and reliably in prospective settings. However, multiple threads of recent research have argued how models trained on retrospective data do not necessarily translate well to the prospective setting due to issues of bias and non-representativeness [10;11;12]. Olteanu et al. [13] argued that the validity and in-practice reliability of human-centered

big data technologies suffer due to the unpredictability and complexity in human behavior along-with unaccounted confounds. Ruths and Pfeffer[12] noted that studies harnessing social media data might misrepresent or be ineffective in the real-world due to people's changing behaviors. Lazer et al.[11] similarly unpacked how the Google Flu predictor algorithm, which used Google search data, overestimated the number of flu visits in real-time, despite performing exceptionally well on historical data.

In addition, privacy concerns may arise when retrospective data, often without participant consent, is employed in making sensitive predictions from archival social media data. Fiesler and Proferes[14] surveyed how Twitter users felt about their historical data being used for research without their knowledge or awareness and found that the majority of respondents felt that researchers should not use postings without their consent. Duffy and Chan[15] found that social media users can alter their online self-presentation based on "imagined surveillance" on the platforms. In fact, scholars fear perceptions of surveillance when prospective research designs are adopted without participant awareness. For example, the Facebook emotion contagion study[16], which did not seek consent from people whose Facebook feeds were modified for experimental purposes, was heavily critiqued on ethical grounds[17]. Pertinent here is the position of boyd and Crawford, who noted that experiments conducted without participant awareness can reinforce the troubling perception of the technologies as "Big Brother, enabling invasions of privacy, decreased civil freedoms, and increased state and corporate control"[10].

An advocated solution to the issues centering around prospective research designs, where individuals are recruited with informed consent for data to be used in algorithms to infer behaviors and psychological states[18]. However, the prospective use of social media-based assessments poses new challenges, which are yet to be studied and addressed. It is to be noted that social media use is a form of intentional and conscious behavior or a behavior that individuals can alter at their will if they feel "observed" — changes that would be consistent with theories of social desirability, psychological reactance, self-presentation, and self-monitoring, to name a few[19;20;21]. The *observer effect* is the phenomenon that individuals might deviate from typical behaviors, attributed to the awareness of being "watched" or studied[22;23]. This phenomenon is also called the "research participation effect", the "experimenter effect", and the "Hawthorne effect"[24].

The social ecological model posits that human behavior is embedded in the complex interplay between an individual and their relationships, communities, and society[25]. While this theory explains the promise of social media as a viable source of naturalistic behavioral data, it points out a caveat—the observers (or researchers), who become a part of a subject's ecology, may affect the subject's behavior. Likewise, the ecological validity of these technologies and measurements remains unattested because the observer effect is not typically accounted for. The observer effect has been commonly cited to affect the reliability of observations in studies because it concerns research participation[26]. Consequently, McCambridge et al.[22] noted, "If there is a Hawthorne effect, studies could be biased in ways that we do not understand well, with profound implications for research[27]".

Social media experiments are also quite unique in comparison to traditional experiments. For instance, social media experiments are sensitive to people's social media use, and social media use happens in a naturalistic setting, which is intentional and conscious behavior that individuals can alter at their will. The likelihood of behavior change attributed to observer effect increases for conscious behaviors[28], as explained in prior research—Arkin and Shepperd noted self-consciousness influences one's strategic self-presentation[29] and Snyder noted people are likely to self-monitor their self-presentations, expressive behaviors, and non-verbal affective displays[21]. These are relevant and important aspects of social media use. Additionally, social media use comprises "social activity" and verbal and expressive behaviors. In contrast, traditional experiments primarily comprise personal activities that are undertaken in somewhat non-natural or even artificial settings. These differences together warrant studying the observer effect in social media experiments.

Motivated by the above, in this research, we ask — **does observer effect present itself in prospective studies of social media, and if so, to what extent and how?** We posit that quantifying the presence and degree of the observer effect can improve social media data-driven measurements. A better understanding of if and how observer effect exists in social media use, would further provide clarity to researcher expectations and support developing measures to account for this effect in study designs and findings in the computational social science field and its in-practice adaptations[12].

Our investigation used as a case study, a longitudinal, multi-disciplinary research effort, where 572 participants consented to social media (Facebook) data collection over a retrospective period of 82 months and a prospective period of 5 months from their enrollment date in the study. We operationalized observer effect along two dimensions of social media use comprising 266,320 Facebook postings, 1) behavioral changes and 2) linguistic changes. Our analytic approach draws on two lines of research: first, causal inference methods[30] to minimize the impacts of confounding factors on changes in social media use, and second, modeling approaches in psychology that use clustering on psychological traits to derive person-centric changes. In particular, we employed time-series and statistical modeling to measure how participants deviated from their expected behaviors after enrolling in the above study, or in response to their awareness of being "observed".

Our findings reveal that observer effect was indeed present, with posting behaviors of participants changing 17-34%, and linguistic attributes changing 4-57%. However, its occurrence varied across participants. For instance, individuals with high cognitive ability and low neuroticism showed an immediate decrease in social media posting after enrollment, but their behaviors got closer to expected over time. In contrast, individuals with high openness significantly increased posting quantity

despite not showing any immediate posting changes following enrollment. Linguistically, most individuals decreased using first person pronouns, which reflects reduced sharing of intimate and self-attentional content. This research bears implications for studies that harness prospective social media data, and we discuss directions to account for observer effect in social media study designs.

2 Results

Theoretically, the observer effect is a change in behavior because of being "observed"[22]. However, there are no established means to operationalize the observer effect, particularly in the context of social media use[1]. Our study, by design, considers enrollment in the study as the *treatment*, and therefore, does not include a comparison/control group as enrolling this group would have subjected them to the same treatment and likely introduced biases of measuring the observer effect. Instead, we draw on synthetic control based causal approaches[30] that suggest addressing the challenge of comparison group's unavailability by *synthetically* preparing control data through data-driven means. For all the participants, we employed time series modeling to predict *expected* post-enrollment social media use or the counterfactual data had they not enrolled in the study. Then, we measure the difference between the *expected* and *observed* post-enrollment data, which we operationalize as the observer effect in the social media use of individuals. Equation. 1 shows the operationalization of observer effect (α) for a participant i and time period T, where Y^o and Y^e represent observed and expected social media use, respectively.

$$\alpha_i[T] = Y_i^o[T] - Y_i^e[T] \tag{1}$$

To measure the observer effect in social media use, we needed to investigate individual-level changes due to the heterogeneity of social media behaviors. However, social media data is sparse and is prone to high variance across individuals; so, it is challenging to extrapolate from individual behaviors, but the extrapolation at a group level is more reliable[31]. Therefore, we adopted a middle ground between fully personalized and fully generalized approaches by clustering individuals on self-reported intrinsic traits and examining the changes per cluster. The clustering led to five clusters (C_0 - C_4) in our dataset. Materials and Methods gives an overview of the data used in this work, the analysis, and our validation procedures, and the Supplementary Information provides further details about the data and the analyses. Figure 1 shows the average distribution of the traits and Table 1 summarizes the characteristics of the five clusters.

This section describes our results— 1) first, we show our findings in terms of deviation in social media behaviors and language use where we report two kinds of results, short-term (two-weeks period) and long-term (100-days period) deviation, and 2) then, we validate and explain our findings with respect to intrinsic traits of individuals.

2.1 Findings of Observer Effect in Social Media Use

We exmaine if enrollment in our study *caused* the participants to change their social media use—the quantity and language of posts. We operationalized the observer effect as the post-enrollment deviation in the participants' *observed* social media use from *expected* use. We measured the deviation in social media use along the dimensions of (1) behavioral changes: posts made and engagement sought; and (2) linguistic changes: topics and psycholinguistics. We obtained expected post-enrollment social media use by extrapolating pre-enrollment behavioral trends into the 100-day post-enrollment period through time series-based modeling (auto-regressive integrated moving average or ARIMA).

Placebo Tests To conclude that the observed effects are *caused* by the *treatment* (study enrollment), we aimed to rule out the likelihood of effects by chance, by conducting placebo tests[32]. Here, the placebo tests are meant to rule out the likelihood that significant changes in social media use could also happen around dates other than the enrollment date (or placebo dates). We obtained 150 random permutations of "placebo" dates in the pre-enrollment data of the participants, and experimented with the same suit of time series analyses. We measured the statistical significance as per t-test in the deviation in observed and predicted time series data for each of the placebo date for each cluster. Out of 150 permutations, two clusters, C_0 and C_4, showed significance in 2 and 1 permutations, respectively, and the other three clusters showed no significant permutations. Therefore, the probability of a significant placebo effect is close to 0 for all the clusters, revealing that the significance observed around the *actual enrollment dates* (or *treatment*) is not by chance. This test also validates our extrapolation of expected behaviors in the post-enrollment period.

2.1.1 Deviation in Behavior

First, we extrapolated expected behaviors using ARIMA models using the pre-enrollment data of the participants, accounting for trends and seasonalities in time series. Then, we measured the deviation in the actual post-enrollment measures from the

[1] Note that this paper uses "social media use" as a phrase encompassing social media posting behaviors, engagements sought, and language.

expected measures. Table 2 and Table 3 summarize the model metrics and observations of changes in participants' social media use. Table 3 summarizes the slope changes in the time series of social media use from pre- to post-enrollment periods, along with causal impact computed as per[33]. High posterior probabilities of causal impacts (CI) indicate that the behaviors changed after enrollment in the study.

Changes in Posting Behavior To obtain expected posting behaviors, the ARIMA models predicting number of posts and words show mean symmetric mean absolute percentage errors (SMAPE) of 6.27 and 13.05, respectively. However, the deviation in the post-enrollment data between predicted and actual values is higher. In the 100-days post-enrollment data, clusters C_2 and C_3 showed statistically significant deviations in both quantity and verbosity of posts, i.e., they posted significantly more frequently and longer than their expected behaviors — C_2 showed average 17% higher and C_3 showed average 24% higher than expected quantity of posts. Focusing on the initial two-weeks post-enrollment, C_2 and C_3 showed similar (36% and 70%) increases in posting. C_0 and C_1 showed respectively 44% and 26% lower frequency of posting in the first two weeks, but, their posting behavior became closer to their expected posting behavior after the initial two weeks period. It is interesting to note that even though C_4 seemed to post greater than expected, their posting behavior had a decreasing trend (negative slope).C_4 individuals posted 41% shorter than expected posts in the initial two weeks. Figure 2 show cluster-wise deviations in actual and expected time series of number of posts.

Changes in Engagement Received The ARIMA models of engagements received predicting the expected number of comments and likes show mean SMAPEs of 20.76 and 15.15, respectively. In the 100-days post-enrollment period, C_3 received a mean 25% higher than expected likes and 22% higher than expected comments, and C_2 received a mean 29% higher than expected likes. The received engagements are likely correlated to these individuals' higher posting activity as noted above. Considering two-weeks' deviations, we find that C_2's posts received immediately higher quantity of comments (67%) and likes (96%), and C_4 received 27% lower than expected comments. C_0 and C_1 did not receive any significant deviations in the engagements received. Figure 3 shows example time series plots of how the number of likes received evolved per cluster.

2.1.2 Deviation in Language Use

Changes in Topical Themes We adopted a semi-automated approach of Latent-Dirichlet Allocation or LDA-based topic modeling[34] followed by manual annotation and interpretation to identify 10 topical themes in our dataset: 1) *Travel and Locations*, 2) *Food and Drinks*, 3) *Holiday Plans*, 4) *News and Information*, 5) *Work-Life Balance*, 6) *Family Gathering*, 7) *Social and Sports*, 8) *Greetings and Celebration*, 9) *Friends and Family*, and 10) *Activities and Interests*. Table 4 summarizes the relative change in topical prevalence from pre- to post-enrollment for each cluster.

Cluster C_0 increased posting about public-facing topics, such as travel, food, and news, increased posting about family gatherings but decreased posting about sports and celebratory events. Cluster C_1 increased posting about holiday plans, family gatherings, and celebratory events, but decreased posting about news-related content. Cluster C_2 shows the least changes in the expressiveness of content, with only decreased posting about food and social events. Cluster C_3 shows varied changes, with increased sharing about travel, food, and sports related content, whereas a decrease in more personal content such as holiday plans, work-life balance, family, and celebratory events. Finally, Cluster C_4 increased posting about food and family gatherings, whereas decreased posting about holiday plans, news, and interests-related content.

Changes in Psycholinguistic Use We examined the psycholinguistic changes in the clusters of participants. Table 5 shows the changes in psycholinguistic use, which we describe below:

Cluster C_0 (routine-oriented) individuals did not show any significant change in affective expressions except *anger*. In cognitive expressions, these participants increased using words related to *certainty*. In perception, *feel* and *see* decreased, whereas *hear* increased. They also decreased *first person singular pronoun* use but increased in *first personal plural pronoun* use. We also find a decrease in several function words, including *adverbs, verbs, auxiliary verbs, quantifiers*, and *relatives*. Among personal and social concerns, these individuals increased language relating to *achievement, home*, and *religion*.

Cluster C_1 (emotionally stable and innovative) individuals did not significantly change affective, cognitive, and perceptive expressions. Among function words, they decreased *second person pronouns* use, and increased *conjunction* and *inclusive* use. These participants also significantly increased the use of social words, including the categories of *family, friends, home*, and *religion*. This aligns with their topical changes post-enrollment. Therefore, C_1 individuals did not significantly change non-content word usage, but significantly changed content word usage: we could assume that they did not change "how" they write, but changed "what" they write.

Cluster C_2 (withdrawn, disagreeable, and prone to stress and irrtability) individuals significantly decreased language relating to a majority of affective, cognitive, and perceptive expressions, including *anger, anxiety, negative affect, positive affect, causation, certainty, cognitive mechanics, inhibition, percept*, and *see*. They decreased the use of first-person pronouns. In other function words, they decreased in *past* and *present tense*, *article, verbs, inclusive, preposition*, and *relative* use. Again, in personal and social concerns, they decreased the use of *friends, family*, and *home*. Together, these psycholinguistic changes

indicate that C_2 individuals inhibit sharing personal and self-expressive content, or prefer to share more about public-facing and less subjective content. This could be a sign of self-regulation among these individuals.

Cluster C_3 (positive, friendly, and wellbalanced) individuals significantly decreased using several affective, cognitive, and perceptive attributes. They also decreased using *first person singular pronouns*, suggesting lowered self-attentional focus, however, the use of *third person pronouns* significantly increased. These participants also decreased using many function words, including *adverbs, verbs*, and *prepositions*. In contrast to C_2, C_3 showed decreased *negative affect* and *swear* words and increased *positive affect* and *inclusive* keywords. We also find an increase in social words, such as *family*, *humans*, and *social*. These could be a manifestation of participants in this cluster wanting to self-present in a more socially desirable or positive way. The decrease in *work* keywords might suggest that the participants chose not to share work-related events on social media, particularly given that our study recruitment happened in a workplace context.

Cluster C_4 (curious and adventurous) individuals increased multiple affective expressions, including *anger*, *negative affect*, and *swear*, whereas a decrease in *positive affect*. Most cognitive and perceptive categories did not change, except for a significant decrease in *negation* and *feel*. These participants showed decreased *first person singular pronouns* usage, but increased *past tense* usage. Most other function words and social words did not significantly change, except there was a significant reduction in the use of *adverbs, preposition, relative*, and *bio*.

2.2 Validation of Observer Effect on Intrinsic Traits

Now, we aim to explain our observations through theories relating to individual differences and psychological traits. For each cluster, we examine the intrinsic characteristics, and evaluate the behavioral and linguistic changes as observed in the social media use, presumably subject to observer effect. We contextualize and interpret the findings by drawing upon the literature in psychology and behavioral science[35;36;37]. Table 6 summarizes the observations from this analysis.

Cluster C_0 (routine-oriented) individuals significantly decreased posting immediately after enrollment; however, their posting behaviors got closer to expected behaviors over time. This behavior change could be explained by their traits of high conscientiousness, which is known to be associated with self-monitoring[35]. The behavioral amendments over time is a form of *habituation* explained in behavioral science[38]. Linguistically, they decreased the use of first-person singular pronouns and increased the use of first-person plural pronouns and posting about public-facing events, which together could be considered to be reduced self-attentional focus and increased collective-identity-based language and increased posting about events attended as a part of a group[39].

Cluster C_1 (emotionally-stable and innovative) individuals significantly decreased posting in the immediate two weeks after enrollment, but their posting behaviors became closer to the expected behaviors subsequently. Their social media language showed an increase in **sociality** after enrollment[40]. As noted earlier, their use of content words increased, but their linguistic style remained similar. A possible explanation of their observed behaviors could be based on Middleton et al.'s observation that individuals with higher cognitive ability are less likely to show psychological reactance[41]. Again, the increased use of family-related keywords is known to be associated with lower self-monitoring[36]. They likely employ lower self-monitoring skills, are less bothered by the aspect of being "observed", and are comfortable to continue sharing their social and personal life on social media.

Cluster C_2 (withdrawn, disagreeable, and prone to stress and irritability) individuals decreased posting about social topics such as food and drinks, sports, and social events. This is also reflected in their psycholinguistic use of fewer personal and social words such as family, friends, and home. However, they increased their posting activity post-enrollment. Their higher volume of post-enrollment posting behavior could be associated with higher self-monitoring skills as per prior work[42]. They also received greater engagement in terms of likes and comments — plausibly a function of heightened information seeking on social media, which is known to be associated with higher neuroticism[43], as also in the case of C_2.

Cluster C_3 (positive, friendly, and well-balanced) individuals increased posting after enrollment. Extraversion is known to positively correlate with public self-consciousness[44] and self-monitoring[45]. Like C_2, greater posting behavior in C_3 could be manifested by high self-monitoring skills[42]. Also, high conscientiousness could indicate a desire to appear as "good" participants or self-present in a more desirable way[46]—likely reflected in their increased social media activities, increased positive affect, and decreased negative affect and swear words, as explained by the self-presentation literature[19;47]. Then, high agreeableness is known to associate with people's likelihood to seek acceptance and maintain social connections[43]. A similar phenomenon is observable as their posts elicited a greater number of likes and comments, compared to before enrollment.

Cluster C_4 (curious and adventurous) individuals did not significantly change posting behaviors immediately but significantly increased it over time. They also showed significant linguistic changes in the post-enrollment period. They increased posting about many personal and social aspects of life, despite a significant reduction in first-person singular pronouns and many function words. They lowered the use of negations and exclusives, suggesting lowered cognitive complexity in language — which could be associated with less personal content[48]. These changes may suggest that C_4 individuals are likely to self-regulate their social media use to present selective aspects of life without sharing too intimate content. Again, greater openness is

known to be associated with high psychological reactance[37], which could be manifested in detached sharing about personal and first-person singular content. Openness is known to be associated with greater resiliency and externally induced behavioral changes[49], however, its interplay with observer effect remains to be examined further.

3 Discussion

Theoretically, this work advances our knowledge about how participants varying in psychological traits could change social media use differently in prospective research design settings. These behavioral changes are explained by behavioral science and psychology theories, including self-monitoring[21], public self-consciousness[29], and psychological reactance[50]. Methodologically, this work contributes a computational and causal framework for modeling and assessing observer effect in prospective research studies in general, and those involving the monitoring of social media use in particular. Our work is motivated by person-centered approaches of clustering individuals on intrinsic traits and studying the behavior changes per cluster[51]. An advantage of person-centered approach is that it views each cluster as an integrated totality[51;52], and helps us draw within-person (or within-clusters, here) insights and interpretations, i.e., given an individual with a certain combination of traits, how are they likely to behave after an intervention. This work provides insights into how the observer effect occurs, how long it lasts, and how its occurrences vary across participants. We discuss this work's implications in recommending strategies to correct for biases arising as a consequence of the observer effect in social media studies.

3.1 Theoretical Implications

This study advances our knowledge in observer effect research. Typically, the observer effect has been hard to study because researchers could only access data generated after participant recruitment[22]. This has precluded researchers from measuring observer effect since it necessitates access to and comparison with a subject's otherwise normative and non-observed behavior, or the counterfactual how they would have behaved without the presence of an observer. In addition, there is no established gold-standard for measuring observer effect. This is the first study of measuring the observer effect in social media use. The longitudinal and historical nature of the social media data stream allowed access to extended periods of an individual's behavior on the platform, including pre-enrollment data. This enabled us to build behavioral models on typical or expected behaviors, which we leveraged in this work.

This study provides insights regarding the prevalence and degree of the observer effect in social media use, by intrinsic traits of participants. We draw a new understanding of how people varying with different combinations of these traits could behave when subjected to the observer effect. These findings inform research about correcting data, biases, and models when implementing practical and prospective data-driven assessments and interventions. In this regard, this study contributes to the recommendations made by Ruths and Pfeffer[12] in correcting biases of big data technologies. Specifically, this study helps us to gauge what to expect when social media is used to assess human behaviors in a prospective setting. For instance, this work informs us that composed and reasonable individuals (Cluster $\mathtt{C}_1$) are likely to decrease posting in the immediate period, but might show habituation, or return to expected behaviors over time, whereas those with high openness (Cluster $\mathtt{C}_4$) may not show any immediate change but increase posting over a period of time. These findings help us be more cognizant about which individuals might significantly deviate from their otherwise expected behaviors, and accordingly build personalized models that are robust to people's baseline traits and tendencies to be impacted by the observer effect.

Our findings can also help to generate hypotheses relating to observer effect in social media. For instance, in Section 2.2, we explained the findings through theories in psychology and behavioral science literature. These associations can be formulated as testable hypotheses in future research. Future research can also incorporate other intrinsic and social processes, such as self-censorship and privacy perceptions, which may also interact with social media behavioral change[53;54].

Due to the lack of direct means to measure success and construct validity of this research, we evaluated and situated the findings with the literature. While our work targeted to obtain passive and objective forms of assessment, it would also be interesting to examine self-reported assessments about the observer effect. Therefore, this work motivates us to design and conduct surveys and interviews to help us gauge complementary information about how the observer effect manifests in social media behavior.

In this study, observers were a group of researchers with whom the participants willingly shared their data based on a data-sharing protocol. These participants self-selected themselves in the study, for which they were compensated. However, the observer effect can manifest in other scenarios involving a variety of observers and data-sharing terms, such as clinicians observing the health of patients, employers observing the productivity of workers, or social media platforms monitoring user activities to control policy violations. Olteanu et al.[13] connected "online" observer effect with people's disclosure behaviors, in terms of how individuals are more likely to share unpopular, sensitive, and more personal opinions in private and anonymous spaces than public ones.[55;56]. Therefore, it remains important to understand the role of different factors in influencing observer effect in real-world situations.

3.2 Implications for Researchers and Practitioners

This research showed that individuals who deviated from their expected behaviors when subjected to real-time and prospective data collection settings — attributed as some form of observer effect. This effect needs to be accounted for to successfully instrument real-time applications that use social media to derive psychological assessments. The computational approaches adopted in this study can be used to measure observer effects in various contexts. Researchers can use such approaches to identify cases of observer effect-based deviations and build predictive models robust to such effects in a person-centric fashion. This study reflects that self-reported psychological traits can not only be used to stratify and cluster individuals, but also to explain their behavioral changes due to the observer effect. Similar approaches can be used to build person-centric models of correction for different groups of individuals. Relatedly, we noted in the Introduction how a majority of social media-based studies of human behaviors are retrospective and observational in nature. However, the major implication of these research studies is to conduct practical and real-time interventions. Our work bears an implication that it would be worth redoing and revisiting the retrospective analyses along with corrections for observer effect before significant efforts and resources are invested in making the interventions.

Besides highlighting the potential methodological biases, this study reinforces an ethical question about social media-based studies of human behaviors in general (both retrospective and prospective). It motivates us to critically reflect and rethink the implications surrounding individuals' autonomy in using social media technologies. People primarily use social media to share and connect with others. However, if external interventions interfere with their social media use or make them feel uncomfortable or surveilled – as revealed to be the same for at least some participants in this study – then the fundamental goals and expectations of using social media platforms can be compromised. Such an unintended consequence needs to be evaluated by researchers, practitioners, as well as the owners of social media platforms. To this end, this work encourages us to critique the trade-offs between the harms and benefits of using social media-based technologies for deriving psychological assessments, and also reinforces the necessity of consenting to individuals' social media data and their specific use.

3.3 Limitations and Future Directions

It is also important to note how our findings are an artifact of the domain and the participant pool. This study is conducted on a specific participant pool of information workers in the context of workplace settings. Such a factor may have an effect on the changes observed in the work-related language (in Table 4 and Table 5). In addition, our study is not devoid of biases due to self-selection[13], and our work adopts a person-centered approach to somewhat mitigate this challenge[57]. While our clustering-based approach helped us examine and understand how observer effect impacts different individuals' social media use, our study population does not include all possible combinations of intrinsic traits. Future experiments can explore more conclusive and generalizable evidence about the observer effect, and whether these are opportunities or challenges in other situations and contexts. We also note that even though it would have been interesting (and possibly more accurate) to include the demographic attributes of individuals in clustering, we excluded these attributes to primarily steer away from "demographic profiling" related interpretations and ethical concerns—demographic attribute-based stratified modeling has been associated with reinforcing and exacerbating stereotypes and existing societal biases[58;59]. In addition, given that our dataset is not representative of all demographic and marginalized groups, the non-demographic intrinsic traits are more robust for studying as well as for reproducibility and applicability of research.

4 Materials and Methods

4.1 Study and Data

The data for this study comes from the Tesserae project[60]. The Tesserae project was approved by the Institutional Review Boards (IRBs) at all the involved researcher institutions. The participants responded to initial survey questionnaires related to demographics, and trait-based measures relating to personality, affect, sleep, and executive functions. The participants were requested to remain in the study for either upto a year or through April 2019. The study enrollment was conducted from January 2018 through July 2018. Participants either received a series of staggered stipends totalling $750 or participated in a set of weekly lottery drawings (multiples of $250 drawings) depending on their employer restrictions.

Given the scale, duration, and nature of the project, the recruitment was challenging. Participants were recruited through both in-person as well as remote enrollment in early 2018. In-person recruitments included researchers in the team doing multiple rounds of corporate company site-visits to speak about and recruit participants. The remote enrollments were conducted via Zoom video conferencing. The participant onboardings included explaining the study protocol, consenting process, and clarifying participant questions through researcher proctoring sessions. This was followed by participants responding to the survey questions. The sensors were provided to the participants either via shipping or via in-hand delivery. More details about the participant recruitment, study protocol, and challenges and lessons learned about setting up the study can be found in Mattingly et al[60].

4.1.1 Social Media Data

The Tesserae project asked consented participants to authorize their Facebook data, *unless they opted out, or did not already use Facebook*. The enrollment briefing and consent process explicitly explained that the study participation did not necessitate them to use social media in a particular fashion, and they were expected to continue with their typical social media use. The participants authorized access to social media data through an Open Authentication (OAuth) based data collection infrastructure developed in Saha et al.[61]. OAuth protocol is an open standard for access delegation, commonly used as a way for internet users to log in and grant third party access to their information, without sharing passwords. The OAuth protocol provides a more privacy-preserving and convenient means of data collection at scale, over secured channels without the transfer of any personal credentials.

Given that Facebook is the most popular social media platform[62] and its longitudinal nature has enabled several of human behavior[4;63], it suits our problem setting of understanding observer effect in social media behavior. Out of the total 572 participants who provided access to Facebook data, 532 made at least one post on their Facebook timeline. Table 7 summarizes the Facebook dataset of Tesserae participants, and we find that there is roughly 82 months data per participant in the pre-enrollment period, and roughly 5 months data per participant in the post-enrollment period.

We filtered participants with at least 60 days of post-enrollment data, leading to 316 participants, whose data was used in this work. We note that we also conducted a sensitivity analysis by varying the threshold of the availability of minimum post-enrollment data (15 days, 30 days, 45 days) to see if the quantity of available data introduced any biases in our findings. We note that for each of the other thresholds, we repeated the experiments for 365, 344, and 335 participants, respectively; however, the findings did not significantly change compared to what we have for the 60 days threshold. In addition, a Cox Proportional-Hazards regression models[64]for all the examined measures (posts made and engagements received) confirmed no statistical significance with respect to the quantity of time of data in the post-enrollment period. This suggests that our findings are not sensitive to the minimum threshold of post-enrollment data considered.

4.2 Statistical Power

Power analysis in statistics estimates the minimum sample size for a study to make significant inferences on a given population[65]. Likewise, we used power analysis to examine if this study has sufficient sample size of participants to make reasonable inferences about the population. This study's participant pool belongs to information workers in the United States. According to U.S. Census Bureau, a rough estimate on the number of information workers in the U.S. is 4.6 million[66]. We calculated a sample size that is representative of this population with a 95% confidence interval and 5% margin of error, this comes out to be a sample size of 385. Given that the net social media sample size is 574 participants, out of which, usable data for studying observer effect is for 316 participants, this study assumes to have a reasonable sample of information workforce in the U.S.

Therefore, while we cannot claim absolute representativeness of the U.S. information workforce, we see a diversity of participants across the demographic and intrinsic traits of participants (Table 8). Statistical power analysis also revealed that we have a reasonable sample of the U.S. information workforce (more details in the Supplementary Information).

4.3 Analytic Approach

4.3.1 Clustering Participants on Intrinsic Traits

Typically, prediction tasks are modeled on the entire dataset of participants, also termed as variable-centric approaches, where a single model is built for the entire training data available. However, in contrast to many other datasets, social media data presents unique challenges, as it is sensitive to people's social media use and may significantly vary across individuals.

Although personalized approaches can help overcome the above challenge[67;68], it is hard to conduct personalized examinations on social media data because of sparsity issues, which compromises the statistical power. Therefore, drawing on prior work[31], we clustered individuals with self-reported intrinsic traits and then examined the outcomes per cluster. This approach is known to account for both between-individual homogeneity and within-individual heterogeneity in our behaviors[31]. As a robustness test of our findings, we also conducted variable-centered regression analyses with the intrinsic traits, described in Supplementary Information.

Given that demographic information are often privacy-intrusive and demographically discriminatory and non-inclusive[58], our clustering only accounted for self-reported intrinsic traits of cognitive ability (abstraction and vocabulary)[69], Big-5 personality traits (openness, conscientiousness, extraversion, agreeableness, and neuroticism)[70], and affect and wellbeing (positive affect, negative affect, anxiety, and sleep quality) measures[71;72;73]. Using these traits as features, we conducted k-means clustering on the individuals.

We employed the elbow heuristic to obtain the optimal number of clusters (k) in our approach[74]. Figure 4 shows the elbow plot of the mean sum of squared distances to the cluster centroids with respect to the number of clusters (k), roughly estimating an optimal number of clusters at k=5. This led us to cluster the initial 532 individuals in the dataset into five clusters (C_0 to C_4), containing 98, 121, 92, 146, and 83 members, respectively.

Evaluating Cluster Heterogeneity We evaluated if our clustering actually reduces the heterogeneity in data per cluster. Table 9 shows a comparison of the standard deviation of traits in the entire data against that per cluster and one-way ANOVA (F-statistic). We find that the standard deviation of each trait per cluster is lower than that in the entire data. One-way ANOVA essentially measures the ratio of between-group variance and within-group variance, and we find that the between-cluster variance in each of the traits is higher than the within-cluster variance. Therefore, we note cluster validity[51] in our approach.

Characterizing and Describing the Clusters of Individuals Figure 1 shows the average distribution of the traits and Table 1 summarizes the characteristics of the five clusters. We draw on the literature[46;75] to assign persona characterization for these clusters, which we describe below:

Cluster C_0 has individuals with high conscientiousness and sleep quality, and low openness and cognitive ability, suggesting the likelihood of them to be *routine-oriented*[46].

Cluster C_1 has individuals with high cognitive ability and low neuroticism, so they are more likely to be *emotionally stable and innovative*[46;69].

Cluster C_2 comprises individuals with high neuroticism, cognitive ability, negative affect, and anxiety, and low extraversion, agreeableness, conscientiousness, positive affect, and sleep quality. These characteristics suggest that they are likely to be more *withdrawn, disagreeable, and prone to stress and irritability*[46]. The ARC taxonomy describes this cluster of individuals as "overcontrolled", who would likely show obsessive-compulsive and avoidant shymptoms[75].

Cluster C_3 consists of individuals with high extraversion, agreeableness, conscientiousness, positive affect, and sleep quality, but low neuroticism, negative affect, and anxiety. They can be characterized to be *positive, friendly, and well-balanced*, i.e., resistant and less likely to experience stress, anxiety, and negative emotions. The ARC taxonomy describes their combination of personality traits as "resilient", and they likely show high psychological adjustments[75].

Cluster C_4 has individuals with high openness. People with high openness tend to be *curious and adventurous*—more open-minded and willing to embrace new things, fresh ideas, and novel experiences[76].

4.3.2 Conducting Placebo Tests

We needed to ensure that the effects observed in the study were an artifact of study enrollment and not due to other confounds or at chance. We conduct placebo tests drawing on permutation test approaches from prior work[77;78]. Within the pre-enrollment data, we permuted (randomize) on several *placebo* dates. We assigned 150 placebo dates, and repeated the above time series comparison around the placebo dates — for every placebo date, we computed the t-tests in the post-placebo date actual and predicted time series data. Then, over all the permutations of placebo dates, we computed the probability (p-value) of significant differences around placebo dates. A p-value lower than 0.05 would reject the null hypothesis that the significance is by chance, also revealing the credibility of any significant changes observed around the (real) enrollment date.

4.3.3 Measuring Behavioral Changes

Measures to Quantify Behavioral Changes We quantified the participants' post-enrollment behavioral changes on social media as the changes in quantity and verbosity of the posts. Additionally, social media use is also characterized by social networking and engagement received from others. Therefore, we also examined the changes in the quantity of likes and comments received on the participants' posts.

Posting Behavior. We examined social media posting behavior in two measures — 1) *quantity of posting*, i.e., the daily average number of posts, and 2) *verbosity of posting*, i.e., the daily average number of words.

Engagement Received. We examined the engagements received on social media posts, in terms of 1) *likes*, i.e., the daily average number of likes received, and 2) *comments*, i.e, the daily average number of comments received.

Modeling and Quantifying Behavioral Changes Drawing on interrupted time series and synthetic control based causal approaches[79;80], we computed the deviation in actual behavior from the expected behavior of the participants as modeled on their historical behavior. For each cluster, we built autoregressive models (ARIMA) to extrapolate post-enrollment expected behaviors of the participants. We built the models accounting for trends and seasonalities in the time series. We trained the models on the pre-enrollment data, using an 80:20 split (80% for training and 20% held-out for testing), and applied grid search to optimize for the best parameters of the time series prediction models. The models were evaluated on the 20% held-out data as symmetric mean absolute percentage error (SMAPE), which quantifies errors in the range of 0 to 100, where lower values indicate a better predictive model. We studied the differences between observed and expected behaviors in the short-term (two-weeks) and long-term (100-days) post-enrollment period and measured the statistical significance of the differences using paired t-tests and effect size (Cohen's d). We also computed the slope changes in the time series of social media use from pre- to post-enrollment periods, along with causal impact computed as per Brodersen et al.[33]. Higher values of the posterior probability of causal impacts (CI) would indicate a significant behavioral change after enrollment in the study.

4.3.4 Measuring Topical Changes

We conducted topic modeling in our dataset to examine how the prevalence and diversity of topics evolve following study enrollment. To extract topics automatically, we employed the widely adopted Latent Dirichlet Analysis (LDA) on the dataset[34].

Building Topic Models and Assigning Topic Labels To identify the optimal number of topics in our dataset, we drew recommendations from prior work[81;82] to vary the number of topics up to 25, and semi-automatically evaluated the quality of topic models, by combining the use of topical coherence scores as well as manual evaluations. Topical coherence score quantifies the degree of semantic similarity between high-scoring words within a topic[83]. Guided by both the highest coherence score, followed by manual evaluation, we used the topic modeling for the number of topics (n) as 10 in our study. Then, three members of the research team adopted interpretative annotation followed by thematic labeling into the interpretable labels of 1) *Travel and Locations*, 2) *Food and Drinks*, 3) *Holiday Plans*, 4) *News and Information*, 5) *Work-Life Balance*, 6) *Family Gathering*, 7) *Social and Sports*, 8) *Greetings and Celebration*, 9) *Friends and Family*, and 10) *Acivities and Interests*. Table S1 shows the 10 thematic categories and top occurring keywords per topic, along with an example paraphrased post from our dataset. The Supplementary sections provide more details about our topic modeling and manual evaluations.

4.3.5 Measuring Psycholinguistic Changes

Another dimension to understand people's expressiveness is through psycholinguistics[3;4]. We used the psychologically validated and widely adopted lexicon of Linguistic Inquiry and Word Count (LIWC)[84]. LIWC allows categorizing the pre- and post-enrollment social media data into psycholinguistic categories of: 1) *affect* (anger, anxiety, negative and positive affect, sadness, swear), 2) *cognition* (causation, inhibition, cognitive mechanics, discrepancies, negation, tentativeness), 3) *perception* (feel, hear, insight, see), 4) *interpersonal focus* (first person singular, second person plural, third person plural, indefinite pronoun), 5) *temporal references* (future tense, past tense, present tense), 6) *lexical density and awareness* (adverbs, verbs, article, exclusive, inclusive, preposition, quantifier), and 7) *personal and social concerns* (achievement, bio, body, death, health, sexual, home, money, religion, family, friends, humans, social).

Data Availability

As per the consenting process and IRB requirements, the data cannot be publicly shared. However, consented and de-identified data collected in the project can be made available upon request, subject to an appropriate data use agreement, if applicable.

Acknowledgements

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2017-17042800007. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. We further thank the contributions from Jordyn Seybolt and Nick Jaffe in annotating the topics. We thank the entire Tesserae team for their continued invovlement and valuable feedback. KS conducted a majority of this research at Georgia Tech.

Ethics Declaration

This research was approved by the Institution Review Board (IRB) at the research institutions working on the Tesserae project. In addition, utmost care was taken to secure the privacy of the individuals in the dataset. This paper used de-identified data for analyses and presents paraphrased quotes to minimize traceability.

Contributors

KS designed the research; KS, EK, and MDC conceptualized and developed the analytic techniques; KS and PG gathered and analyzed the data; KS interpreted the results; KS drafted the paper; and GM, EK, and MDC read, edited, and provided feedback on the paper.

References

1. Shelley Boulianne. Social media use and participation: A meta-analysis of current research. *Information, communication & society*, 18(5):524–538, 2015.

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

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

4. Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *ICWSM*, 2013.

5. Prerna Chikersal, Danielle Belgrave, Gavin Doherty, Angel Enrique, Jorge E Palacios, Derek Richards, and Anja Thieme. Understanding client support strategies to improve clinical outcomes in an online mental health intervention. In *Proc. CHI*, 2020.

6. Holly Korda and Zena Itani. Harnessing social media for health promotion and behavior change. *Health promotion practice*, 14(1):15–23, 2013.

7. Emma Grace, Parimala Raghavendra, Julie M McMillan, and Jessica Shipman Gunson. Exploring participation experiences of youth who use aac in social media settings: Impact of an e-mentoring intervention. *Augmentative and Alternative Communication*, 35(2):132–141, 2019.

8. Stevie Chancellor and Munmun De Choudhury. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):1–11, 2020.

9. Sharath Chandra Guntuku, H Andrew Schwartz, Adarsh Kashyap, Jessica S Gaulton, Daniel C Stokes, David A Asch, Lyle H Ungar, and Raina M Merchant. Variability in language used on social media prior to hospital visits. *Scientific Reports*, 10(1):1–9, 2020.

10. Danah Boyd and Kate Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679, 2012.

11. David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

12. Derek Ruths and Jürgen Pfeffer. Social media for large studies of behavior. *Science*, 346(6213):1063–1064, 2014.

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

14. Casey Fiesler and Nicholas Proferes. "participant" perceptions of twitter research ethics. *Social Media+ Society*, 2018.

15. Brooke Erin Duffy and Ngai Keung Chan. "you never really know who's looking": Imagined surveillance across social media platforms. *New Media & Society*, 21(1):119–138, 2019.

16. Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

17. Jukka Jouhki, Epp Lauk, Maija Penttinen, Niina Sormanen, and Turo Uskali. Facebook's emotional contagion experiment as a challenge to research ethics. *Media and Communication*, 4, 2016.

18. Ruth Faden, Nancy Kass, Danielle Whicher, Walter Stewart, and Sean Tunis. Ethics and informed consent for comparative effectiveness research with prospective electronic clinical data. *Medical Care*, pages S53–S57, 2013.

19. Erving Goffman. The presentation of self in everyday life. 1959.

20. Christina Steindl, Eva Jonas, Sandra Sittenthaler, Eva Traut-Mattausch, and Jeff Greenberg. Understanding psychological reactance. *Zeitschrift für Psychologie*, 2015.

21. Mark Snyder. Self-monitoring processes. In *Advances in experimental social psychology*, volume 12, pages 85–128. Elsevier, 1979.

22. Systematic review of the hawthorne effect: new concepts are needed to study research participation effects. *Journal of clinical epidemiology*, 67(3):267–277, 2014.

23. David Oswald, Fred Sherratt, and Simon Smith. Handling the hawthorne effect: The challenges surrounding a participant observer. *Review of social studies*, 1(1):53–73, 2014.

24. Luke F Chen, Mark W Vander Weg, David A Hofmann, and Heather Schacht Reisinger. The hawthorne effect in infection prevention and epidemiology. *Infection control & hospital epidemiology*, 36(12):1444–1450, 2015.

25. Ralph Catalano. *Health, behavior and the community: An ecological perspective*. Pergamon Press New York, 1979.

26. Alan E Kazdin. Evidence-based treatment research: Advances, limitations, and next steps. *American Psychologist*, 66(8): 685, 2011.

27. John D Holden. Hawthorne effects and research into professional practice. *Journal of evaluation in clinical practice*, 7(1): 65–70, 2001.

28. Augustine Brannigan and William Zwerman. The real "hawthorne effect", 2001.

29. Robert M Arkin and James A Shepperd. Strategic self-presentation: An overview. 1990.

30. Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

31. 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. Person-centered predictions of psychological constructs with social media contextualized by multimodal sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1–32, 2021.

32. Jasjeet S Sekhon. Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science*, 12: 487–508, 2009.

33. Kay H Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L Scott. Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics*, pages 247–274, 2015.

34. David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *JMLR*, 3(Jan):993–1022, 2003.

35. Jonathan A Shaffer, Andrew Li, and Jessica Bagger. A moderated mediation model of personality, self-monitoring, and ocb. *Human Performance*, 28(2):93–111, 2015.

36. Qiwei He, Cees AW Glas, Michal Kosinski, David J Stillwell, and Bernard P Veldkamp. Predicting self-monitoring skills using textual posts on facebook. *Computers in Human Behavior*, 33:69–78, 2014.

37. Eric A Seemann, Walter C Buboltz, Adrian Thomas, Barlow Soper, and Lamar Wilkinson. Normal personality variables and their relationship to psychological reactance. *Individual Differences Research*, 3(2), 2005.

38. Luke F Chen, Charlene Carriker, Russell Staheli, Pamela Isaacs, Brandon Elliott, Becky A Miller, Deverick J Anderson, Rebekah W Moehring, Sheila Vereen, Judie Bringhurst, et al. Observing and improving hand hygiene compliance implementation and refinement of an electronic-assisted direct-observer hand hygiene audit program. *Infection Control & Hospital Epidemiology*, 34(2):207–210, 2013.

39. Michael A Cohn, Matthias R Mehl, and James W Pennebaker. Linguistic markers of psychological change surrounding september 11, 2001. *Psychological science*, 15(10):687–693, 2004.

40. Sindhu Kiranmai Ernala, Kathan H Kashiparekh, Amir Bolous, Ali Asra, John M Kane, Michael L Birnbaum, and Munmun De Choudhury. A social media study on mental health status transitions surrounding psychiatric hospitalizations. *Proceedings of the ACM on Human-Computer Interaction*, (CSCW), 2021.

41. Jay Middleton, Walter Buboltz, and Barlow Sopon. The relationship between psychological reactance and emotional intelligence. *The Social Science Journal*, 52(4):542–549, 2015.

42. Jeffrey A Hall and Natalie Pennington. Self-monitoring, honesty, and cue use on facebook: The relationship with user extraversion and conscientiousness. *Computers in Human Behavior*, 29(4):1556–1564, 2013.

43. Gwendolyn Seidman. Self-presentation and belonging on facebook: How personality influences social media use and motivations. *Personality and individual differences*, 54(3):402–407, 2013.

44. Daniel R Stalder. Need for closure, the big five, and public self-consciousness. *The Journal of social psychology*, 147(1): 91–94, 2007.

45. Murray R Barrick, Laura Parks, and Michael K Mount. Self-monitoring as a moderator of the relationships between personality traits and performance. *Personnel psychology*, 58(3):745–767, 2005.

46. Murray R Barrick and Michael K Mount. The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, 44(1):1–26, 1991.

47. Bernie Hogan. The presentation of self in the age of social media: Distinguishing performances and exhibitions online. *Bulletin of Science, Technology & Society*, 30(6):377–386, 2010.

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

49. Claude H Miller, Michael Burgoon, Joseph R Grandpre, and Eusebio M Alvaro. Identifying principal risk factors for the initiation of adolescent smoking behaviors: The significance of psychological reactance. *Health communication*, 19(3): 241–252, 2006.

50. Jack W Brehm. A theory of psychological reactance. 1966.

51. Sang Eun Woo, Andrew T Jebb, Louis Tay, and Scott Parrigon. Putting the "person" in the center: Review and synthesis of person-centered approaches and methods in organizational science. *Organizational Research Methods*, 21(4):814–845, 2018.

52. Roseanne J Foti, Nicole J Thompson, and Sarah F Allgood. The pattern-oriented approach: A framework for the experience of work. *Industrial and Organizational Psychology*, 4(1):122–125, 2011.

53. Sauvik Das and Adam Kramer. Self-censorship on facebook. In *Proc. ICWSM*, 2013.

54. Alice E Marwick and Danah Boyd. I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society*, 13(1):114–133, 2011.

55. Michael S Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Gregory G Vargas. 4chan and/b: An analysis of anonymity and ephemerality in a large online community. In *International Conference on Weblogs and Social Media (ICWSM)*, 2011.

56. Sarita Yardi Schoenebeck. The secret life of online moms: Anonymity and disinhibition on youbemom. com. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

57. Katy Jordan. Validity, reliability, and the case for participant-centered research: Reflections on a multi-platform social media study. *International Journal of Human–Computer Interaction*, 34(10):913–921, 2018.

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

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

60. Stephen M Mattingly, Julie M Gregg, Pino Audia, Ayse Elvan Bayraktaroglu, Andrew T Campbell, Nitesh V Chawla, Vedant Das Swain, Munmun De Choudhury, Sidney K D'Mello, Anind K Dey, et al. The tesserae project: Large-scale, longitudinal, in situ, multimodal sensing of information workers. 2019.

61. Koustuv Saha et al. Social media as a passive sensor in longitudinal studies of human behavior and wellbeing. In *CHI Ext. Abstracts*. ACM, 2019.

62. Shannon Greenwood, Andrew Perrin, and Maeve Duggan. Demographics of social media users in 2016. pewinternet.org/2016/11/11/social-media-update-2016/, 2016. Accessed: 2017-02-12.

63. Robert E Wilson, Samuel D Gosling, and Lindsay T Graham. A review of facebook research in the social sciences. *Perspectives on psychological science*, 7(3):203–220, 2012.

64. David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34 (2):187–202, 1972.

65. KP Suresh and S Chandrashekara. Sample size estimation and power analysis for clinical research studies. *Journal of human reproductive sciences*, 5(1):7, 2012.

66. Press Release. Census bureau reports. https://www.census.gov/newsroom/press-releases/2016/cb16-139.html.

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

68. Reza Safdari, Elham Maserat, Hamid Asadzadeh Aghdaei, et al. 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.

69. Walter C Shipley. *Shipley-2: manual*. WPS, 2009.

70. Christopher J Soto and Oliver P John. 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): 117, 2017.

71. David Watson and Lee Anna Clark. The panas-x: Manual for the positive and negative affect schedule-expanded form. 1999.

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

73. Yuriko Doi, Masumi Minowa, Makoto Uchiyama, Masako Okawa, Keiko Kim, Kayo Shibui, and Yuichi Kamei. 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):165–172, 2000.

74. Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *ICDCS*, 2011.

75. Paul T Costa Jr, Jeffrey H Herbst, Robert R McCrae, Jack Samuels, and Daniel J Ozer. The replicability and utility of three personality types. *European Journal of Personality*, 16(S1):S73–S87, 2002.

76. Kendra Cherry and Paul G Mattiuzzi. *The Everything Psychology Book: Explore the human psyche and understand why we do the things we do*. Simon and Schuster, 2010.

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

78. Koustuv Saha, Manikanta D Reddy, Vedant Das Swain, Julie M Gregg, Ted Grover, Suwen Lin, Gonzalo J Martinez, Stephen M Mattingly, et al. 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)*, 2019.

79. David McDowall, Richard McCleary, and Bradley J Bartos. *Interrupted time series analysis*. Oxford University Press, 2019.

80. Sebastian Bauhoff. The effect of school district nutrition policies on dietary intake and overweight: a synthetic control approach. *Economics & Human Biology*, 12:45–55, 2014.

81. Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112, 2009.

82. Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Neural information processing systems*, volume 22, pages 288–296. Citeseer, 2009.

83. Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 889–892, 2013.

84. Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

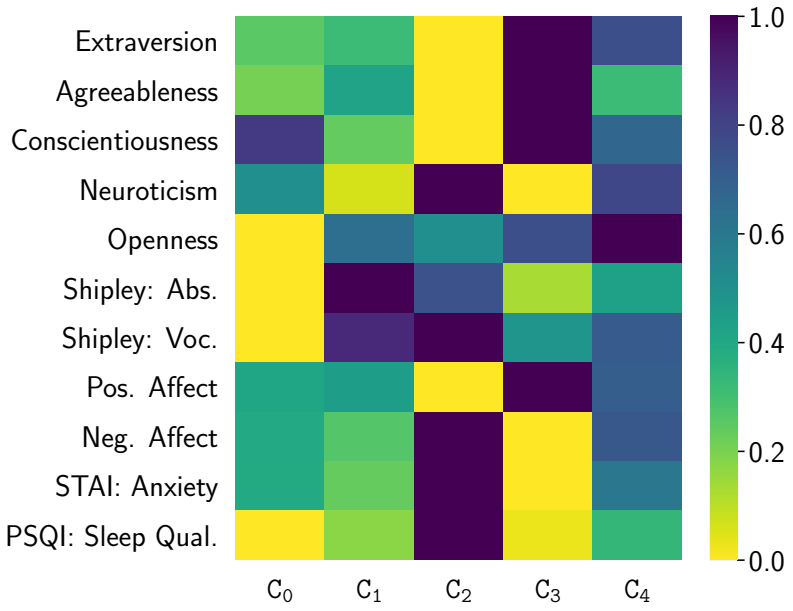


Figure 1. Distribution of traits across clusters of individuals.

Table 1. Summary of descriptions of clusters on psychological traits.

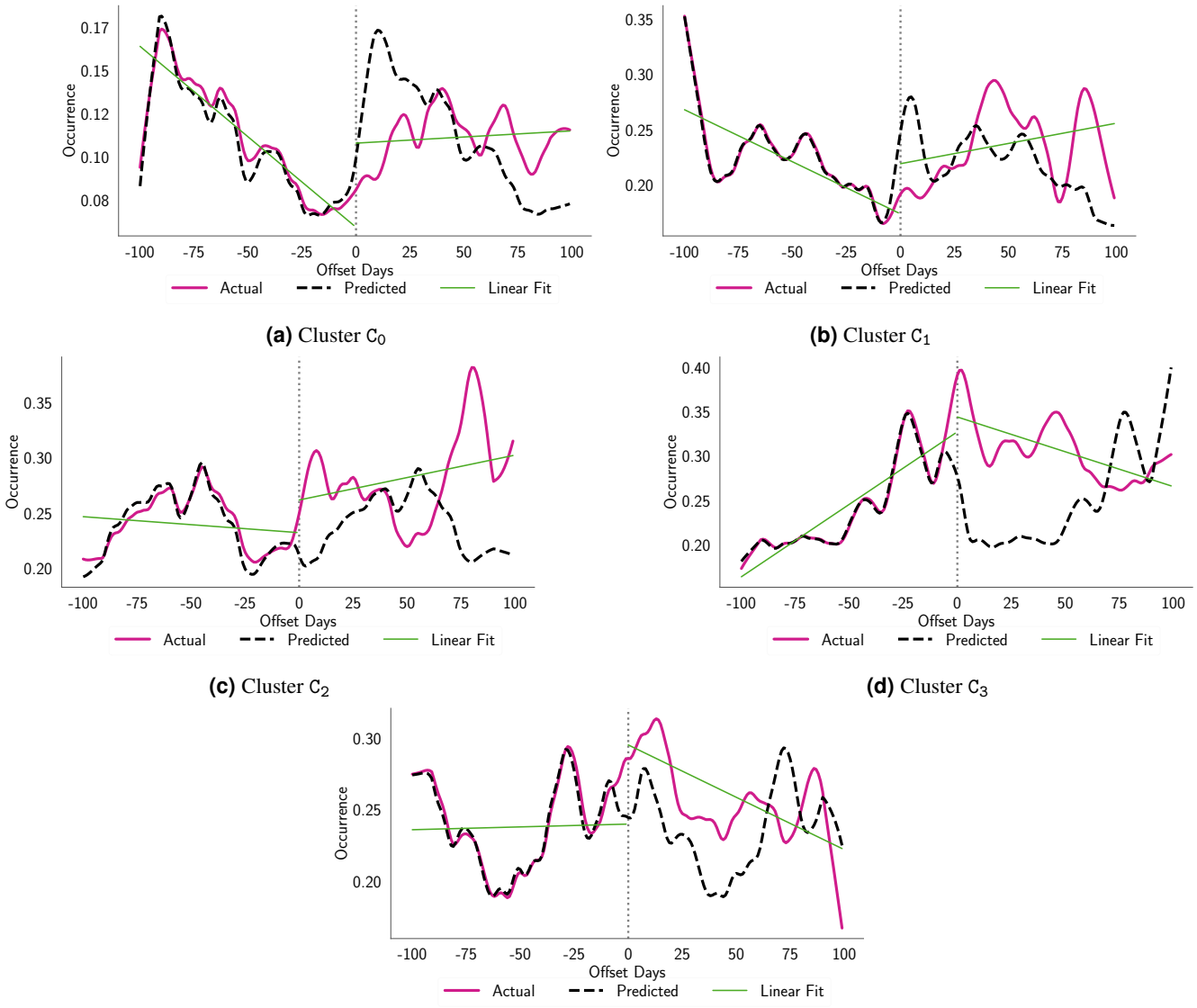
Cluster	N	Trait Overview	Persona characterization
Cluster c_0	60	High (Conscientiousness, Sleep Quality), Low (Openness, Cognitive Ability)	Routine-oriented [46]
Cluster c_1	66	High (Cognitive Ability), Low (Neuroticism)	Emotionally-stable and innovative [46;69]
Cluster c_2	44	Low (Extraversion, Agreeableness, Conscientiousness, PA, Sleep Quality), High (Neuroticism, Cognitive Ability, NA, Anxiety)	Withdrawn and prone to stress and irritability [46;75]
Cluster c_3	97	High (Extraversion, Agreeableness, Conscientiousness, PA, Sleep Quality), Low (Neuroticism, NA, Anxiety)	Positive, friendly, and well-balanced [75]
Cluster c_4	49	High Openness	Curious and adventurous [76]

Table 2. Summary of behavioral deviations in post-enrollment compared to expected (or predicted) behavior per cluster in terms of SMAPE, paired t-tests, and effect size (Cohen's d). Statistical significance reported as p-value, *<0.05, **<0.01, ***<0.001. Positive t or d indicates higher values in actual time series compared to the predicted time series. Significant values are shaded in blue to indicate an increase and red to indicate a decrease during the post-enrollment period.

Cluster	Model	100-days post-enrollment					2-weeks post-enrollment				
	SMAPE	Mean (Act.)	Mean (Exp.)	SMAPE	t-test	Cohen's d	Mean (Act.)	Mean (Exp.)	SMAPE	t-test	Cohen's d
Posting Behavior											
Average Daily Number of Posts											
Cluster c_0	11.09	0.11	0.11	24.45	-0.05	-0.01	0.09	0.16	30.73	-4.31 ***	-1.59
Cluster c_1	4.42	0.24	0.22	14.85	1.52	0.21	0.20	0.27	17.82	-3.68 ***	-1.35
Cluster c_2	5.78	0.28	0.24	17.45	3.49 ***	0.49	0.30	0.22	19.77	3.93 ***	1.44
Cluster c_3	4.20	0.31	0.25	18.00	5.76 ***	0.82	0.34	0.20	27.1	4.99 ***	1.84
Cluster c_4	5.85	0.26	0.24	16.25	2.02 *	0.29	0.31	0.28	17.04	1.07	0.39
Average Daily Number of Words											
Cluster c_0	22.69	0.67	0.59	42.85	1.10	0.16	0.58	0.75	54.22	-0.97	-0.36
Cluster c_1	11.24	1.68	1.86	24.99	-1.44	-0.2	1.66	1.76	19.46	-0.38	-0.14
Cluster c_2	11.31	2.14	1.80	24.5	2.60 *	0.37	2.24	1.72	23.28	1.46	0.54
Cluster c_3	6.40	1.85	1.60	17.86	3.03 ***	0.43	1.84	1.40	18.15	1.96 *	0.72
Cluster c_4	13.65	2.20	2.23	25.37	-0.18	-0.03	2.03	3.46	34.4	-3.72 ***	-1.37
Engagement Received											
Average Daily Number of Comments Received											
Cluster c_0	39.29	0.16	0.13	51.71	1.16	0.16	0.16	0.18	56.14	-0.41	-0.15
Cluster c_1	11.20	0.21	0.21	30.22	-0.35	-0.05	0.21	0.22	27.45	-0.41	-0.15
Cluster c_2	18.61	0.26	0.25	33.23	0.26	0.04	0.35	0.21	32.51	2.57 *	0.94
Cluster c_3	9.08	0.33	0.27	24.93	2.78 *	0.39	0.25	0.25	18.18	-0.18-	-0.07
Cluster c_4	25.63	0.30	0.29	31.38	0.46	0.07	0.24	0.33	26.55	-2.13 *	-0.78
Average Daily Number of Likes Received											
Cluster c_0	25.59	1.06	0.88	41.90	1.64	0.23	1.06	0.75	52.27	1.10	0.40
Cluster c_1	10.74	1.20	1.13	28.27	0.68	0.10	1.13	1.35	29.33	-0.97	-0.36
Cluster c_2	17.66	1.26	0.98	31.37	3.01 ***	0.43	1.49	0.76	33.87	3.10 ***	1.14
Cluster c_3	8.04	1.84	1.47	18.43	4.74 ***	0.67	1.47	1.35	17.45	0.72	0.26
Cluster c_4	13.74	1.97	1.73	28.2	1.70 *	0.24	1.57	1.94	32.34	-1.28	-0.47

Table 3. Summary of behavior changes in terms of causal impact estimation post-enrollment in the study, showing the slope in pre- and post- enrollment data, relative change in slope, Kolmogorov–Smirnov-test (KS-test), and posterior probability of causal impact (PP% CI)[33].

Cluster	Pre-enrollment	Post-enrollment	Rel. Change %	KS-test	PP% CI
Posting Behavior					
Average Daily Number of Posts					
Cluster c_0	-1.05×10^{-3}	7.13×10^{-5}	106.80	0.47***	65.83
Cluster c_1	-9.39×10^{-4}	3.65×10^{-4}	138.92	0.48***	96.20
Cluster c_2	-1.44×10^{-4}	4.05×10^{-4}	380.74	1.0***	99.60
Cluster c_3	1.63×10^{-3}	-7.85×10^{-4}	-148.01	0.63***	100.00
Cluster c_4	3.93×10^{-5}	-7.29×10^{-4}	-1954.83	0.76***	91.11
Average Daily Number of Words					
Cluster c_0	-3.60×10^{-3}	-4.19×10^{-4}	88.37	0.66***	82.82
Cluster c_1	6.04×10^{-5}	-7.03×10^{-3}	-11740.23	0.73***	77.52
Cluster c_2	2.20×10^{-4}	3.53×10^{-3}	1501.28	1.0***	98.40
Cluster c_3	3.23×10^{-3}	-2.30×10^{-3}	-171.11	0.91***	96.30
Cluster c_4	-1.35×10^{-2}	2.08×10^{-3}	115.45	0.46***	52.35
Engagement Received					
Average Daily Number of Comments Received					
Cluster c_0	-8.80×10^{-5}	-1.84×10^{-4}	-108.92	1.0***	58.94
Cluster c_1	-1.12×10^{-4}	1.72×10^{-4}	252.91	0.47***	53.35
Cluster c_2	3.42×10^{-4}	2.18×10^{-4}	-36.30	0.36***	99.10
Cluster c_3	7.75×10^{-4}	5.50×10^{-5}	-92.90	1.0***	65.23
Cluster c_4	-2.22×10^{-5}	2.32×10^{-4}	1144.37	0.89***	91.71
Average Daily Number of Likes Received					
Cluster c_0	5.29×10^{-4}	-2.06×10^{-3}	-489.06	1.0***	75.62
Cluster c_1	5.86×10^{-4}	-1.08×10^{-3}	-284.34	0.84***	99.10
Cluster c_2	3.67×10^{-3}	4.36×10^{-3}	18.96	0.75***	99.80
Cluster c_3	2.78×10^{-3}	2.41×10^{-3}	-13.33	1.0***	89.31
Cluster c_4	3.49×10^{-3}	1.98×10^{-3}	-43.22	0.89***	86.91



(a) Cluster C_0

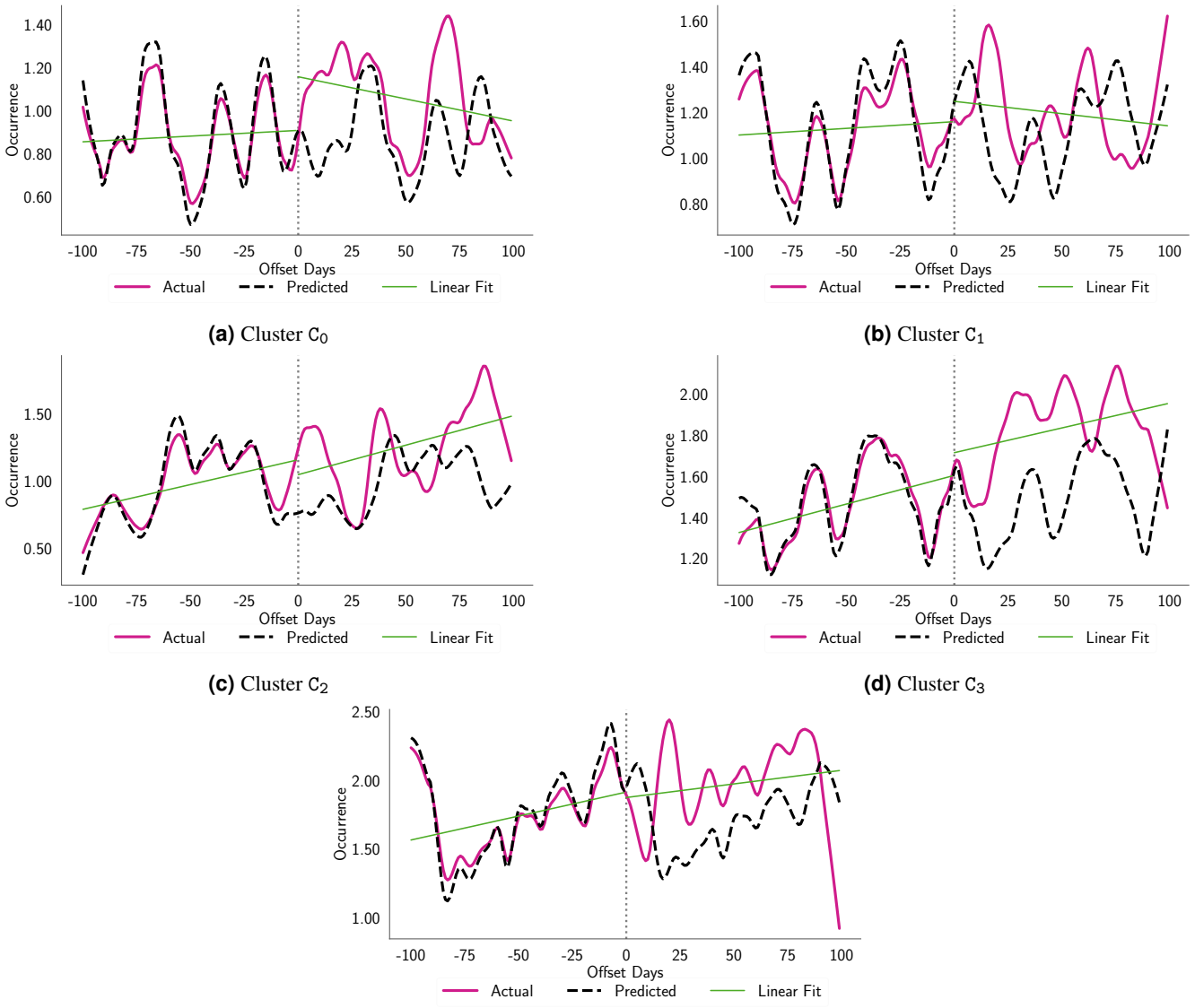
(b) Cluster C_1

(c) Cluster C_2

(d) Cluster C_3

(e) Cluster C_4

Figure 2. Evolution of the daily average number of posts per cluster in 100-days pre- and post- enrollment period. The dotted line in the center of each plot represents the date of enrollment (day 0).



(a) Cluster C_0

(b) Cluster C_1

(c) Cluster C_2

(d) Cluster C_3

(e) Cluster C_4

Figure 3. Evolution of the daily average number of likes per cluster in 100-days pre- and post- enrollment period. The dotted line in the center of each plot represents the date of enrollment (day 0).

Table 4. Changes in topical prevalence post-enrollment in the study. Statistical significance is computed as per independent-sample t-tests (* p<0.05, ** p<0.01, *** p<0.001). Significant values are shaded in blue for increased sharing, i.e., higher average value in post-enrollment, and red for decreased sharing, i.e., lower average value in post-enrollment period.

Topic	% Change in Cluster				
	C_0	C_1	C_2	C_3	C_4
Travel & Locations	28.38 ***	-0.98	-7.69	25.14 *	-1.94
Food & Drinks	37.16 ***	2.28	-13.85 **	3.39 *	14.20 **
Holiday Plans	18.22 *	18.65 *	-7.22	-12.10 *	-10.21 *
News & Information	33.89 **	-14.25 ***	-6.19	-19.29 ***	-17.36 *
Work-Life Balance	-0.05	1.28	-8.93	-8.30 *	0.88
Family Gathering	56.72 ***	11.99 *	-7.43	3.41	36.54 ***
Social & Sports	-29.13 **	12.21	-4.54 *	66.77 ***	-14.62
Greetings & Celebrations	-11.58 ***	23.62 ***	-7.64	-28.64 ***	18.51
Friends & Family	-1.37	-5.98	-10.48	-12.79 ***	-9.33
Activities & Interests	-0.38	6.10	-21.42	-16.39	-30.58 **

Table 5. Independent-sample *t*-tests in pre- and post- enrollment psycholinguistic (LIWC) use per cluster (* $p<0.05$, ** $p<0.01$, *** $p<0.001$). Significant values are shaded in blue for positive changes, i.e., higher average occurrence in post-enrollment, and red for negative changes, i.e., lower average occurrence in post-enrollment period.

LIWC	*t*-test				
	C_0	C_1	C_2	C_3	C_4
Affect					
Anger	2.01 *	-1.07	-2.02 *	-1.02	4.00 ***
Anxiety	1.41	0.03	-2.27 *	-1.957	1.93
Neg. Affect	0.83	-0.93	-2.60 **	-2.83 **	2.09
Pos. Affect	0.08	1.18	-4.49 ***	1.30 *	-2.06
Sadness	1.427	-0.42	1.52	-1.46	-0.61
Swear	1.134	0.60	-0.12	-3.06 **	7.53 ***
Cognition					
Causation	0.234	0.87	-2.69 **	-1.97	0.20
Certainty	4.08 ***	1.91	-2.12	-1.11	0.28
Cog. Mech.	1.32	0.86	-3.80 ***	-0.80	-0.93
Inhibition	-1.13	-1.37	-3.53 ***	-0.02	0.60
Discrepancies	-1.20	-1.61	1.08	-0.05	-0.55
Tentativeness	0.43	-1.17	1.79	-1.83	1.23
Feel	-2.31 *	0.87	-1.66	-3.12 **	-2.51
Hear	5.48 ***	0.50	2.39 *	1.27	1.41
Insight	-1.23	-0.141	-0.32	-2.39	0.90
Percept	-0.07	0.35	-4.74 ***	-1.23	-1.50
See	-2.31 *	-0.80	-4.77 ***	-1.41	-0.80
Interpersonal Focus					
1st P. Sing.	-7.29 ***	-1.00	-5.78 ***	-2.35 *	-4.17 ***
1st P. Plu.	2.25 *	0.47	-2.34 *	1.86	1.31
2nd P.	-1.43	-3.32 ***	5.71 ***	1.16	-0.70
3rd P.	-0.12	-0.63	-0.26	4.61 ***	-0.03
Indef. Pron.	-3.29 **	-1.33	-1.30	-3.43 ***	0.92
Fut. Tense	0.32	-0.65	2.32	-0.69	-0.36
Past Tense	1.85	0.28	-1.99 *	-0.158	2.61 **
Prs. Tense	-5.54 ***	0.19	-3.15 **	-6.49 ***	-1.90

LIWC	*t*-test				
	C_0	C_1	C_2	C_3	C_4
Lexical Density and Awareness					
Adverb	-3.00 **	-0.44	0.61	-3.36 ***	-2.46 *
Article	0.10	1.94	-3.60 ***	0.27	-1.34
Verb	-4.78 ***	0.47	-2.77 **	-5.53 ***	-1.36
Aux. Verb	-4.61 ***	0.40	0.10	-7.00 ***	-1.30
Conjun.	1.82	2.43 *	3.01 **	0.88	-0.15
Exclusive	1.05	-1.56	0.80	-1.92	-0.33
Inclusive	2.17 *	2.99 **	-3.47 ***	3.32 ***	-1.53
Negation	-1.38	-1.09	-0.90	-4.73 ***	-2.68 **
Preposition	1.47	-1.01	-3.27 **	-3.11 **	-2.28 *
Quantifier	-2.34 *	0.96	0.71	-0.06	0.50
Relative	-2.20 *	-1.16	-3.65 ***	-1.57	-2.98 **
Personal and Social Concerns					
Achvmt.	3.28 **	-0.91	-2.61**	0.14	-1.08
Bio	1.57	2.57 *	0.09	-0.20	-2.77 **
Body	-1.72	0.74	1.03	-1.34	-1.73
Death	0.43	1.99 *	-0.931	-0.162	-0.45
Family	1.14	2.64 **	-2.06 *	3.66 ***	0.29
Friends	-2.08	0.35 *	-1.46 *	-2.01	-1.17
Health	0.52	0.47	-0.34	-0.11	-0.81
Home	3.14 **	2.77 **	-2.19	1.39	0.08
Humans	-2.94 **	-1.67	0.51	2.85 **	-0.62
Money	-1.81	-1.06	-1.05	1.01	-1.24
Religion	2.29 *	2.48 *	-1.06	-0.87	-0.14
Sexual	1.62	-0.27	1.07	-0.62	0.56
Social	-1.02	-0.63	-1.542	2.79 **	0.30
Work	0.29	-0.58	-4.57 ***	-2.96 **	-1.74

Table 6. Summary of Findings.

Cluster	Traits	Behavior	Topics	Psycholinguistics	Notes / Descriptor
c_0	High (Conscientiousness, Sleep Quality), Low (Openness, Cognitive Ability)	Posting significantly reduces in the initial few days, then back to expected behaviors (2a)	Increased sharing about public-facing information (4)	Increased (anger, achievement, home, religion), Decreased (feel, first person singular, present tense, function words, friends, humans) (5)	High conscientiousness is associated with self-monitoring. Habituation in posting behavior. Decreased self-attentional focus.
c_1	High (Cognitive Ability), Low (Neuroticism)	Posting significantly decreased in the first two weeks, then closer to expected behaviors (2b)	Increased sharing about family gathering, social, and online greeting related activities (4)	Increased (social words), Decreased (2nd person) (5)	These participants are trait-wise more reasonable and composed. They show high sociality post-enrollment. Low psychological reactance and low self-monitoring skills; less bothered about being "observed".
c_2	Low (Extraversion, Agreeableness, Conscientiousness, PA, Sleep Quality), High (Neuroticism, Cognitive Ability, NA, Anxiety)	Posting significantly increased throughout. Greater engagement received. (2c)	Decreased sharing about food and social topics (4)	Increased (hear, future tense), Decreased (affective, cognitive, perceptive, 1st person pronouns, function words, social words) (5)	Trait-wise, they may be more withdrawn, and prone to stress and irritability. High self-monitoring skills, and heightened information seeking (associated with high neuroticism).
c_3	High (Extraversion, Agreeableness, Conscientiousness, PA, Sleep Quality), Low (Neuroticism, NA, Anxiety)	Posting significantly increases throughout. Greater engagement received. (2d)	Decreased sharing about personal events (4)	Increased (social words, third person pronouns), Decreased (affective, cognitive, perceptive, first person pronouns, function words) (5)	Desire to self-present in a more desirable way. Likelihood to seek acceptance and maintain social connections.
c_4	High Openness	No immediate significant difference in posting frequency, but posting significantly increases throughout. More likes received. (2e)	Decreased sharing about news and holiday plans. Increased sharing about food/family gathering (4)	Increased (anger, NA, swear, past tense), Decreased (PA, negation, feel, 1st person singular, function words) (5)	Self-regulation. Less personal-content. High psychological reactance, manifested in detached sharing about personal content.

Table 7. Summary of pre- and post- enrollment Facebook datasets.

	Before Enrollment		After Enrollment	
Type	**Range**	**Mean**	**Range**	**Mean**
Posts	26-4,472	865	8-964	101
Comments	34-10,228	1,593	5-1,104	175
Likes	62-52,139	6,536	15-4,540	940
Duration (months)	0-160.27	82.52	0-12.87	4.59

Table 8. Summary of demographics and individual differences of 316 participants whose data we study for observer effect.

Covariates	Value Type	Values / Distribution	
Demographic Characteristics			
Gender	Categorical	Male | Female	
Age	Continuous	Range (21:63), Mean = 36.36, Std. = 10.28	
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)			
Fluid (Abstraction)	Continuous	Range (5:24), Mean = 16.53, Std. = 3.32	
Crystallized (Vocabulary)	Continuous	Range (18:40), Mean = 33.82, Std. = 3.63	
Personality Trait (BFI)			
Extraversion	Continuous	Range (1.7:5.0), Mean = 3.43, Std. = 0.71	
Agreeableness	Continuous	Range (2.3:5.0), Mean = 3.97, Std. = 0.57	
Conscientiousness	Continuous	Range (1.9:5.0), Mean = 3.90, Std. = 0.65	
Neuroticism	Continuous	Range (1.0:4.6), Mean = 2.52, Std. = 0.82	
Openness	Continuous	Range (2.2:5.0), Mean = 3.88, Std. = 0.59	
Affect and Wellbeing			
Pos. Affect	Continuous	Range (13.0:49.0), Mean = 33.91, Std. = 5.84	
Neg. Affect	Continuous	Range (10.0:40.0), Mean = 17.14, Std. = 5.24	
Anxiety	Continuous	Range (20.0:67.0), Mean = 39.01, Std. = 10.00	
Sleep Quality	Continuous	Range (1.0:16.0), Mean = 7.14, Std. = 2.75	

Table 9. Comparison of standard deviation in traits in the entire data and that per cluster, one-way ANOVA (F-statistic), statistical significance reported as p-value, *<0.05, **<0.01, ***<0.001.

Trait	All	$\mathbf{C_0}$	$\mathbf{C_0}$	$\mathbf{C_0}$	$\mathbf{C_0}$	$\mathbf{C_0}$	F-stat.
Extraversion	0.71	0.68	0.56	0.61	0.50	0.62	22.23***
Agreeableness	0.57	0.42	0.52	0.52	0.41	0.55	15.94***
Conscientiousness	0.65	0.53	0.48	0.52	0.39	0.44	46.67***
Neuroticism	0.82	0.58	0.39	0.46	0.48	0.58	78.74***
Openness	0.59	0.48	0.42	0.43	0.52	0.35	25.85***
Shipley: Abs.	3.32	3.02	2.84	3.32	2.83	3.15	2.95**
Shipley: Voc.	3.63	3.62	2.85	3.44	3.50	3.60	1.70*
Pos. Affect	5.84	4.51	4.78	4.88	3.91	5.11	25.23***
Neg. Affect	5.24	3.35	3.70	4.32	2.55	4.63	23.09***
STAI: Anxiety	10.00	5.29	5.08	7.01	5.32	7.64	79.28***
PSQI: Sleep Qual.	2.75	2.29	2.58	2.92	2.59	2.08	9.12***

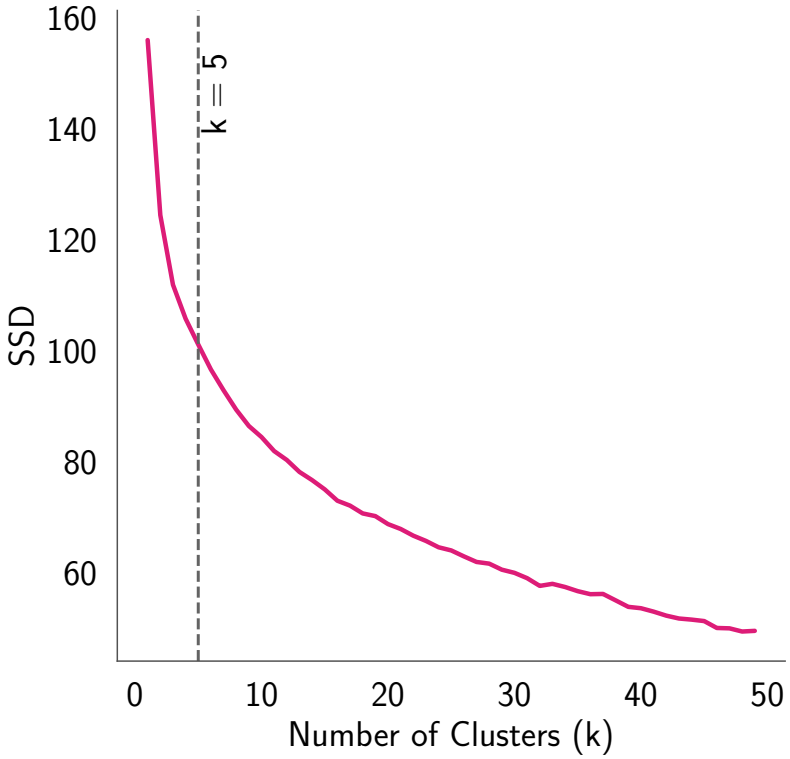


Figure 4. Elbow plot to estimate the optimal number of clusters by varying number of clusters (k) and mean sum of squared distances to the cluster centroids (SSE).

Supplementary Information Appendix

This file includes supporting information (SI) for the manuscript "**Observer Effect in Social Media Behavior**."

This PDF file includes:

Fig. S1-S5
Tables S1-S4
SI Appendix References

Preliminary Analyses

We conducted some feasibility and preliminary tests on the data for our study.

S4 Quantity of Posting

Posting behavior is a prominent social media behavior that has revealed significant signals of human behavior in prior work[? ? ? ?]. I measure the average posting behavior of the participants over time and around their enrollment in the study. Figure S1 shows the daily average posting behavior of the participants relative to the day of enrollment, where day=0 corresponds to the enrollment day for the participants. We notice an apparent bump in the average number of posts per day post-enrollment in the study.

S5 Expressive Behavior

We examined the changes in the expressive behavior of the participants. For this, we used the psycholinguistic lexicon LIWC[84] to obtain the psycholinguistic changes in the participants' post-following enrollment in the study. Figure S2 reports the effect sizes comparing pre- and post- enrollment normalized use of psycholinguistic categories across the participants. A positive effect size indicates greater use of the category post-enrollment, whereas a negative effect size indicates lower use in the post-enrollment period. Effect size (Cohen's d) is considered to be a significant difference if its magnitude is greater than 0.15.

We find that at an aggregated level, multiple psycholinguistic categories show significant changes. For example, considering pronoun use, first-person pronoun use decreases, which might indicate a decreased sharing of intimate content and decreased self-attentional focus[39]. In contrast, the use of first-person plural, second, and third-person pronouns increases. We also find a decrease in the use of cognition-related words (such as cognitive mechanics, discrepancies, inhibition, negation, etc.). We also find a significant decrease in affective categories of anger, sadness, and swear.

The above preliminary analyses indicate certain changes people's behavioral and expressive social media use following enrollment in the study at an aggregated level. This motivates us to examine the changes in a much more rigorous and robust fashion. Given that not all individuals are the same, this study borrows from person-centric approaches to examine the changes in cohorts (clusters) of similar individuals on psychological constructs[?].

Methodological Details

S6 Clustering Individuals on Intrinsic Traits

We adopted a k-means clustering approach to cluster individuals on intrinsic traits as collected via ground-truth surveys (personality traits, cognitive ability, affect, anxiety, and wellbeing). We employed the elbow-heuristic to obtain the optimal number of clusters (k) in our approach[74]. Fig. 4 shows the elbow plot of mean sum of squared distances to the cluster centroids with respect to the number of clusters (k), roughly estimating an optimal number of clusters at k=5. This led us to clustering the initial 532 individuals in the dataset into five clusters ($\mathtt{C}_0$ to $\mathtt{C}_4$), containing 98, 121, 92, 146, and 83 members respectively. We also evaluated if our clustering approach actually reduces the heterogeneity in data per cluster. Table 9 shows a comparison of the standard deviation of traits in the entire data against that per cluster, and one-way ANOVA (F-statistic). We find that the standard deviation of each trait per cluster is lower than that in the entire data. One-way ANOVA essentially measures the ratio of between-group variance and within-group variance, and we find that the between-cluster variance in each of the traits is higher than within-cluster variance. Therefore, we note cluster validity[51] in our approach.

S7 Building Topic Models

Finding Optimal Number of Topics Figure S3 plots the coherence scores on varying the number of topics from 2 to 26, suggesting that the highest coherence is achieved at around the number of topics (n) as 10. In addition, the first author and two collaborators in the research team manually evaluated the topical distribution for n=8, n=10, and n=12. We found the topical distributions at n=8 and n=12 to be less semantically coherent, with a substantial increase in noisy keywords. Therefore, as guided by both coherence scores and manual examination, we used topic modeling for n=10 topics for our study.

Interpreting Topics After building the topic models, we assigned interpretable labels to topics and keywords. For this purpose, three members of the research team designed an interpretive annotation to identify coherent themes in the keywords per topics. The topics were first inductively and independently coded with implied themes. Then the codes were compared and agreed upon to assign final thematic labels per topic. The thematic category of a topic was implied from the within-topic coherence and between-topic separation of keywords. These themes are 1) *Travel and Locations*, 2) *Food and Drinks*, 3) *Holiday Plans*, 4) *News and Information*, 5) *Work-Life Balance*, 6) *Family Gathering*, 7) *Social and Sports*, 8) *Greetings and Celebration*, 9) *Friends and Family*, and 10) *Acivities and Interests*. Table S1 shows the 10 thematic categories and top occurring keywords per topic, along with example paraphrased post from our dataset.

S8 Robustness of Findings

In this work, we adopted a person-centered approach by first clustering individuals, followed by observing and interpreting the findings on their behavioral changes by situating them in the literature. However, we also wanted to evaluate the robustness of our observations with respect to our chosen study design. As an additional analysis, we built linear regression models of intrinsic traits and observer effect deviation (SMAPE) for each of the metric (quantity of posts and words, and quantity of likes and comments received). Table S2 summarizes the coefficients of the regression models, and we describe our observations here. We find significance in several of the traits with the participants' deviations in post-enrollment social media use. All the regression models show high goodness-of-fit (adjusted R^2) with statistical significance.

Among personality traits, we find that extraversion shows a negative coefficient with the immediate likes and comments received, i.e., extroverted individuals were less likely to show deviation in the engagement received immediately after study enrollment. Agreeableness only shows for a positive coefficient with quantity of posts deviation (2 weeks), i.e., more agreeable individuals were more likely to deviate the most immediately post-enrollment—this aligns with behaviors of $\mathtt{C}_3$ (which included people with high agreeableness). Conscientiousness shows high coefficient along with significance for all the metrics. We also see that its coefficient is higher for the 2-weeks period as compared to the corresponding coefficient in the 100-day period, indicating that more conscientious individuals are likely to show greater deviation immediately post-enrollment, but the deviation also sustains in long-term post-enrollment. $\mathtt{C}_0$ and $\mathtt{C}_3$ consisted individuals with high conscientiousness, and we found that $\mathtt{C}_3$ showed high deviation in all the metrics, and $\mathtt{C}_0$ showed deviation in the engagement received immediately after enrollment. Neuroticism is another trait that shows positive coefficient across all the metrics, indicating that individuals with high neuroticism are likely to show high deviation in social media use—consistent with what we saw for $\mathtt{C}_2$. In contrast to the other personality traits, openness shows a negative coefficient across all the metrics, suggesting that high-openness individuals are less likely to show deviation in social media use—which we also observed for $\mathtt{C}_4$ who showed small deviations in the behaviors. Next, cognitive ability generally shows a positive coefficient across the metrics—these individuals were likely to be in $\mathtt{C}_1$ and $\mathtt{C}_2$ clusters. We see positive affect showing a positive coefficient across the metrics—$\mathtt{C}_3$, which was characterized as the happy and expressive individuals showed high deviation across the metrics. Trait anxiety also shows a positive coefficient across the metrics, and this trait likely correlates with neuroticism and occurred for $\mathtt{C}_2$ individuals, showing deviations in post-enrollment social media use. Finally, for sleep quality, PSQI's directionality is interpreted as higher PSQI values indicating lower sleep quality and lower PSQI values indicate higher sleep quality. Therefore, negative coefficients can be interpreted as greater sleep quality associates with greater deviations (e.g., in $\mathtt{C}_3$), who showed high deviation in the measures.

The regression examinations not only provide convergent findings as our cluster-based approach, but also provide additional observations into how the intrinsic traits associate with people's deviations in social media use post-enrollment into our study. We note that these are interesting insights into how different traits relate to the observer effect impacts. These observations bear theoretical explanations into how different personality traits are associated with self-monitoring[21;35], psychological reactance[37;41], self-presentation[19], and resilience to behavioral changes[49]. However, the variable-centered regression analyses, alone, cannot explain how different combinations of traits interact, and how different individuals with these combinations of traits behave when subjected to the observer effect. Therefore, we believe that person-centered or variable-centered approaches are not necessarily substitutes for each other, but rather, provide complementary lenses for inferring the findings. The interplays of the traits need to be further examined and evaluated in different populations and study samples as well.

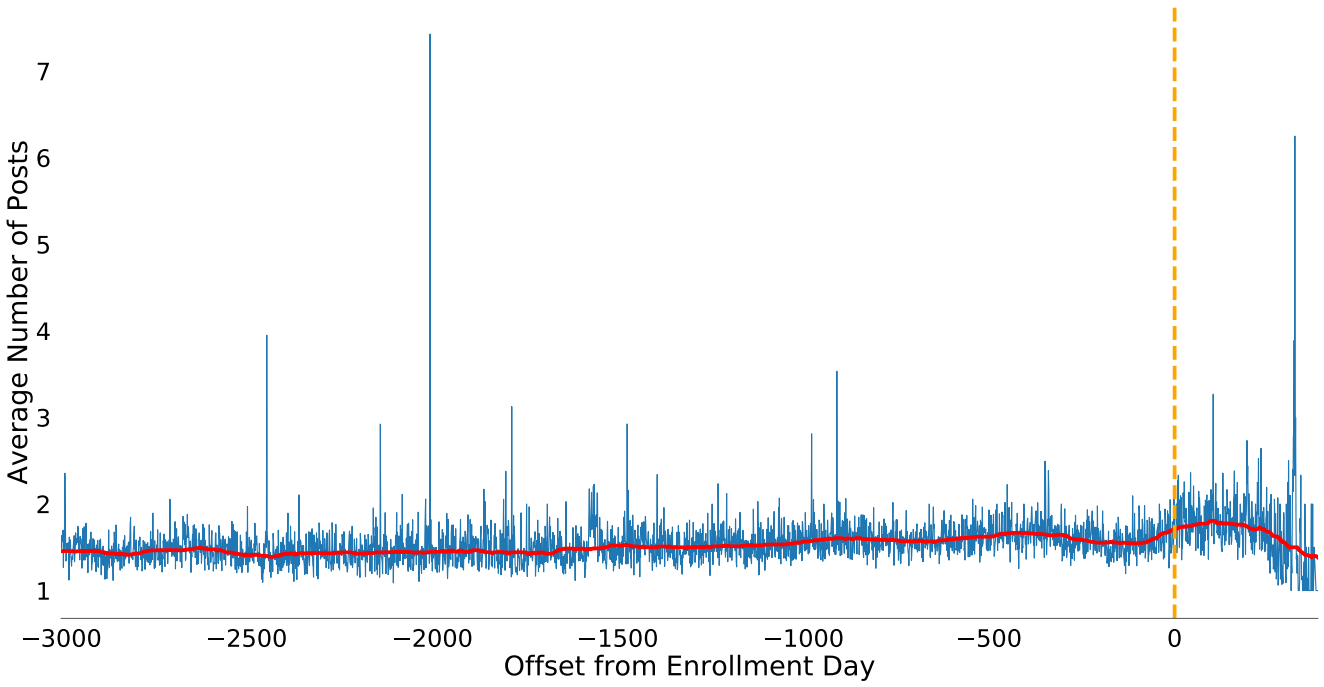


Figure S1. Average number of posts per day across all participants on relative offset from their day of enrollment. Day 0 indicates the day of enrollment.

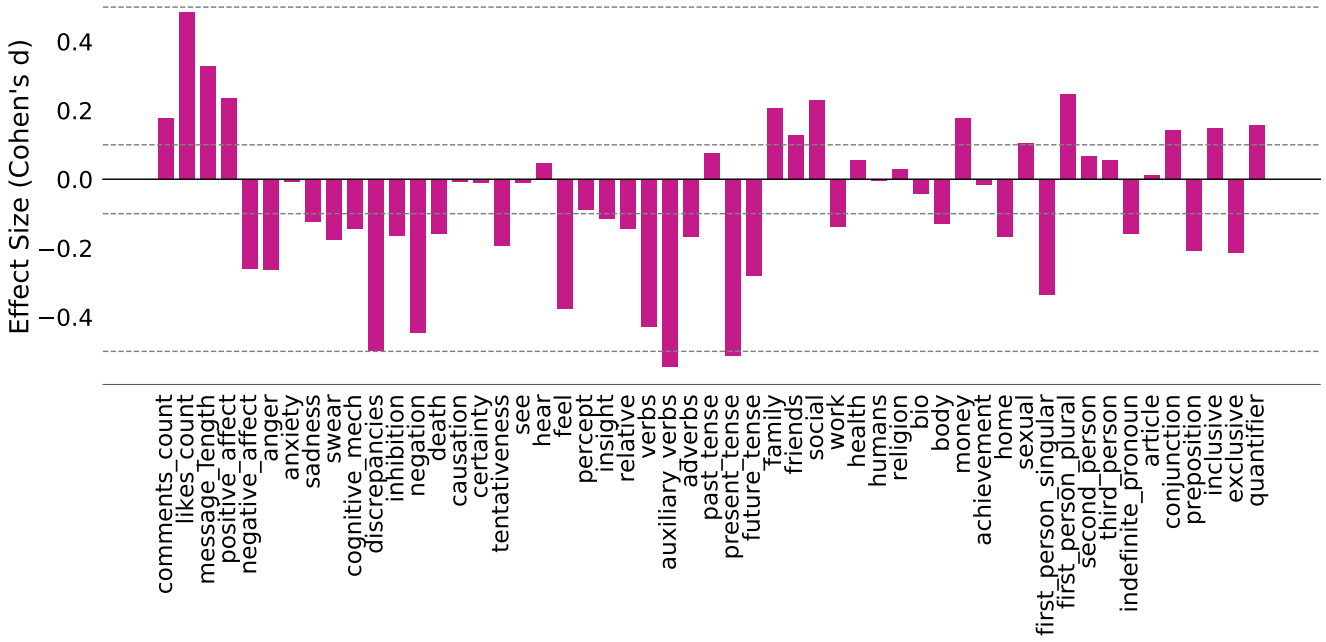


Figure S2. Effect size (cohen's *d*) comparing before and after enrollment datasets of users across psycholinguistic attributes. A positive cohen's *d* indicates that post- enrollment data Cohen's *d* magnitude smaller than 0.20 is considered to be small difference.

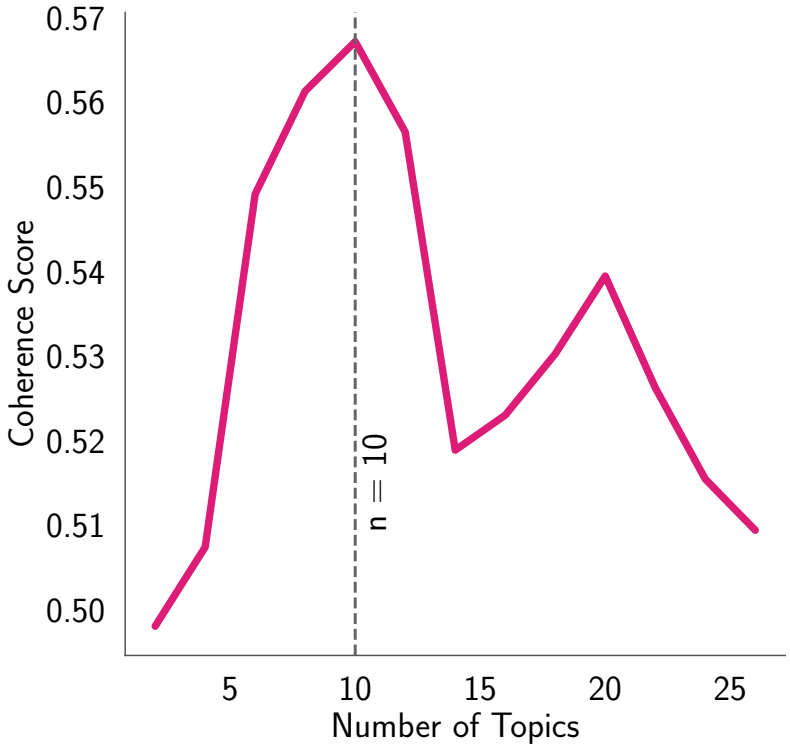


Figure S3. Topical coherence scores on LDA topic modeling with varying number of topics.

Table S1. Thematic categories of topics identified in our dataset.

Theme	Topic Words	Example post
Travel & Locations	country, green, baby, miss, right, chicago, sad, need, let, denver, mean, airport, hello, way, win, begin, yum, national, cubs, joanie	*Smiles all around after a good ATD conference together in Denver.*
Food & Drinks	lol, new, ready, room, sweet, boy, getting, waiting, finally, time, chicken, need, delicious, chicken, got, cheese, food, beer, gotta, yeah, guess	*Chicken on the grill, beef roast on the cutting board, regular and sweet potatoes in the oven. Guess who's not cooking tomorrow!*
Holiday Plans	christmas, school, vote, today, true, high, trip, look, season, awesome, johnson, merry, news, summer, party, check, raise, mom, family	*Morning hike, trip to the beach, and relaxing at our rental!*
News & Information	like, people, time, things, trump, think, watch, know, looks, right, got, thing, want, need, good, going, bad, stop, run, better, org	*Climate models want to change the way we live ... should we listen? It's a short video, watch it.*
Work-Life Balance	home, work, day, got, yes, new, today, time, tomorrow, little, house, going, like, car, snow, hours, bed, dog, night, way	*After work. Only one thing on my mind.*
Family Gathering	good, morning, great, night, fun, day, time, weekend, dinner, week, friday, today, tonight, party, work, family, team, going, view, date, girls, weekend	*Had a great visit with Otto & family!*
Social & Sports	game, want, tony, retweeted, play, south, come, bend, dame, notre, it's, tulio, tickets, world, need, free, shit, dace, wants	*Watched my team in India play a friendly cricket match last night and got a lesson on the difference between batting in baseball versus cricket.*
Greetings & Celebration	day, happy, love, birthday, wedding, today, anniversary, halloween, disney, beautiful, mom, http, best, little, year, life, wish, challenge, thank	*Wishing my beautiful daughter a wonderful birthday. Love you baby girl.*
Friends & Family	years, time, love, family, friends, year, life, thanks, kids, amazing, best, know, today, old, wait, great, ago, days, help, people	*Enjoying St Helena, brunch and wine tasting with my son and friends.*
Activities & Interests	like, read, years, wow, know, love, good, think, people, music, interesting, post, facebook, ago, copy, itś, wheels, place, favorite, book	*First book Iv́e read in a long time that I couldn't put down. The Life We Bury*

Table S2. Coefficients of intrinsic traits in linear regression models with intrinsic traits as independent variables and deviation (SMAPE values) of the measures of social media use. Statistical significance reported as p-values (*<0.05, **<0.01, ***<0.001).

Trait	Posts		Words		Likes		Comments	
	SMAPE (100)	SMAPE (2W)	SMAPE (100)	SMAPE (2W)	SMAPE (100)	SMAPE (2W)	SMAPE (100)	SMAPE (2W)
Adj. R^2	0.98 ***	0.97 ***	0.93 ***	0.88 ***	0.95 ***	0.92 ***	0.94 ***	0.90 ***
Extraversion	0.14	0.48	-1.05 *	0.13	-1.09	-1.60 *	-0.94	-1.73 *
Agreeableness	0.44	1.22 **	-0.32	0.37	-0.40	-1.07	-0.19	-0.84
Conscientiousness	2.62 ***	4.35 ***	3.65 ***	2.61 ***	1.97 ***	3.45 ***	4.40 ***	5.10 ***
Neuroticism	0.73 *	0.28 *	2.24 **	0.73 **	2.38 ***	3.94 ***	2.66 ***	3.55 ***
Openness	-2.64 ***	-3.64 ***	-5.95 ***	-2.65 **	-5.00 ***	-7.45 ***	-6.89 ***	-9.76 ***
Shipley: Abs.	0.08 *	0.08	0.20	-0.01	0.27 *	0.28	0.24	0.28
Shipley: Voc.	0.09 **	0.04	0.39 ***	0.28 *	0.45 ***	0.54 ***	0.42 ***	0.51 ***
Pos. Affect	0.14 ***	0.22 ***	0.24 **	0.27 *	0.20 **	0.26 *	0.28 ***	0.32 **
Neg. Affect	-.05	-0.09	-0.06	-0.03	-0.02	-0.03	-0.06	-0.08
STAI: Anxiety	0.11 ***	0.08	0.32 ***	0.29 **	0.37 ***	0.46 ***	0.37 ***	0.48 ***
PSQI: Sleep Qual.	-0.01 *	0.10	-0.28	-0.59 *	-0.19	-0.35	-0.22	-0.27