HUMAN AND COMPUTERIZED PERSONALITY INFERENCES FROM DIGITAL FOOTPRINTS ON TWITTER

by

CORY K. COSTELLO

A DISSERTATION

Presented to the Department of Psychology
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2020

DISSERTATION APPROVAL PAGE

Student: Cory K. Costello

Title: Human And Computerized Personality Inferences From Digital Footprints On Twitter

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Psychology by:

Sanjay Srivastava	Chair
Nicholas Allen	Core Member
Robert Chavez	Core Member
Ryan Light	Institutional Representative

and

Kate Mondloch	Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2020.

DISSERTATION ABSTRACT

Cory K. Costello

Doctor of Philosophy

Department of Psychology

June 2020

Title: Human And Computerized Personality Inferences From Digital Footprints On Twitter

The increasing digitization of our social world has implications for personality, reputation, and their social consequences in online environments. The present dissertation is focused on how personality and reputation are reflected in digital footprints from the popular online social network Twitter, and the broader implications this has for the expression and perception of personality in online spaces. In three studies, I demonstrate that personality is reflected in the language people use in their tweets, the accounts they decide to follow, and how they construct their profile. I further examine moderators of accuracy including the number of users' tweets, the number of accounts they follow, and the density of their follower networks. Finally, I examine intra- and interpersonal consequences of being perceived accurately or ideally, speaking to the social functions of self-presentation in online environments. This multi-method investigation provides insight into how personality is represented online, how it can be recovered using computers and human judges, and the consequences this has for individuals.

CURRICULUM VITAE

NAME OF AUTHOR: Cory K. Costello

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
Wake Forest University, Winston-Salem, NC
New College of Florida, Sarasota, FL

DEGREES AWARDED:

Doctor of Philosophy, Psychology, 2020, University of Oregon
Master of Arts, Psychology, 2014, Wake Forest University
Bachelor of Arts, Psychology, 2012, New College of Florida

AREAS OF SPECIAL INTEREST:

Personality and Interpersonal Perception
Reputation
Data Science

PROFESSIONAL EXPERIENCE:

Graduate Employee, Department of Psychology, University of Oregon, Eugene, OR 2014-2020

PUBLICATIONS:

Costello, C. K. & Srivastava, S. (2020). Perceiving personality through the grapevine: A network approach to reputations. *Journal of Personality and Social Psychology.* Advance online publication. https://doi.org/10.1037/pspp0000362

Thalmayer, A. G., Saucier, G., Srivastava, S., Flournoy, J. C., & Costello, C. K. (2019). Ethics-Relevant Values in Adulthood: Longitudinal Findings from the Life and Time Study. *Journal of Personality* 87, 6,https://doi.org/10.1111/jopy.12462

Costello, C. K., Wood, D., & Tov, W. (2018). Examining East-West Personality Differences Indirectly Through Action Scenarios. *Journal of Cross-Cultural Psychology*, 49, 554-596. https://doi.org/10.1177/0022022118757914

Wood, D., Tov, W., & Costello, C. K. (2015). What a _____ Thing to Do! Formally Characterizing Actions by their Expected Effects, *Journal of Personality and Social Psychology*, 108, 953-976. http://dx.doi.org/10.1037/pspp0000030

ACKNOWLEDGEMENTS

I'd like to start by thanking my advisor and mentor, Sanjay Srivastava. I'm grateful for your unwavering support and encouragement over the past six years, for the invaluable advice, for the many great conversations, and most of all for training me to conduct rigorous science. I'll always look back fondly on the days sitting in your office discussing Mehl, trying to figure out a particularly tricky path model, or pouring over R output.

I'd like to thank my committee members, Rob Chavez, Nick Allen, and Ryan Light, for their helpful feedback in the design and development of this dissertation.

I'd like to thank the support staff in the University of Oregon's Psychology department for all of the help over the years, especially Lori Olsen, for making life as a graduate student so much more manageable.

I'd like to thank the members of the Personality and Social Dynamics Lab I had the good fortune of overlapping with – Pooya, Bradley, Cianna, Nicole, John, and Allison – for all of the help and feedback over the years. I always knew an idea had legs if it could survive a round of criticism in a lab meeting. I want to say a special thanks to Pooya who assisted with collecting the data presented here, and Cianna who assisted with data collection and blinded screening of the Study 3 data.

I'd like to thank my cohort mates - Grace Binion, Melissa Latham, Rita Ludwig, Adam Kishel (honorary member), and Brett Mercier. Thanks for making this place that is at the opposite end of the country from my family feel like home.

The experiences we shared - from Thanksgiving festivities, to grabbing beers at a local brewery, to the trips to conferences - were some of the best of my life, and I know I'll always look back fondly on them.

I'd like to thank my Mom, for always believing in me (maybe too much), supporting me, and helping me get to where I am today. There is too much to say thank you for on this page, but know that I know that none of my accomplishments would have been possible without you. I'd like to thank my Dad for always encouraging me to study hard and for teaching me the value of hard work. I'd like to thank my stepfather Peter for all of his support, and for always providing a word of encouragement when I needed it. I'd like to thank my stepmother Tracy and stepsister Taylor for the fun times over the years, especially the trip out to Oregon. I'd like to thank Keith and Lynn Oglesby for treating me like a member of the family and raising my favorite person. My extended family - but especially my grandma Nancy and late grandfather - also deserve a special thanks for their love, support, and encouragement.

I'd finally like to thank my wife and best friend, Katherine Oglesby. I can't begin to express how grateful I am for our life together, for your willingness to uproot and move across the country with me several times over, and all that you've done to help me achieve this goal. I could fill up all of the pages of this document and hardly scratch the surface of how grateful I am to you and for our relationship.

Dedication

For my grandfather, Jean Claude Bensimon, for all the laughs, the insightful conversations, and for teaching me to approach all things - including myself - with an open yet critical mind.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

Figure Page

LIST OF TABLES

TABLE Page

TABLE Page

I. INTRODUCTION

Our social world is becoming increasingly digitized, with much of our daily behavior and social interactions taking place in online environments. One unique aspect of behaving and interacting in online environments is that much of this behavior is recorded and stored in more or less permanent digital records. People and organizations use these *digital footprints* to draw inferences about the users that generated them. For example, it is commonplace to look someone up online and form (or update) an impression of them based on what turns up, a practice which has founds its way into formal processes like hiring decisions (Grasz, 2016). In addition to inferences made by people, machine learning algorithms are being used to infer psychological characteristics from digital footprints, with some research suggesting that they can outperform knowledgeable human perceivers (Youyou, Kosinski, & Stillwell, 2015). While previous work generally finds some degree of accuracy in human and computerized inferences from online behavior, there is considerable variability (Back et al., 2010; Kosinski, Stillwell, & Graepel, 2013; Park et al., 2015; Qiu, Lin, Ramsay, & Yang, 2012; Youyou et al., 2015). In my dissertation, I build on this work with a multi-method investigation into inferring personality from digital footprints available on Twitter. In three studies, I examine computerized inferences from tweets (Study 1), outgoing network ties (Study 2), and human inferences from profiles (Study 3), furthering our understanding of how personality is manifest in and recoverable from different digital footprints.

Digital Footprints, Identity Cues, and Behavioral Residue

Human- and computer-based personality judgments differ in many ways, but each require inferring a target person's standing on an unobservable psychological construct (e.g., how extraverted a person is) from observable cues the target produces in a given environment (e.g., a Tweet). The Brunswik (1955) Lens Model formalizes this as two underlying processes. *Cue validity* refers to the extent to which the construct produces valid and available cues in a particular environment and *cue utilization* refers to the extent to which judges use the cues correctly to render their judgment. Likewise, Funder's (1995) Realistic Accuracy Model (RAM) holds that accurate judgments of a construct require relevant cues be made available to the judge, which the judge then detects and properly utilizes. According to both models, accurate inferences from digital footprints - whether by a human perceiver or a computer - require access to valid cues and knowledge of how cues relate to the underlying psychological characteristics being judged.

Cues are often differentiated between those that incidentally vs. intentionally communicate aspects of ourselves' to others, generally referred to as *behavioral residue* and *identity claims* respectively (Gosling, Ko, Mannarelli, & Morris, 2002). Although typically discussed as a property of cues, it may be more fruitful to consider them as two theoretical processes that link underlying psychological characteristics to observed behavior. On the one hand, behaviors have certain predictable effects on the environment, which accumulate in frequented physical or digital spaces. This

accumulated behavioral residue incidentally provides insight into the psychological mechanisms that could have produced it and is thus characteristically *not* self-presentational. On the other hand, people use signals to intentionally communicate aspects of the self to others or to reinforce their own self-views. These identity claims are overt, self-presentational, and part of the broader identity negotiation process (Hogan, 2010; Swann, 1987) in which targets and perceivers mutually determine targets' identities.

Different approaches to inferring personality from digital footprints likely differ with respect to the how much they draw on behavioral residue vs. identity claims. Analyzing network ties like followed accounts probably relies more heavily on behavioral residue, since they are not prominently displayed and are a byproduct of following accounts. For example, following the American Psychological Association on Twitter may reflect a user's interest in psychology and even more distal characteristics (e.g., higher levels of a personality characteristic like openness), but it is unlikely that users follow this account specifically to communicate these aspects of their identity. At the other extreme, inferences based on profiles and their constituent parts (e.g., profile picture, bio, etc.) likely rely more heavily on identity claims, given that profiles are displayed prominently and function to communicate users' identities. Indeed, features like the bio exist primarily so that users can provide information about who they are to perceivers. Tweets likely rely on an even mix of behavioral residue (e.g., typos in tweets) and identity claims (e.g., statements of one's value).

Thus, differences between judgments made with tweets, network ties, and profiles might reflect different proportions of behavioral residue and identity claims.

Inferring Personality from Digital Footprints with Machine Learning Algorithms

Personality can be effectively inferred from digital footprints common across many OSNs, including the linguistic content of what a user posts online (e.g., Park et al., 2015) and their network ties (e.g., Facebook-like ties; Kosinski et al., 2013). Each are discussed below with a particular eye towards their points of difference.

Inferring Personality from Language. Personality and other psychological constructs (e.g., depression) can be effectively inferred from the language people use online, including Facebook status updates (Park et al., 2015) and tweets (Coppersmith, Harman, & Dredze, 2014; De Choudhury et al., 2013a, 2013b; Qiu et al., 2012; DeChoudhury2016; Dodds, Harris, Kloumann, Bliss, & Danforth, 2011; Golbeck, Robles, Edmondson, & Turner, 2011; Nadeem, 2016; Reece et al., 2017; Schwartz et al., 2013; Sumner, Byers, Boochever, & Park, 2012). Accuracy varies substantially across studies, likely due to several factors including the use of different techniques for quantifying and analyzing text, differences across different platforms due to what the technological architecture affords (e.g., length of Facebook posts vs. tweets), and norms that emerge on different platforms.

Within psychology, the two most common approaches to date for automated text analysis are dictionary-based and open-vocabulary approaches, which are

occasionally combined. Dictionary-based approaches generally work by matching the linguistic content a person produced with entries in a dictionary, which typically form one or more higher-order groups of words. For example, the Linguistic Inquiry Word Count software (LIWC; Tausczik & Pennebaker, 2010) is a commonly used dictionary-based approach which counts up the number of words associated with 69[1] different psychologically meaningful categories (e.g., first person singular pronouns, positive emotion words, biological processes, etc.). Other examples include sentiment analysis, where words are either counted (like LIWC) or scored for their relative positivity or negativity based on a pre-trained dictionary (Mohammad & Kiritchenko, 2015). While useful, dictionary-based approaches can miss important features of a text if those features aren't in the *a priori* dictionary. This might be especially concerning in online environments like Twitter, where abbreviations, slang, and terminology unique to the platform may be important features. This could explain the relatively poor accuracy found when predicting personality from Tweets using only dictionary-based approaches (e.g., r's from .13 to .18 in Golbeck et al., 2011; see also De Choudhury et al., 2013a, 2013b; DeChoudhury2016; Qiu et al., 2012; Reece et al., 2017; Sumner et al., 2012).

In contrast, the open-vocabulary approach is a data-driven alternative where words, phrases, and empirically-derived topics (e.g., from probabilistic topic models using Latent Dirichlet Allocation or LDA; Blei, 2012) are extracted from the text

[1] The exact number of categories depends on the version; I use the 2003 version in this dissertation, which has 69 categories.

without an *a priori* dictionary involved. This does require substantially more data than using pre-defined dictionaries, but it has the advantage of discovering non-obvious or unexpectedly important features in the text that might be missed by dictionary-based approaches. This advantage has proven to be worth the increased cost of training. Park et al. (2015), for instance, used an open vocabulary approach to predict personality from Facebook status updates with considerable accuracy (r's from .38 to .41; see also Coppersmith et al., 2014; Nadeem, 2016), outperforming the dictionary-based work mentioned above. While a substantial innovation over dictionary-based approaches, open-vocabulary approaches have significant limitations as well. One common to many text analytic approaches is the bag-of-words assumption, which holds that the order of words is irrelevant. This assumption, while absurd on its face, was necessary to make text analysis tractable for most purposes.

More advanced techniques have overcome this simplifying assumption by training vector embeddings of words using neural network architechtures, a set of techniques which have demonstrated superior performance in a variety of natural language processing tasks (Mikolov, Chen, Corrado, & Dean, 2006; Pennington, Socher, & Manning, 2014). These methods represent relations between words in semantic space with real valued vectors, based on word-word co-occurrences (e.g., "grad" and "student" often co-occurring) and word-context co-occurrences (e.g., "grad" and "undergrad" often preceding "student"). More recent approaches go even further, taking subword information into account by training n-gram character

embeddings and representing words as the sum of the n-grams they contain (Bojanowski, Grave, Joulin, & Mikolov, 2017). The major drawback of vector embeddings is that they require a substantial amount of data for training, which can be circumvented by using pre-trained vector embeddings.

Of course, the extent to which language use online predicts personality might vary across different OSN environments. Differences could emerge due to how the architecture of the platform shapes behavior. For example, the highest predictive accuracy for predicting personality from language use online was observed with Facebook status updates (Park et al., 2015), and one reason for this could be that the stricter character limits imposed by Twitter relative to Facebook make tweets more noisy (and therefore less predictive) than status updates. Thus, it's possible that tweets are less predictive of personality even when using more sophisticated analytic techniques.

Inferring Personality from Network Ties. Ties or connections on Twitter are directed, meaning that users can initiate outgoing ties (called "following" on Twitter) and receive incoming ties (called "being followed" on Twitter) which are not necessarily reciprocal. I'll refer to the group of users that a person follows as their followed accounts and the group of users that follow a person as their followers. While both ties are likely rich in psychological meaning, they almost certainly require different approaches. I'll focus exclusively on followed accounts within the context of inferring personality, treating them as individual features or predictors in predictive

models.

Although the psychological meaning of followed accounts is perhaps less immediately obvious than the psychological meaning of tweets, there are several reasons to suspect that it may be rich. One theory anticipating links between individuals' psychology and network ties is homophily, which holds that people like and therefore seek out others who are similar to themselves. For example, relatively extraverted individuals would be anticipated to differentially follow other similarly extraverted individuals or accounts. Homophily has been consistently observed (offline) for individual differences in emotion (Anderson, Keltner, & John, 2003; Watson, Beer, & McDade-Montez, 2014; Watson et al., 2000a, 2000b, 2004), mental health status such as depression (Schaefer, Kornienko, & Fox, 2011), and recently observed for personality among Facebook friends (Youyou, Stillwell, Schwartz, & Kosinski, 2017). We might thus expect some degree of personality homophily on Twitter, where people follow accounts based in part on perceived similarity. We can't examine this directly in the present study, but homophily would promote followed-account-based predictive accuracy.

More generally, following accounts is the primary way users' curate their feed or what they see when they log into the platform. Followed accounts thus likely reflect the kinds of information or experiences people are seeking out on Twitter, a broad expression of interest that likely reflects users' standings on personality characteristics to some degree. For example, Openness might be expressed by following accounts

that post intellectually stimulating content - such as artists, scientists, and other public thinkers. Considering followed accounts as an expressions of preferences and interests highlights their similarity to Facebook likes, a digital footprint which has been previously demonstrated to predict psychological characteristics with moderate accuracy (Kosinski et al., 2013; Youyou et al., 2015).

Language vs. Ties. The language in users' tweets and the accounts they follow are both promising predictors that differ in practical and theoretical terms relevant to the present investigation. Two critical theoretical differences are worth pointing out. The first stems from the distinction between active and passive social media use (Burke, Kraut, & Marlow, 2011). Active use refers to using social media to actively provide content, which on Twitter primarily includes tweeting and replying to others' tweets. Passive use refers to using social media to passively consume content provided by others. Active users differ with respect to tweet-frequency by definition, and so tweet-based approaches may achieve better accuracy predicting psychological characteristic of active users than passive users. The theoretical distinction between active and passive use does not make predictions about outgoing ties. However, it is possible that users that follow more accounts are more accurately captured by followed-accounts-based predictions, which would be consistent with prior work on Facebook likes (Kosinski et al., 2013; Youyou et al., 2015). I will examine the extent to which number of tweets and number of followed accounts affects accuracy in Studies 1 and 2 respectively, speaking to the extent to predictive

accuracy of different cues depends on how target users use the platform.

The Second critical theoretical difference between tweet content and followed accounts stems from the distinction between behavioral residue and identity claims. Although inferences from tweets and followed accounts are not strictly the product of behavioral residue or identity claims, it seems likely that tweets would rely more heavily on identity claims than followed accounts. Tweets are more overt and observable; once a user posts a tweet, it will appear in their followers' feeds, it might invite replies or interactions, and it will later be prominently displayed within the timeline feature of their own profile. Moreover, Tweets are language, and language is inherently social, intended to serve communicative and social functions (Tomasello, 2010). Tweets are thus intended to be consumed by an audience of perceivers. Followed accounts on the other hand are relatively less observable (though still viewable in a user's profile), and aren't generally intended to be consumed by others. Because of these differences, tweeting, relative to following accounts, may heighten public self-awareness, thereby increasing efforts to convey a particular impression via tweets (Leary & Kowalski, 1990). Digital footprints with relatively more identity claims than behavioral residue have been theorized to be better predictors of personality (Gladstone, Matz, & Lemaire, 2019), which would suggest that tweet-based predictions may be more accurate in general than followed-account-based predictions.

Another distinct possibility is that identity claims and behavioral residue are

better or worse predictors of different personality domains based on their level of evaluativeness (i.e., the desirability of being higher or lower on the dimension; John & Robins, 1993). Among the Big Five, Openness, Agreeableness, and Conscientiousness are relatively more evaluative, whereas Extraversion and Neuroticism are relatively less evaluative (John & Robins, 1993); Honesty-Propriety, the added sixth domain in the Big Six, is probably among the most evaluative dimensions. Desires to be seen positively will be expressed across all of the Big Six, but should be more heightened for the relatively more evaluative traits. These self-presentation efforts would affect identity claims more than behavioral residue, potentially leading to lower accuracy for tweet-based predictions for more evaluative traits (e.g., Openness). However, differences in accuracy across differently evaluative personality characteristics could also arise because self-reports, the accuracy criterion in this study, are worse indicators of evaluative constructs (Vazire, 2010).

Practically speaking, tweets and followed accounts have a lot in common. For example, they're both relatively sparse and noisy predictors (Kosinski, Wang, Lakkaraju, & Leskovec, 2016; Schwartz et al., 2013). There are also practical differences between them. Followed accounts can be relatively more straightforward to analyze, with the ties either included as individual predictors (e.g., Youyou et al., 2015) or subject to a data reduction technique like Singular Value Decomposition (SVD) or Principal Components Analysis (PCA; Kosinski et al., 2013). As outlined above, methods for quantifying text differ substantially (e.g., dictionary-based, open

vocabulary, embeddings, etc.), and the choice of method can drastically impact the accuracy of the corresponding model.

Inferring Personality from Digital Footprints with Human Judges

Digital footprints also provide a rich source for human perceivers to use in judging others' personalities, an opportunity recognized by the many hiring managers that report using social media searches in their decisions (Grasz, 2016). Indeed, human perceivers achieve considerable consensus and some degree of accuracy when judging targets' personalities based on their Facebook profiles or collections of their tweets (Back et al., 2010; Qiu et al., 2012). Personality judgments from digital footprints are thus moderately reliable and valid. Moreover, judgments based on Facebook profiles are closer to targets' real self (i.e., what they say they're really like) than their ideal self (i.e., how they wish they'd be seen by others), providing further evidence that Facebook profiles provide valid cues to targets' real (offline) personalities. This matched what Back and colleagues' (2010) referred to as the *extended real-life hypothesis*, which holds that people use online social networks as an extension of their offline lives, and thus present themselves relatively accurately online. Do we expect the extended real life hypothesis to hold for Twitter?

The only work to my knowledge that has examined personality judgments made by human perceivers from digital footprints on Twitter was conducted by Qiu et al. (2012), which does demonstrate accuracy for Big Five Agreeableness and Neuroticism. However, instead of providing perceivers with targets' profiles like the study by Back

et al. (2010) on Facebook profiles, Qiu et al. (2012) provided perceivers with pre-processed tweets in a text file. This is a serious shortcoming. On many OSNs, including Twitter, profiles are the hub of information about a user and include more information than what is available in Tweets, including profile and background pictures, screen names, the presence of a link to a professional blog, and other psychologically rich information provided by the target user. Thus, the use of tweets is a threat to ecological validity and likely provides lower-bound estimates of accuracy. Additionally, unlike Back et al. (2010), they did not measure how users want to be seen, preventing them from examining the extended real life hypothesis. Finally, a small methodological issue common to both studies is the use of small samples of undergraduate RAs for perceiver ratings instead of randomly sampling perceivers, which potentially limits the generalizability of their findings. Study 3 will address these shortcomings, examining the extent to which Twitter profiles provide human judges insight into target users' real or ideal selves.

Target Self-Presentation and Its Impact on Accuracy. Various theories hold that individuals want to be seen positively by others, engaging in idealized self-presentation to bolster their reputations and self-esteem (Hogan, 2010; Leary, 2007; Leary & Kowalski, 1990; Paulhus & Trapnell, 2008; Swann, Pelham, & Krull, 1989). As mentioned above, personality dimensions have a more and less desirable end (John & Robins, 1993) and the desire to present an idealized image would therefore affect how people present their personality. However, self-verification

theory holds that people have an even stronger desire to maintain their self-images, even if those images are less positive or desirable (Swann et al., 1989; Swann & Read, 1981). Twitter profiles, like Facebook profiles, might provide insight into users' true personalities, either because they engage more in self-verification than idealized self-presentation or because they fail at presenting their ideal self (e.g., they can't help but make many typos despite their attempt to present as highly conscientious). At the same time, it's possible that the public nature of Twitter heightens individuals' public self awareness (Leary & Kowalski, 1990), leading them to present an idealized front. Of course, there may be stable individual differences in both the extent and content of self-presentation (Paulhus & Trapnell, 2008). I will examine the extent to which profiles lead human perceivers to inferences more similar to target users' real or ideal self, and the extent to which this varies across targets.

Audience, Accountability, and Accuracy. In addition to targets' self-presentation, accuracy may vary as a function of the context users are in (Funder, 1995). In particular, I'm focusing on a users' followers, which constitute their audience of (known) perceivers online, as a contextual factor that might affect accuracy through its impact on target behavior.

Boyd (2007) and Hogan (2010) note that online interactions are unique in that they take place in front of an unimaginably large audience, unbounded by time and space. For example, when a person decides to tweet, what constitutes their audience? If the account is public, then the potential audience consists of anyone who has or

ever will have access to the internet (and can reach Twitter's servers), thus far exceeding the largest spatial-temporal boundaries one might encounter in even the most public offline contexts. Both Boyd (2007) and Hogan (2010) note that this large, unbounded audience has consequences for how people manage impressions online. Hogan (2010) suggests that rather than attempt to understand the scope and boundaries of their audience and how they might negotiate their identity given that audience, people instead consider two groups of perceivers: those whom they want to present an ideal self to, and those that might take issue with it (whom Hogan calls the lowest common denominator). This approach, called the *lowest common denominator* approach, suggests that understanding identity negotiation online requires considering the relative composition of target users' audience.

Hogan's (2010) lowest common denominator approach, Back and colleagues (2010) extended real life-hypothesis, and Swann's (1987) identity negotiation all place importance on the audiences' role in constraining self-presentation strategies. Moreover, these theories would predict differences, across individuals or OSN platforms, to the extent that the composition of the audience differs. Indeed, it's possible that people use Twitter differently than Facebook, using it to follow news and current events rather than connect with their offline friends and family. This could result in differences in audience composition, and therefore differences in self-presentation strategy. We can examine this indirectly by comparing our findings to that of Back and colleagues. Differences in audience composition across users

within a site may relate to self-presentation strategy, which we can examine directly in this study. We focus presently on the structure (rather than content) of one's audience on Twitter, focusing specifically on the density of users' follower networks as a moderator for accuracy of human inferences. Density captures the extent of interconnectedness in a network; denser networks are thought to enhance social support and trust, in part because they can more readily rally collective action to offer support or sanction bad behavior (Kadushin, 2012). This sanctioning of bad behavior might include dishonest self-presentation, leading to users in denser networks presenting themselves more honestly. We will examine this in Study 3 by assessing the relation between density and judgeability (i.e., how accurately people are able to judge a target user).

Consequences of Being Perceived Accurately or Ideally. What are the consequences for being perceived accurately or ideally? In a classic study, Swann and colleagues (1989) demonstrated that people have a desire for self-enhancement and self-verification, meaning they want to be seen positively and self-verifyingly (i.e., consistent with their self-perception), but prioritize self-verification over positivity. Do people have a desire for their profile to convey positive and self-verifying impressions? What happens when this desire is or is not satisfied? One possibility is that being perceived self-verifyingly and positively increases individuals' overall well-being, both by satisfying their identity negotiation goals and by providing the benefits that come along with it (e.g., being treated how one wants and expects to be treated by others).

However, given that people are simultaneously motivated to be seen positively and self-verifyingly, the relation between well-being and how one is perceived may not be so simple. Indeed, one can easily imagine that being perceived self-verifyingly might be more or less beneficial depending on where one lands in the distribution of a personality trait. For example, being mis-perceived on Agreeableness might have different implications for people higher or lower in Agreeableness. Being perceived accurately or ideally likely has *interpersonal* consequences as well. One example is likability, where individuals might be perceived as more or less likable in part based on how they're perceived. Indeed, recent evidence suggests that perceivers like targets that they perceive accurately, supporting the accuracy fosters liking hypothesis (Human, Carlson, Geukes, Nestler, & Back, 2018). However, it's also possible that accuracy's relation to liking depends on the target's personality. For example, it's possible that accurately perceiving a target is less associated with likability when the target is highly disagreeable. Response surface analysis (RSA; Barranti, Carlson, & Cote, 2017) can be used to examine these potentially complex effects of accurate or idealized perception, providing insight into how different kinds of (in)accuracy and idealization impact targets' well-being and likability. This will be the focus on Aim 3c in Study 3.

Overview of Present Studies

In three Studies, I examine personality inferences from digital footprints including computerized inferences from tweets (Study 1), outgoing network ties

(Study 2), and human perceivers' inferences from targets' profiles (Study 3). All three studies draw upon two samples we've collected as part of an NIMH- (Grant # 1 R21 MH106879-01) and an NSF- (GRANT # 1551817) funded project. The methodological details common to the three studies are described next, followed by the specific methods and results of each study.

Samples & Procedure. General data collection includes two samples, I'll refer to as the NIMH sample and the NSF sample. Data collection for the NIMH and NSF samples were approved by the University of Oregon Institutional Review Board (Protocol # 12082014.013 for NIMH; Protocol # 10122017.011 for NSF) and were conducted in a manner consistent with the ethical treatment of human subjects.

In both samples, our inclusion criteria required participants to provide an existing unlocked Twitter account, to currently reside in the US, to primarily tweet in English, and to meet minimum thresholds for being an active Twitter user. Minimally active twitter users were defined as having at least 25 tweets, 25 friends, and 25 followers. Using two-stage prescreening, we attempted to first screen participants for eligibility before they completed the main survey; participants had to affirm that they met the inclusion criteria before they proceeded with the main survey. However, since participants could erroneously state that they met the inclusion criteria, each participant was individually screened to verify that they indeed met the criteria, and to further assess whether the Twitter handle belonged to the participant whom provided it. This consisted of manually searching each Twitter account provided,

ensuring it met the activity thresholds, and assessing whether the account provided was obviously fake (e.g., one participant provided Lady Gaga's account and was subsequently excluded). When it was especially difficult to verify that the accounts provided belonged to participants, we asked them to confirm that they owned the account they provided by direct messaging our lab's Twitter account from the account they provided.

For both samples, we then downloaded each eligible participant's data from Twitter's API, including their full friends list, their user data (i.e., the information displayed in their profile), and up to 3200 of their most recent tweets, retweets, and replies.

NIMH Sample. The NIMH sample was collected from the Spring of 2016 until the Fall of 2017, recruiting participants primarily from the "r/beermoney" and "r/mturk" Reddit communities, with additional participants from the University of Oregon Human Subjects Pool (UOHSP), Amazon's Mechanical Turk (mTurk), and Twitter advertising (using promoted tweets).

In all recruitment methods, participants were able to click a link that took them to the Qualtrics survey where they provided their Twitter handles, answered some questions about their Twitter use, completed several self-report measures (described below), and finally completed basic demographics questions. At the end of the survey, participants were thanked, and compensated either with an Amazon gift card or physical check for $10 or with course credit for participants recruited through the

human subjects pool.

This process led to a total of $n_{nimh-initial} = 756$ accounts that we were able to verify met our inclusion criteria. Ineligible prescreen participants contained a mixture of participants who did not provide an existing Twitter account, participants who provided an account that they did not own (e.g., Lady Gaga's account), participants whose Twitter account did not meet the activity thresholds, and participants that provided an eligible but locked account.

Of the 756 eligible accounts, we successfully retrieved tweets for $n_{nih\ tweets} = 487$ and followed accounts for $n_{nih\ followeds} = 638$. Note that these different sample sizes generally arise from being unable to download participants' tweets or followed accounts, because participants either deleted, locked, or changed their account name between the time when they were verified as eligible and when we downloaded their twitter data (a lag which sometimes extended for months).

NSF Sample. The NSF sample was collected from February 2018 to March 2020. Participants were recruited from the "r/beermoney" Reddit community and consisted of an initial sample of $n_{nsf-inital} = 654$ that met inclusion criteria and completed the Big Six questionnaire. Of these participants, we were able to successfully retrieve tweets for $n_{nsf-tweets} = 614$ participants and followed accounts for $n_{nsf-followeds} = 639$ participants. As with the NIH sample, the difference in sample sizes reflects participants who either deleted, locked, or changed the name of their account before we downloaded their twitter data.

Table 1
Participant Gender for Study 1 and Study 2 Samples

gender	n_{S1}	n_{S2}
Female	404	505
Male	673	746
Other	12	13
unknown/not reported	12	15

Note. Targets for Study 3 were also drawn from these samples, but their demographic information is provided in the Study 3 Methods section. The majority of participants provided data for Studies 1 and 2.

Participants in both samples responded to demographic questions reflecting NIH enrollment reporting standards. Gender, race, and ethnicity for both samples are shown in Tables 1, Tables 2, Tables 3, respectively. These are broken down by participants used in tweet-based analyses (Study 1) and followed-account-based analyses (Study 2), but keep in mind that these are mostly the same participants. Study 1 participants ranged in age from 14 to 68 with an average age of 27.12. Study 2 participants ranged in age from 14 to 68 with an average age of 26.85.

Measures. Both sample completed self-reports of the Big Six personality domains using a combination of two instruments. The Big Five (extraversion, agreeableness, conscientiousness, negative emotionality, and openness) were measured using the Big Five Inventory 2 (Soto & John, 2017b), which consists of 60 short statements rated on a scale from one (Disagree strongly) to five (Agree strongly) with a neutral point of three (neither agree nor disagree). We used eight items from the

Table 2
Participant Race for NIH and NSF Samples

race	n_{S1}	n_{S2}
American Indian / Alaskan Native	8	9
Asian	135	147
Black / African American	70	86
more than 1 race	92	107
Native Hawaiian / Pacific Islander	1	1
White	783	913
unknown / not reported	12	16

Note. Targets for Study 3 were also drawn from these samples, but their demographic information is provided in the Study 3 Methods section. The majority of participants provided data for Studies 1 and 2.

Table 3
Participant Race for NIH and NSF Samples

ethnicity	n_{S1}	n_{S2}
hisanic/latino	138	154
not hispanic/latino	951	1110
unknown/not reported	12	15

Note. Targets for Study 3 were also drawn from these samples, but their demographic information is provided in the Study 3 Methods section. The majority of participants provided data for Studies 1 and 2.

Questionnaire Big Six family of measures to measure the sixth domain, honesty-propriety (Thalmayer & Saucier, 2014), rated on the same scale. These scales showed adequate internal consistency, with alphas ranging from a low of .64 for honesty-propriety and .92 for neuroticism. NSF participants completed additional measures relevant to Study 3 describe in its method section below.

Analyses. Unless otherwise noted, all analyses were conducted in R (Version 4.0.2; R Core Team, 2019) and the R-packages *broom.mixed* (Version 0.2.6; Bolker & Robinson, 2020), *caret* (Version 6.0.86; Kuhn et al., 2019), *dplyr* (Version 0.8.5; Wickham et al., 2019), *forcats* (Version 0.5.0; Wickham, 2019a), *ggplot2* (Version 3.3.1; Wickham, 2016), *igraph* (Version 1.2.5; Csardi & Nepusz, 2006), *lattice* (Version 0.20.41; Sarkar, 2008), *lavaan* (Version 0.6.7; Rosseel, 2012), *lme4* (Version 1.1.23; Bates, Mächler, Bolker, & Walker, 2015), *lmerTest* (Version 3.1.2; Kuznetsova, Brockhoff, & Christensen, 2017), *Matrix* (Version 1.2.18; Bates & Maechler, 2019), *papaja* (Version 0.1.0.9942; Aust & Barth, 2018), *purrr* (Version 0.3.4; Henry & Wickham, 2019), *quanteda* (Version 2.0.1; Benoit et al., 2018), *readr* (Version 1.3.1; Wickham, Hester, & Francois, 2018), *rio* (Version 0.5.16; Chan, Chan, Leeper, & Becker, 2018), *RSA* (Version 0.10.0; Schönbrodt & Humberg, 2018), *shiny* (Version 1.4.0.2; Chang, Cheng, Allaire, Xie, & McPherson, 2019), *stringr* (Version 1.4.0; Wickham, 2019b), *tibble* (Version 3.0.1; Müller & Wickham, 2019), *tidyr* (Version 1.1.0; Wickham & Henry, 2019), and *tidyverse* (Version 1.3.0; Wickham, 2017).

II. STUDY 1: PREDICTING PERSONALITY FROM TWEETS

Study 1 examines computerized judgments made from the language people share online in their tweets. In the first of two aims (*Aim 1a*), I examine the extent to which tweets can be used to predict self-reported personality, using a cross-validated machine learning approach. This will combine unsupervised machine learning methods for data reduction and supervised machine learning techniques to predict self-reports (from these reduced data). I'll evaluate tweet-based models in terms of their ability to predict self-reports of new users (from only their tweets), and the extent to which the models are consistent with theoretical understandings of the predicted constructs. In the second aim (*Aim 1b*), I'll examine the extent to which *how* people use twitter affects predictive accuracy, examining both number of tweets and number of followed accounts. This study will provide insight into how personality relates to what people talk about online, how accurately we can infer personality from online language, and the extent to which this depends on the how people engage with the platform.

Methods

Samples & Procedure. Study 1 used all eligible participants from both the NIMH and NSF samples that completed Big Six personality measures and for whom we were able to successfully retrieve tweets. This resulted in a total sample of $N_{combined-twets} = 1101$ (see Tables 1 to 3 for participant gender, race, and ethnicity).

Analytic Procedure. In aim 1a, I predicted personality from the language in users' tweets using a procedure designed to minimize overfitting and data leakage in estimating predictive accuracy, while also providing insight into how different analytic decisions (e.g., scoring with dictionaries vs. vector embeddings) affect predictive accuracy.

Data Partitioning. We first split the final sample (N = 1101) into a training and holdout (testing) set using the Caret package in R (Kuhn et al., 2019). The training and holdout samples consisted of approximately 80% ($n_{training} = 882$) and 20% ($n_{holdout} = 219$) of the data respectively. All feature selection, data reduction, model training, estimation, and selection was determined from the training data. The final model(s), trained and selected within the training data, were tested on the holdout sample to get an unbiased estimate of out-of-sample accuracy.

Preparing & Pre-processing Tweets. Tweets were *tokenized* into individual words and short (two-word) phrases using an emoji-aware tokenizer from the quanteda package in R (Benoit et al., 2018). Then, they were *scored* using three techniques: dictionaries, open-vocabulary, and vector embeddings.

Tweets were scored using the 2003 version of the LIWC dictionary (Tausczik & Pennebaker, 2010), and a sentiment and emotion dictionary designed for and validated with tweets (Mohammad & Kiritchenko, 2015). LIWC scores are proportions of words from each category relative to the total number of words in

users' tweets. The sentiment and emotion dictionaries have continuous scores for each term in their dictionary; sentiment scores in this dictionary theoretically range from negative infinity (maximally negative sentiment) to positive infinity (maximally positive sentiment), and emotion scores range from 0 (not relevant to emotion label) to positive infinity (maximally relevant to emotion label). Each participant received a single score for sentiment and the eight specific emotions, corresponding to the average scores across all the words in their downloadable tweet history (e.g., the average anger score across every word in their downloadable tweet history).

Open-vocabulary analyses included two types of features: 1) Individual words and two-word phrases and 2) topics extracted using Latent Dirichlet Allocation (LDA; Blei, 2012), a data reduction technique that extracts topics based on the extent to which words co-occur across documents (tweet-histories in this case). There were 4.8 Million words and two-word phrases in the training users' tweets, which is far beyond what is computational feasible or efficient. After some trial and error, we limited individual words and two-word phrases to those which were used at least once by 25% of the training sample; this reduced the number of individual words and phrases to 3,060. We then scored individual words and phrases as proportions such that each represents a words' and phrases' frequency in users' tweets relative to their total number of words (across all tweets). We performed LDA topic models on just single words and used a more generous threshold of 1% (i.e., words had to be used at least once by 1% of our participants to be included in the topic models) and extracted 300

topics. LDA topic modeling results in a continuous score for each word in the corpus and each topic extracted that corresponds to a word's probability of belonging to a topic, analogous to a factor loading for each item in a multi-dimensional scale. Each participants' full tweet history was scored for topics using these continuous scores, analogously to factor scoring a set of items based on their loadings.

Tweets were also scored with (pre-trained) vector embeddings from two different approaches. GloVe word embeddings trained on tweets by Pennington et al. (2014) were downloaded from their website (https://nlp.stanford.edu/projects/glove/) and applied to participants' tweets. Likewise, word vectors derived from fastText character embeddings trained by Mikolov, Grave, Bojanowski, Puhrsch, and Joulin (2017) were downloaded from their website (https://fasttext.cc/docs/en/english-vectors.html). Then, word vectors were averaged within participants, resulting in a single score per vector for each participant; though this technique is less optimal than training a weighted model, it works reasonably well for short texts and circumvents the need for large training data sets. This resulted in 500 vector scores corresponding to the 200 GloVe and 300 FastText vectors.

Model training. Dictionary scores, word and phrase proportion scores, topic scores, and vector scores were included as predictors or features in predictive models using two different approaches. Each personality trait was modeled separately, and so the model trained and selected for one construct (e.g., extraversion) could differ in every respect (approach, hyperparameters, parameters) from the model trained and

selected for another construct (e.g., conscientiousness). All models were trained, tuned, and evaluated (within-training evaluation) using k-fold cross-validation. This splits the data into k random subsets called folds, trains the data with k-1 folds, and tests the model's performance on the kth fold; this is repeated until each fold has been the test fold. We set k to 10, which is commonly recommended (Kosinski et al., 2016). This procedure is an efficient means for reducing overfitting during model training and selection (Yarkoni & Westfall, 2017).

Linguistic Feature Selection. We trained models on different subsets of linguistic features. There were five sets of features in total, consisting of (1) dictionary-based scores, (2) all open-vocabulary feature (words, phrases, and topics), (3) topic scores, (4) vector scores from GloVe and FastText word embeddings, and (5) all of the features (dictionaries, open-vocabulary features, and vector scores).

Modeling Approaches. I compared two different modeling approaches: Ridge Regression and Random Forests. Each is described in greater detail below.

Mirroring Park et al. (2015)'s approach to predicting personality from Facebook status updates, I trained models predicting self-reported Big Six personality scores from linguistic features with ridge regression. Ridge regression is a penalized regression model, which minimizes the sum of squared errors *and* the L2 penalty, or the sum of squared coefficient values (i.e., $\lambda * \Sigma_{j=1}^{j=\beta_j} B_j^2$, where λ is a scaling parameter that determines the weight of the penalty). It has the effect of shrinking coefficients

to be closer to zero. Ridge can provide relatively interpretable solutions when predictors are uncorrelated, but can be misleading in the face of correlated predictors.

The second approach was Random Forests algorithm. Random Forests works by iteratively taking a subset of observations (or cases) and predictors, building a regression tree (i.e., a series of predictor-based decision rules to determine the value of the outcome variable) with the subset of predictors and observations, and averaging across the iterations. It is thus an ensemble method, which avoids overfitting by averaging across many models trained on different subsets of participants and features. It works well with sparse predictors (Kuhn & Johnson, 2013), making it a promising candidate for tweet-based predictions, especially using the sparser feature-sets (e.g.„ word- and phrase-proportions). Like ridge regression, interpretation can be difficult in the presence of correlated predictors, though the permutation importance metric (used here) is relatively robust to correlated predictors (Genuer, Poggi, & Tuleau-Malot, 2010).

Model selection. As mentioned above, all models were trained using the training data, and each model's training performance was indexed via root mean squared error (RMSE) and the multiple correlation (R) from 10-fold cross-validation. Although machine learning approaches tend to prioritize predictive accuracy over interpretability (Yarkoni & Westfall, 2017), we aim to maximize both to the extent possible. As such, we based our model selection on both (quantitative) model performance criteria (minimal RMSE, maximal multiple R) and (qualitative)

interpretability. Note that in addition to RMSE/R for the best performing model, we also considered the spread of training results (e.g., we may choose a model that did not have the best single performance, if it has less variability in performance).

Model evaluation. We selected our candidate models based on the training data, completed an interim registration of our model selection (available at https://osf.io/4xbcd/?view_only=2916632373d3410bbf02f94650e50b1d), and then tested the selected models' accuracy using the (heldout) test data. To guard against overfitting, we selected one candidate model per outcome variable. In addition to our candidate models, we tested the out-of-sample accuracy for the non-selected models as exploratory analyses, but we clearly distinguish selected from non-selected models (which can be verified in our registration). This provides an estimate of accuracy that is unbiased by selection (accuracy from selected models) as well as some insight into the extent to which our selection process resulted in the best model.

Aim 1b: moderator analyses. After selecting the model and evaluating it on the holdout set, we used the tweet-based predicted personality scores for all 1102 participants in a series of OLS moderated multiple regressions. In these analyses, actual self-reported personality scores were regressed on tweet-based scores, number of tweets (followed accounts), and their interaction, with a significant interaction indicating an effect of number of tweets (followed accounts) on tweet-based predictive accuracy. Each of the Big Six personality domains were examined separately,

resulting in 12 total moderator analyses.

Results

Aim 1a: Predictive Accuracy. Below I describe our results from model training, which models we selected for the holdout dataset, and how accurate the selected and non-selected models were in the holdout dataset.

Model Training & Selection. First, I examined the accuracy with which each combination of feature set and modeling approach could predict self-reported Big Six Domains, focusing on the average R and RMSE for predicting the holdout-folds in the 10-fold cross-validation procedure. Figure 1 shows the average R (Panel A) and RMSE (Panel B) for each combination of feature-set (y-axes) and modeling approach (color); each dot represents the average R and RMSE for each set of hyperparameters and the bar represents the average (of average Rs or RMSEs) across hyperparameter specifications. Big Six domains are shown in separate panels, indicated with the first letter of the domain name. Note that some specifications of Ridge with LDA topics are omitted from the RMSE plot because they were an order of magnitude greater and beyond the limits set on the x-axis.

Figure 1 demonstrates that personality can be predicted from linguistic features of tweets with at least some degree of accuracy using different combinations features, modeling approaches, and hyperparameter specifications. Moreover, it is apparent in Figure 1 that Random Forests outperformed ridge with only a few exceptions.

Accuracy was relatively similar across feature sets, with the possible exception of LDA topics (on their own), which tended to be less accurate across domains. This is somewhat surprising given that the dictionary models used 77 predictors and the "all" models used over 3,000 predictors. Figure 2 shows these same metrics for the best hyperparameter specification per modeling approach and set of features, and paints a similar picture. Accuracy was thus considerably higher for random forests, and there was little difference between feature sets. Though feature sets had only marginal differences in accuracy, dictionaries were best for agreeableness, and using all features simultaneously was best for the other five domains.

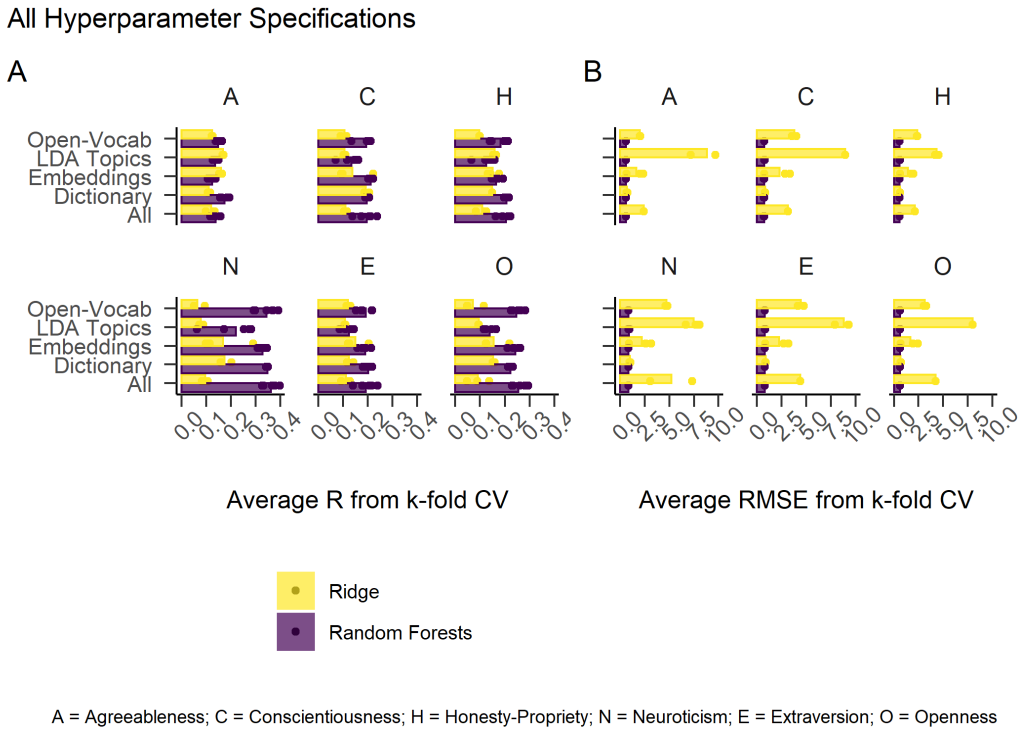

Figure 1. K-Fold CV Accuracy for Predicting Personality from Tweets (All Model Specifications).

Interpretability. Judging models strictly by accuracy, random forests achieved

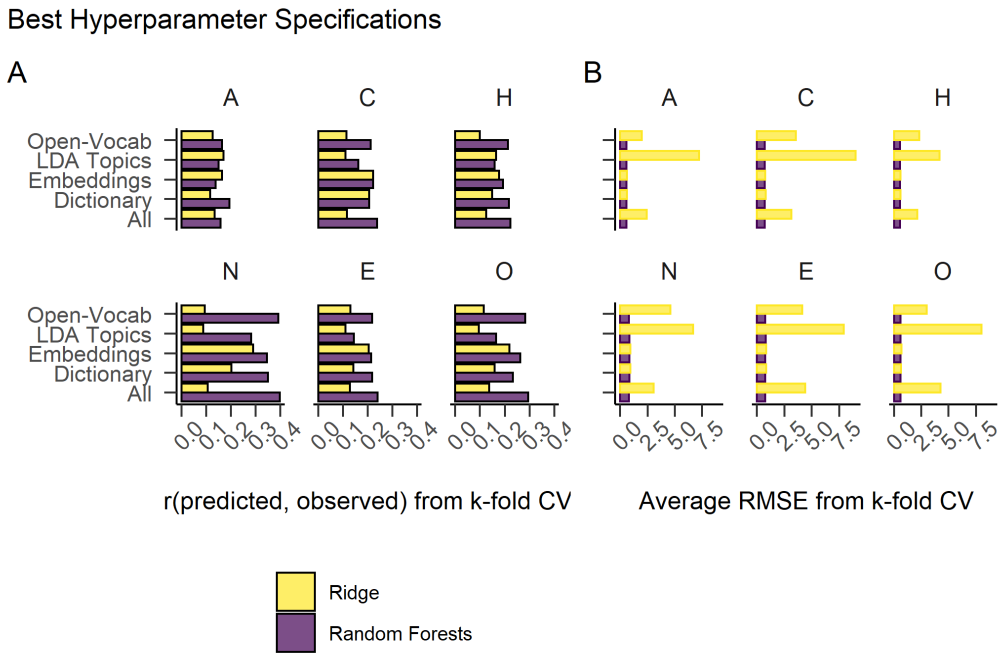

Figure 2. K-Fold CV Accuracy for Predicting Personality from Tweets (Best Model Specifications).

greater accuracy and there was little differentiation among feature sets.

Interpretability proved helpful in this case, as models *did* differ in terms of how apparently consistent with prior theory they were. With the exception of the notably difficult to interpret vector embeddings, models trained with different feature sets all had some degree of consistency with prior theory. For example, the two-word phrase "thanks-much" was one of the most important predictors of agreeableness in the model trained with all features, the swear words category from LIWC was a highly important predictor for conscientiousness in the models trained with dictionaries, "anxieti-" stemmed words were the most important predictor of neuroticism in models trained with open-vocab features, tagging users was one of the most important

predictors of extraversion in the models trained with LDA topics, and the word stem "creativ" was one of the most important predictors of openness using either open-vocab or all features in training. However, the models that generally stood out in terms of interpretability were the models trained with dictionary scores, which are described in greater detail next.

Figure 3 shows the permutation importance scores (from the best-fitting random forests model) of the dictionary categories in predicting agreeableness. To ease interpretation, bars are colored based on whether they are positive (blue) or negative (red) zero-order correlates, though it is important to keep in mind that their role in the random forests prediction algorithm may be less straightforward (e.g., not linear and additive). Dictionaries include both LIWC and NRC sentiment and emotion scores; NRC sentiment and emotion scores are all prefixed with "m_", which can help differentiate the two dictionaries. It seems that the model is picking up on theoretically relevant content, including swear words, LIWC's negative emotion group (e.g., abandon, abuse), LIWC's anger category (e.g., aggressive, agitate), LIWC's positive feelings (e.g., adore, agree*), inclusive words (e.g., altogether), negations (e.g., can't, don't) and other theoretically relevant content. Together, this seems to capture agreeableness's core content of interpersonal warmth vs. antagonism.

Figure 4 shows the same information for conscientiousness, where important features include NRC's anger category, swear words, time words (e.g., age, hour, day), sexual words, NRC's sentiment score, school-related words, negative emotions,

pronouns, and leisure activity. These features may reflect aspects of conscientiousness like industriousness, punctuality, and impulsivity.

Figure 5 shows the same information for honesty-propriety, where you can see that important categories included negative emotion, school, leisure activities, metaphysics (e.g., bless, angels), LIWC's anger category, sexual words, positive emotions, and third-person pronouns (labeled "Other") and second-singular pronouns (labeled "You"). Interestingly, it overlaps somewhat with agreeableness and conscientiousness, but also seems to be picking up on some core moral content with the metaphysics category.

Figure 6 shows the same information for neuroticism, which shows that important categories included core affective content, including negative emotions like NRC anger, LIWC anxiety, NRC disgust, and NRC sadness, positive emotions content such as anticipation and surprise, and sentiment (which ranges from negative to positive). Important categories also included time, friends, pronouns, the up category (e.g., high, on, top), and other indirectly relevant content.

Figure 7 shows the same information for extraversion, which shows that important categories included NRC anger, school, discrepancies (e.g., should, would, could), other (3rd person pronouns), optimism, exclusive words (e.g., but, without), humans (e.g., boy, woman, adult), tentativeness (e.g., anyhow, ambiguous), and causation (e.g., because, affected). This model was harder to interpret than the others, but the models did seem to pick up on an assertiveness vs. tentativeness theme. It is

worth noting that extraversion is one case in which the open-vocab seemed to pick up on relevant themes, with highly important words referring to more mainstream or niche cultural interests (sports-related words like team vs. draw and anime).

Figure 8 shows the same information for openness, where important word-categories included occupation-related words (accomplish, advance, administration), communication words (e.g., admit, suggest, informs), school words, hearing words (e.g., listening, speaking), insight words (e.g., analyze, understand, wonder), negative emotions, optimism, music, achievement, and other relevant content. This might correspond to pursuing and expressing intellectual and aesthetic interests on twitter, behavior highly characteristic of high openness.

Selected models. The choice of algorithm was a simple one here: random forests showed consistently greater accuracy in training than ridge regression and importance scores mapped onto theoretically consistent themes for each domain. Selecting a feature set was more challenging, given the similarity in accuracy achieved with different feature sets. Consequently, we used interpretability as a guiding principle in this selection process and ultimately selected the dictionary-based models. The dictionary-based models were either the most accurate (agreeableness) or a close second or third (difference in R's $\leq .1$), and were often more interpretable than the alternatives. Within this selection, RMSE and R agreed with respect to the most accurate set of hyperparameters, and so we selected these specifications as our final models. The specifications for these final, selected models are shown (alongside

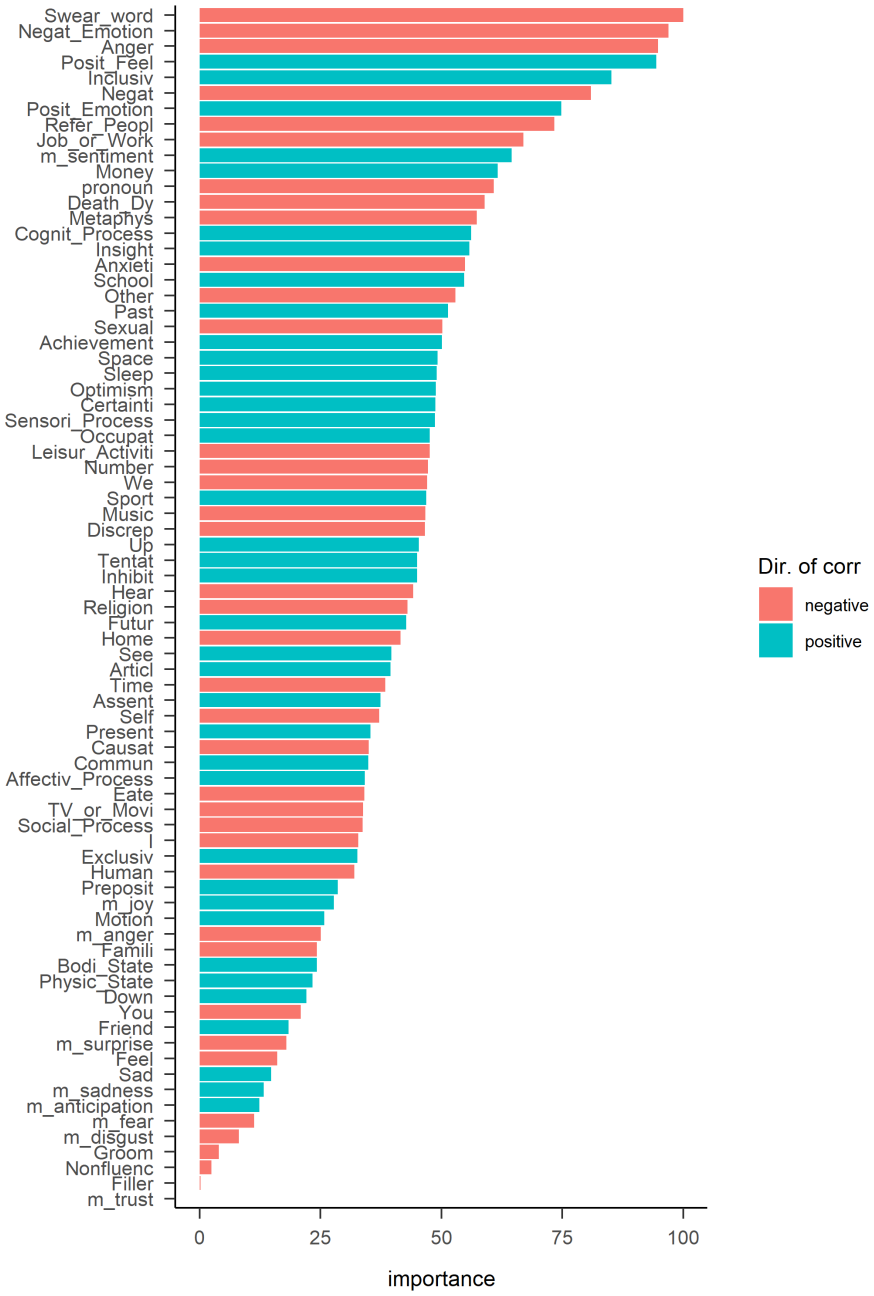

Figure 3. Importance Scores from Random Forests Predicting Agreeableness with Dictionary Scores

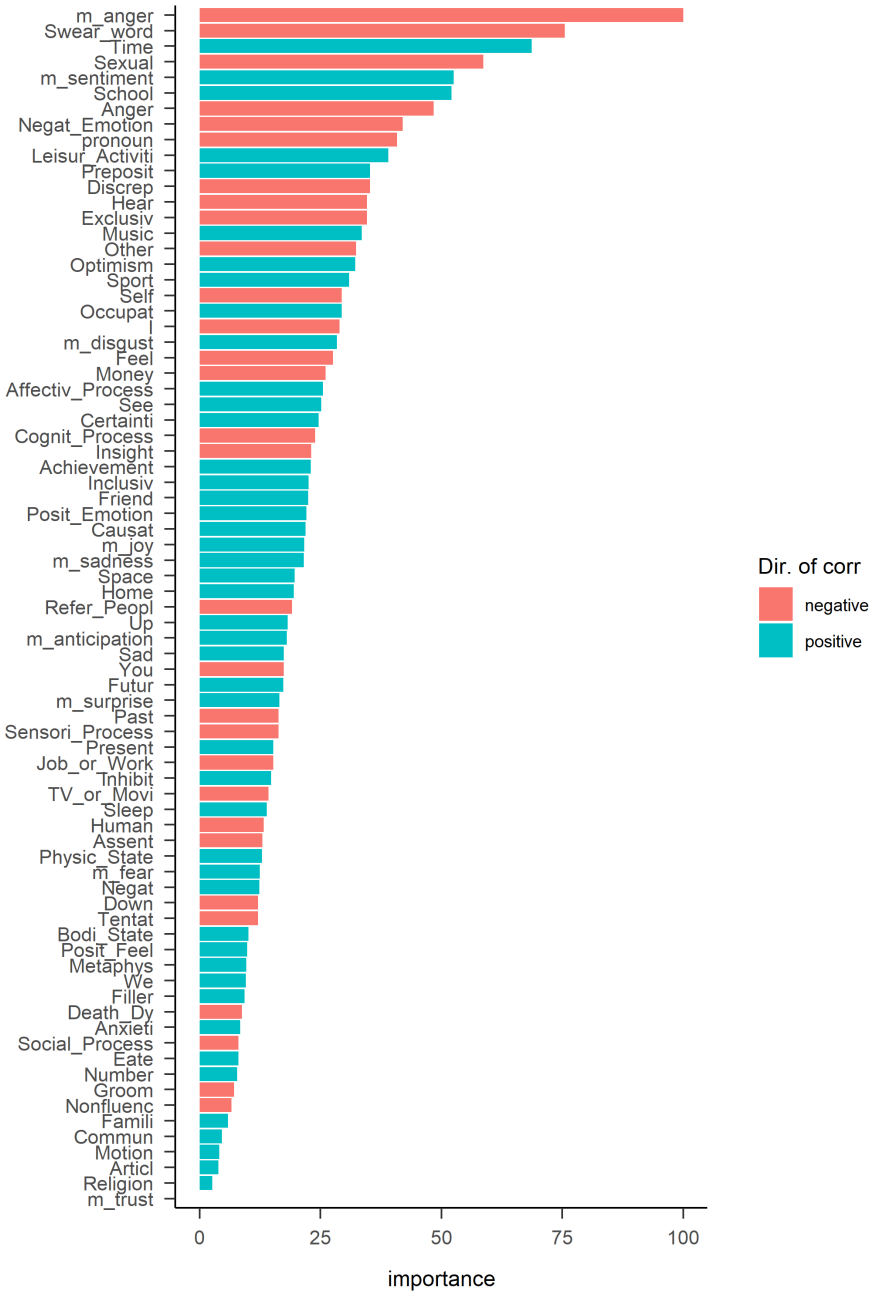

Figure 4. Importance Scores from Random Forests Predicting Conscientiousness with Dictionary Scores

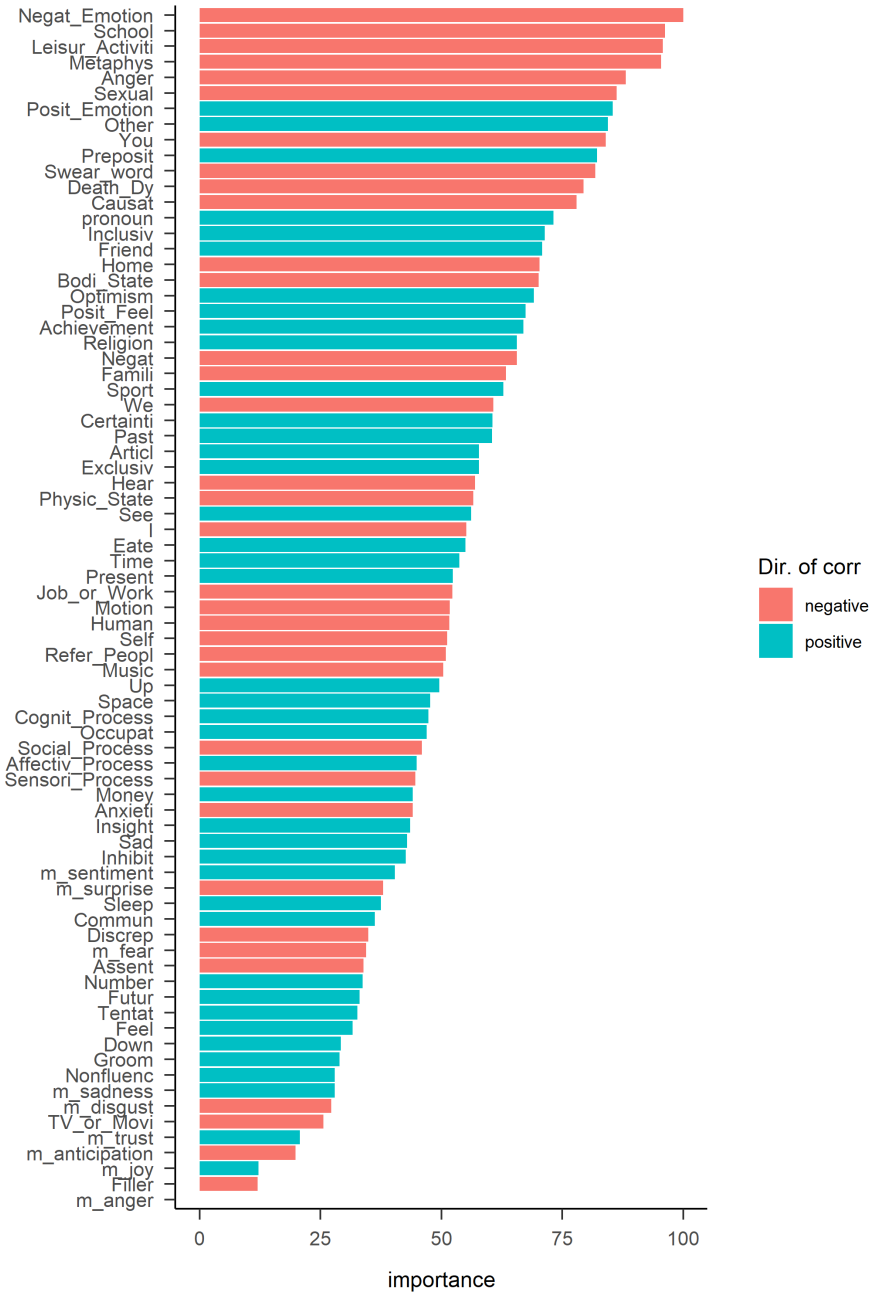

Figure 5. Importance Scores from Random Forests Predicting Honesty-Propriety with Dictionary Scores

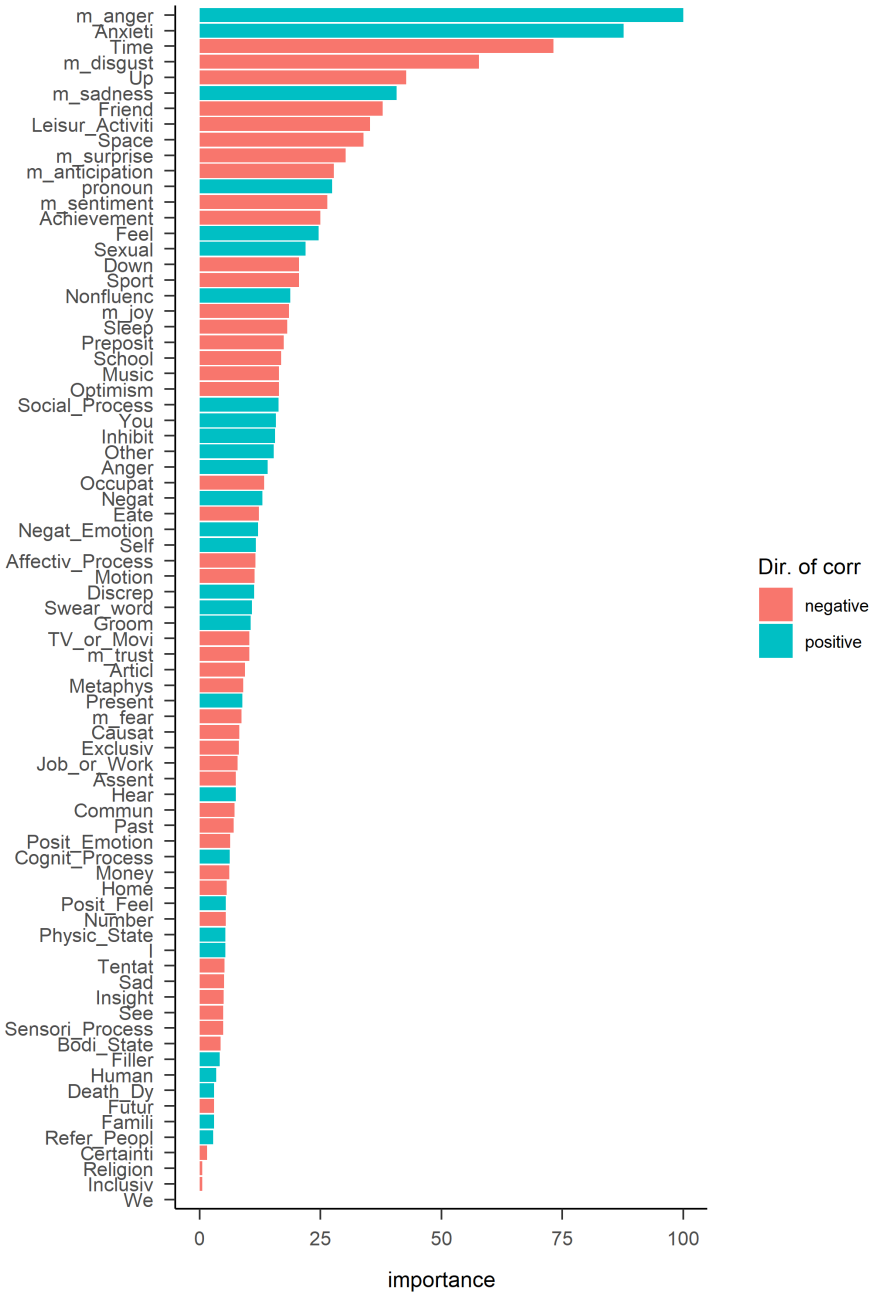

Figure 6. Importance Scores from Random Forests Predicting Neuroticism with Dictionary Scores

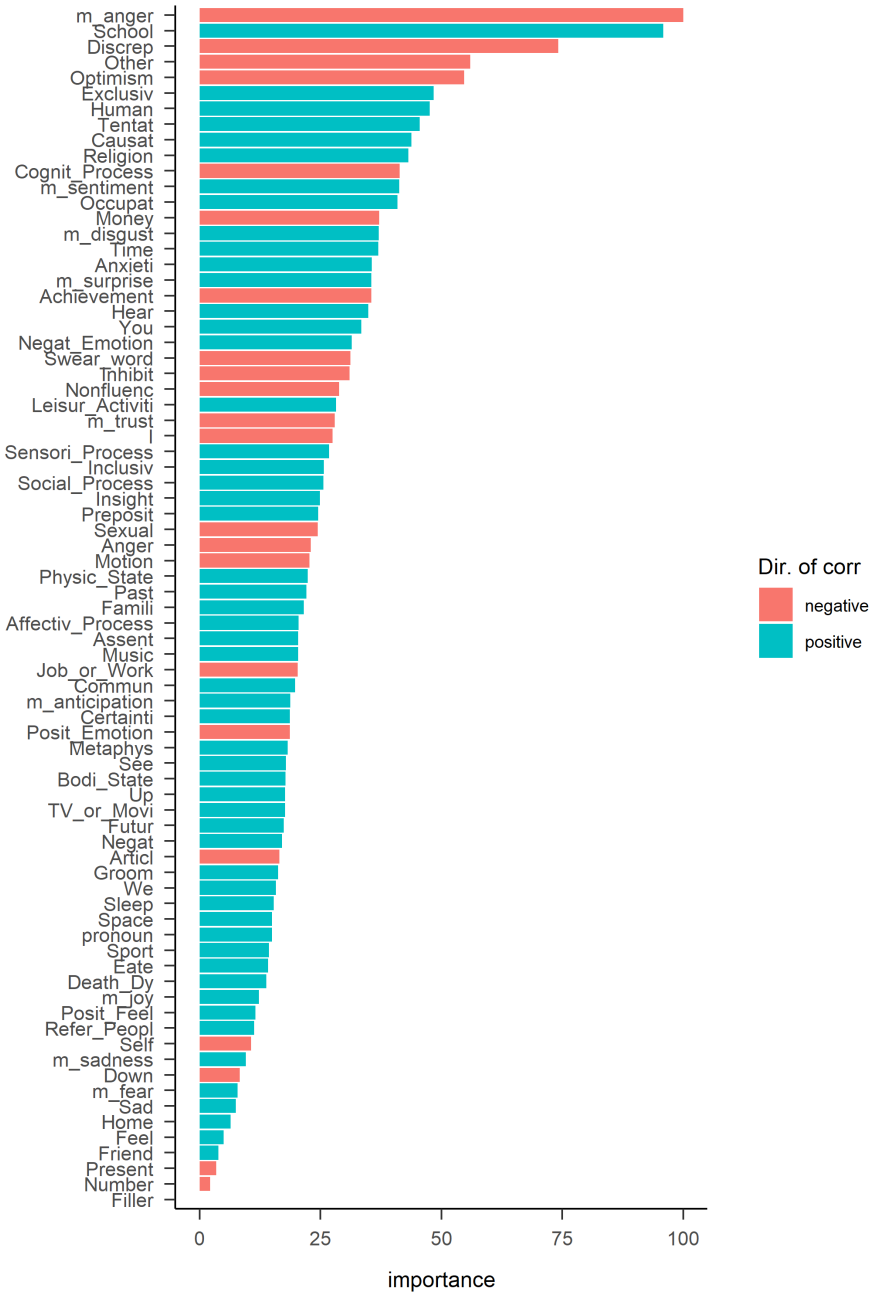

Figure 7. Importance Scores from Random Forests Predicting Extraversion with Dictionary Scores

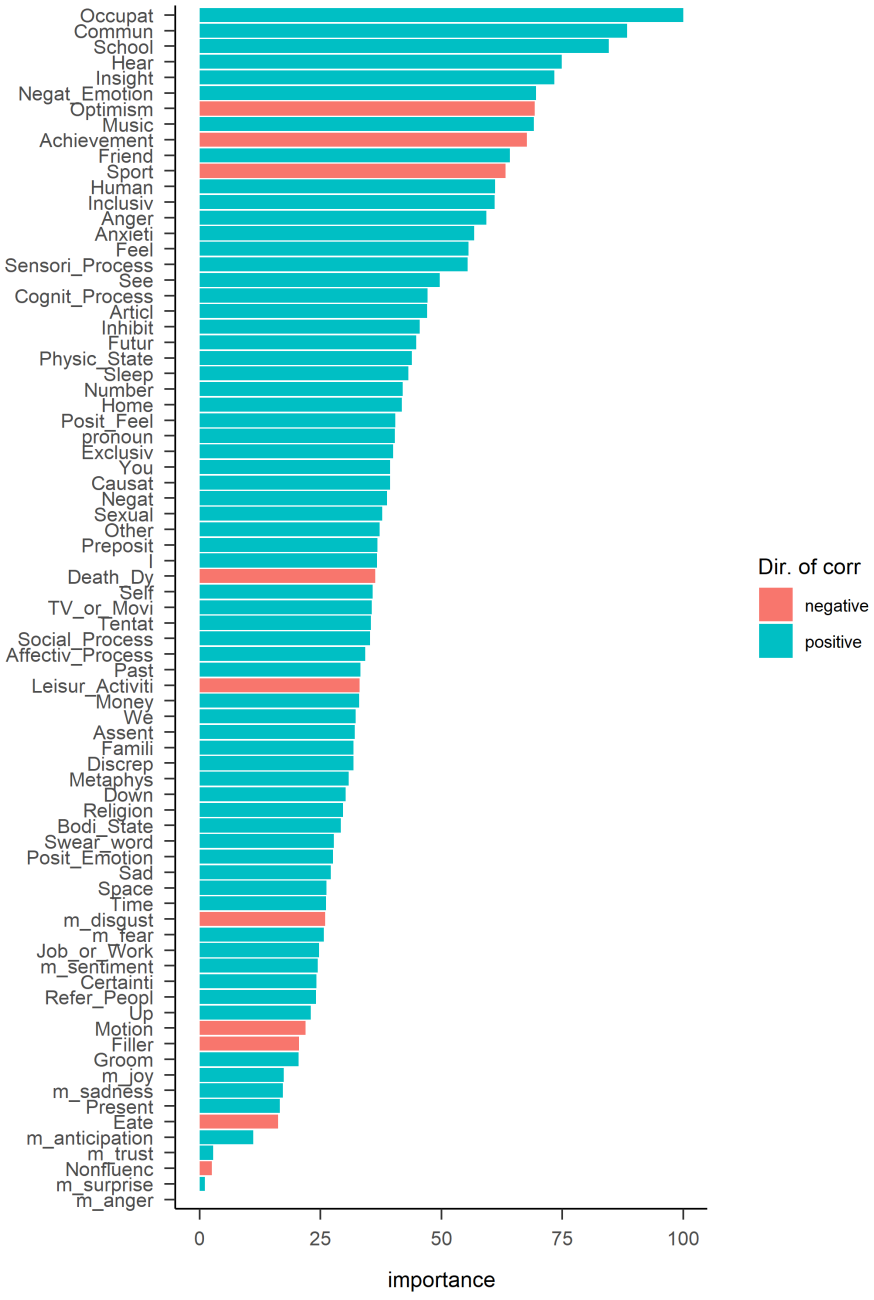

Figure 8. Importance Scores from Random Forests Predicting Openness with Dictionary Scores

Table 4
Specifications for Selected Models for Predicting Personality from Tweets

domain	Modeling approach	mtry	MNS	splitrule	R	RMSE
agreeableness	Random Forests	2	5	variance	0.19	0.58
conscientiousness	Random Forests	39	5	extratrees	0.21	0.71
honesty	Random Forests	2	5	extratrees	0.22	0.57
neuroticism	Random Forests	77	5	variance	0.35	0.84
extraversion	Random Forests	77	5	extratrees	0.22	0.77
openness	Random Forests	2	5	variance	0.23	0.61

Note. The feature set used in the selected models were the dictionary scores. mtry and MNS are hyperparameter specifications. mtry corresponds to how many predictors the algorithm samples to build each tree in the forest. MNS stands for minimum node size and corresponds to the minimum number of observation in each 'node', meaning it won't create a split in the data for fewer observations than MNS.

corresponding accuracy estimates) in Table 4.

Model Evaluation. I next evaluated the models by assessing their accuracy in predicting self-reported personality scores in the holdout data. Correlations between predicted scores derived from the trained models and observed scores for the holdout data are shown for both selected (triangles) and non-selected (circles) models in Figure 9. You can see in Figure 9 that the model selection procedure did not lead to choosing the model with the highest or nearly highest out-of-sample accuracy. Indeed, the selected model was never the highest R, though it was very close to the highest R for openness and conscientiousness. For the other four domains, it was quite a bit lower than non-selected alternatives and even among the lowest for some domains. Importantly, the accuracy estimates from non-selected models should be taken with a grain of salt; all of these estimates are subject to fluctuation and taking

the non-selected models accuracy at face value undermines the principal behind using a separate evaluation set, namely, estimating accuracy removed from a further (biasing) selection effect. Moreover, the differences in correlations are not large, and are similar to differences seen in training (differences of approx. .1 or less). Even still, these results may suggest that larger and less restrictive features sets (e.g., open-vocabulary) are better fit for some domains, perhaps especially when predicting true holdout data.

Figure 10 shows the estimates for selected models compared to predictive accuracy predicting personality from Facebook status updates from Park and colleagues' (2015) study, where it can be seen that tweets predict conscientiousness, neuroticism, and openness with moderate accuracy, and honesty, agreeableness, and extraversion with little accuracy. Moreover, tweet-based predictive accuracy tended to be lower than their Facebook-status-based counterparts, which could stem from their shorter length, how that constrains the text (e.g., increased use of slang), or social norms governing what people post on Facebook vs. twitter.

With the exceptions of agreeableness and extraversion, Big Six personality domains were at least somewhat predictable from tweets. However, it is not clear if the models are picking up on distinctive information about each domain (e.g., how conscientiousness specifically is reflected in tweets) or some more general information relevant across domains (e.g., how general positivity is reflected in followed accounts). To speak to these competing possibilities, I first examined the inter-correlations

between predicted Big Six domains, which can be seen in Table 5. Correlations between domains were generally stronger among tweet-based predictions than among (observed) self-reported scores, but the pattern of correlations were generally similar. One exception was that, among predicted scores, openness was positively correlated with neuroticism and negatively correlated with conscientiousness, whereas these domains are basically uncorrelated among self-reports. These higher intercorrelations suggest that predicted scores may indeed be picking up on more general information, rather than information specific to each domain.

Next, I more directly assessed the specificity of tweet-based predictions by regressing each observed domain on all of the predicted scores simultaneously. If personality domains are distinctly reflected in tweets, we should see a significant slope for the matching predicted score and non-significant (near-zero) slopes for the non-matching predicted scores. The results from these regression analyses are shown in Figure 11, where it is apparent that models picked up on distinctive information for openness and conscientiousness, but not so much for the others, which did show less accuracy to begin with (see Figure 10). Together, the results suggest that openness and conscientiousness are reliably and distinctly reflected in the language people use on Twitter, with the other four domains being generally more difficult to predict (agreeableness, extraversion) or more difficult to predict distinctly (honesty, neuroticism).

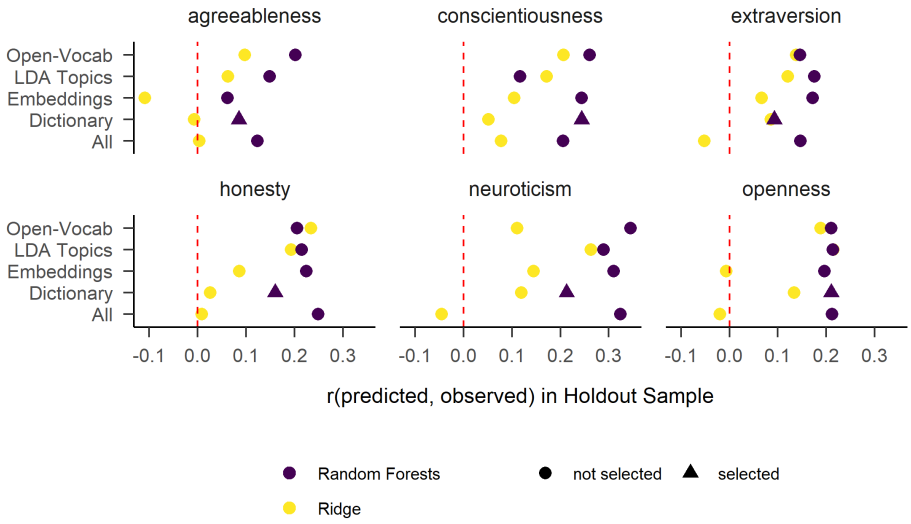

Figure 9. Out-of-sample Accuracy (R) for Selected and Non-Selected Tweet-Based Predictive Models

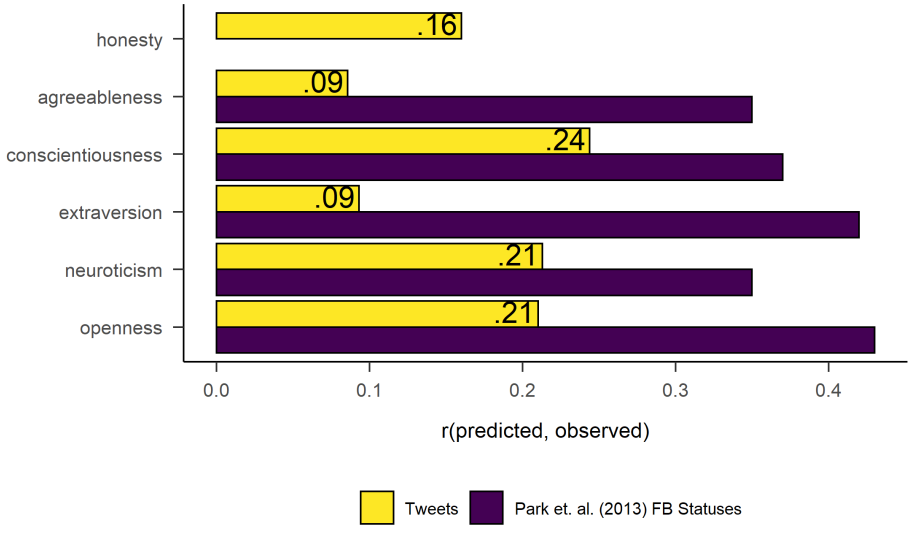

Figure 10. Out-of-sample Accuracy (R) for Tweet-Based Predictions Compared to Facebook Status Updates

Table 5

Correlations Between Tweet-Based Predictions and Observed Big Six Scores

Variable	1	2	3	4	5	6	7	8	9	10	11
1. Obs. A											
2. Obs. C	.27**										
	[.14, .39]										
3. Obs. H	.38**	.42**									
	[.26, .49]	[.31, .52]									
4. Obs. N	-.20**	-.49**	-.11								
	[-.33, -.07]	[-.59, -.39]	[-.24, .02]								
5. Obs. E	.15*	.22**	-.27**	-.38**							
	[.02, .28]	[.09, .34]	[-.39, -.14]	[-.49, -.26]							
6. Obs. O	.31**	.12	.08	-.07	.29**						
	[.19, .43]	[-.01, .25]	[-.05, .21]	[-.20, .06]	[.17, .41]						
7. Pred. A	.09	.11	.12	-.03	-.06	.03					
	[-.05, .22]	[-.02, .24]	[-.01, .25]	[-.16, .10]	[-.19, .07]	[-.11, .16]					

Table 5 continued

Variable	1	2	3	4	5	6	7	8	9	10	11
8. Pred. C	-.03	.24**	.04	-.25**	-.01	-.13	.45**				
	[-.16, .11]	[.12, .36]	[-.09, .17]	[-.37, -.12]	[-.14, .12]	[-.25, .01]	[.34, .55]				
9. Pred. H	.08	.12	.16*	-.09	-.06	.04	.71**	.50**			
	[-.05, .21]	[-.01, .25]	[.03, .29]	[-.22, .05]	[-.19, .07]	[-.09, .17]	[.64, .77]	[.39, .59]			
10. Pred. N	.07	-.17*	.02	.21**	-.04	.12	-.30**	-.68**	-.21**		
	[-.06, .20]	[-.29, -.04]	[-.11, .15]	[.08, .34]	[-.17, .09]	[-.01, .25]	[-.42, -.18]	[-.74, -.60]	[-.33, -.08]		
11. Pred. E	-.02	-.03	-.05	.01	.09	.03	.06	.29**	-.11	-.44**	
	[-.15, .11]	[-.16, .10]	[-.18, .08]	[-.12, .14]	[-.04, .22]	[-.11, .16]	[-.07, .20]	[.17, .41]	[-.24, .02]	[-.54, -.33]	
12. Pred. O	.04	-.21**	-.05	.29**	-.05	.21**	-.05	-.33**	-.14*	.27**	.33**
	[-.09, .17]	[-.33, -.08]	[-.19, .08]	[.17, .41]	[-.18, .08]	[.08, .33]	[-.18, .09]	[-.44, -.21]	[-.27, -.01]	[.14, .39]	[.21, .45]

Note. Pred. are tweet-based predictions and Obs. are (observed) self-reports. *p < .05; **p < .01; ***p < .001; 95 percent CIs are enclosed in brackets.

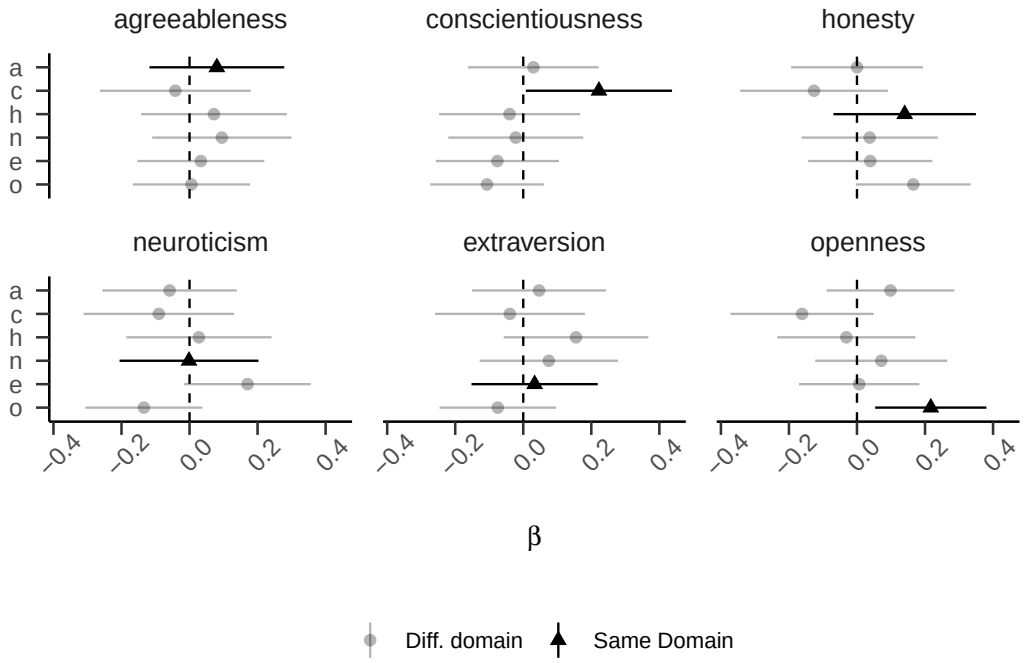

Figure 11. Results from Regressing Observed Big Six from All Tweet-Based Scores Simultaneously

Aim 1b: Does activity moderate tweet-based accuracy? I next examined the extent to which tweet-based predictive accuracy was moderated by how often individuals tweet and how many accounts they follow by regressing self-reported Big Six scores on tweet-based predicted scores (from the selected models), number of tweets (followed accounts), and the interaction term. The standardized results from these models are shown in Table 6, which shows that all of the moderator effects were small and statistically indistinguishable from zero. Tweet-based predictive accuracy does not seem to depend on how much a person tweets or how many accounts they follow, assuming they meet the minimum activity threshold(s) of our sample.

Table 6
Tweet-Based Predictive Accuracy Moderated by Activity

domain	moderator	term	estimate	SE	t	p	CI LL	CI UL
agreeableness	followeds	Intercept	-0.01	0.02	-0.54	.588	-0.04	0.03
agreeableness	followeds	predicted	0.83	0.02	45.66	< .001	0.79	0.86
agreeableness	followeds	num. followed	0.00	0.02	-0.24	.814	-0.04	0.03
agreeableness	followeds	predicted * num. followed	-0.04	0.03	-1.42	.157	-0.09	0.01
conscientiousness	followeds	Intercept	0.00	0.02	0.13	.895	-0.03	0.03
conscientiousness	followeds	predicted	0.86	0.02	54.18	< .001	0.83	0.90
conscientiousness	followeds	num. followed	-0.02	0.02	-1.16	.245	-0.07	0.02
conscientiousness	followeds	predicted * num. followed	-0.01	0.02	-0.85	.393	-0.05	0.02
honesty	followeds	Intercept	-0.02	0.02	-0.92	.357	-0.05	0.02
honesty	followeds	predicted	0.84	0.02	46.64	< .001	0.80	0.87
honesty	followeds	num. followed	-0.04	0.02	-1.76	.079	-0.08	0.00
honesty	followeds	predicted * num. followed	-0.03	0.02	-1.80	.072	-0.07	0.00

Table 6 continued

domain	moderator	term	estimate	SE	t	p	CI LL	CI UL
neuroticism	followeds	Intercept	0.00	0.02	0.25	.801	-0.03	0.04
neuroticism	followeds	predicted	0.84	0.02	49.35	< .001	0.80	0.87
neuroticism	followeds	num. followed	0.03	0.02	1.35	.179	-0.01	0.08
neuroticism	followeds	predicted * num. followed	-0.01	0.03	-0.30	.764	-0.06	0.04
extraversion	followeds	Intercept	0.00	0.02	0.13	.893	-0.03	0.03
extraversion	followeds	predicted	0.86	0.02	53.59	< .001	0.83	0.90
extraversion	followeds	num. followed	0.00	0.02	0.25	.803	-0.03	0.04
extraversion	followeds	predicted * num. followed	-0.01	0.02	-0.25	.800	-0.05	0.04
openness	followeds	Intercept	0.01	0.02	0.33	.740	-0.03	0.04
openness	followeds	predicted	0.80	0.02	42.35	< .001	0.76	0.84
openness	followeds	num. followed	0.01	0.02	0.39	.698	-0.03	0.04
openness	followeds	predicted * num. followed	-0.01	0.02	-0.40	.686	-0.04	0.03

Table 6 continued

domain	moderator	term	estimate	SE	t	p	CI LL	CI UL
agreeableness	tweets	Intercept	-0.01	0.02	-0.60	.546	-0.05	0.02
agreeableness	tweets	predicted	0.83	0.02	45.42	< .001	0.79	0.87
agreeableness	tweets	num. of tweets	0.03	0.02	1.33	.183	-0.01	0.06
agreeableness	tweets	predicted * num. of tweets	0.00	0.02	-0.13	.899	-0.05	0.04
conscientiousness	tweets	Intercept	0.01	0.02	0.41	.683	-0.02	0.04
conscientiousness	tweets	predicted	0.87	0.02	54.11	< .001	0.84	0.90
conscientiousness	tweets	num. of tweets	0.03	0.02	1.61	.107	-0.01	0.06
conscientiousness	tweets	predicted * num. of tweets	0.03	0.02	1.94	.053	0.00	0.07
honesty	tweets	Intercept	-0.02	0.02	-0.90	.367	-0.05	0.02
honesty	tweets	predicted	0.83	0.02	46.19	< .001	0.80	0.87
honesty	tweets	num. of tweets	0.02	0.02	0.91	.361	-0.02	0.05
honesty	tweets	predicted * num. of tweets	-0.01	0.03	-0.25	.802	-0.06	0.04

Table 6 continued

domain	moderator	term	estimate	SE	t	p	CI LL	CI UL
neuroticism	tweets	Intercept	0.00	0.02	-0.01	.988	-0.03	0.03
neuroticism	tweets	predicted	0.84	0.02	49.06	< .001	0.81	0.88
neuroticism	tweets	num. of tweets	-0.03	0.02	-1.16	.247	-0.07	0.02
neuroticism	tweets	predicted * num. of tweets	0.02	0.02	1.26	.209	-0.01	0.06
extraversion	tweets	Intercept	0.00	0.02	0.18	.860	-0.03	0.03
extraversion	tweets	predicted	0.87	0.02	53.77	< .001	0.83	0.90
extraversion	tweets	num. of tweets	0.04	0.02	2.41	.016	0.01	0.07
extraversion	tweets	predicted * num. of tweets	0.01	0.02	0.52	.603	-0.02	0.04
openness	tweets	Intercept	0.01	0.02	0.29	.769	-0.03	0.04
openness	tweets	predicted	0.80	0.02	42.35	< .001	0.76	0.83
openness	tweets	num. of tweets	0.00	0.02	0.03	.975	-0.04	0.04
openness	tweets	predicted * num. of tweets	-0.04	0.02	-1.65	.100	-0.08	0.01

Table 6 continued

domain	moderator	term	estimate	SE	t	p	CI LL	CI UL

Note. num. of tweets and num. of followed accounts were grand-mean-centered. CI LL and CI UL are the lower and upper bound of the 95 percent CI.

Discussion

Our findings indicate that at least some aspects of personality are reflected in the language people use on Twitter, but there is considerable heterogeneity across domains. Conscientiousness and openness could be predicted from tweets accurately and distinctly, honesty and neuroticism showed some accuracy but little distinctiveness, and agreeableness and extraversion showed little of either. Tweet-based predictive models appeared to use features that are both consistent with prior work (Mehl, Gosling, & Pennebaker, 2006; Park et al., 2015; Qiu et al., 2012) and with how the Big Six are thought to manifest in observed behavior. Indeed, inspecting Figures 3 through 8 paints quite the picture, of agreeableness corresponding to swearing angrily vs. expressing positivity and inclusivity, of conscientiousness corresponding to topics more or less suited to a workplace, of honesty corresponding to a metaphysically-tinged blend of agreeableness and conscientiousness, of neuroticism corresponding to greater negative affect, and of openness corresponding to talking about one's aesthetic and intellectual interests. Extraversion was notably difficult to interpret and had the least in common with prior work, which along with the low accuracy estimates suggest that it is more difficult to predict from what people say on Twitter. Finally, tweet-based predictive accuracy appeared to be completely unaffected by how often people tweet or how many accounts they follow. This could suggest that tweet-based predictions are relatively robust to differences in activity above the minimal threshold used here (at

least 25 tweets and followed accounts).

Tweets seem to best capture conscientiousness, openness, and neuroticism, as demonstrated by the higher accuracy in predicting them from tweets, and the relevance of the features important for predicting these domains. This may reflect some mixture of what Twitter affords to its users. Indeed, twitter offers a place for people to talk about their interests (openness), share their feelings (neuroticism), and exercise restraint or not (conscientiousness), and all of these behaviors create cues that could be easily captured with the techniques used here. Twitter may simply afford fewer opportunities to express one's level of agreeableness, honesty, and extraversion via tweets, but this doesn't seem entirely likely. A second possibility is that these domains manifest in more complex ways and require more sophisticated tools, a possibility highlighted by the slightly greater accuracy achieved with the more complex and open-ended approaches (e.g., open-vocab, topics, embeddings). Finally, it is worth considering the possibility that these domains, or at least agreeableness and honesty, are harder to predict via tweets because they are highly desirable (John & Robins, 1993), which could lower accuracy either because people tailor their tweets to convey a more positive impression or because self-reports are a poorer reflection of behavior for these more desirable domains (Vazire, 2010). This would be somewhat at odds with the high accuracy seen for openness, one of the most evaluative Big Six domains.

Interestingly, tweet-based predictions seem to capture something broader and

more generic than the Big Six given the high intercorrelations among predicted scores for different domains. The pattern of intercorrelations was generally similar to self-reports, and corresponds roughly to the higher-order Big Two (Digman, 1997; Saucier & Srivastava, 2015), a structure which has been shown to be more robust across diverse personality lexicons (Saucier et al., 2014) and theorized to correspond to core biological systems (DeYoung, 2015). Despite this, conscientiousness and openness were distinctly recoverable from tweets, which is unsurprising given the accuracy with which they can be predicted. For this reason, it is somewhat surprising that neuroticism was not distinctly recoverable from tweets. However, it seems plausible that cues for neuroticism, like negative emotion words, are highly reliable but not very distinctive, and that the algorithms were unable to differentiate between people that tweet negative affect often because they experience it often (high N) or because they're less able to inhibit the impulse to tweet about it (Low C). This is consistent with the fair amount of overlap in important features for neuroticism and the other domains and the high correlation between predicted neuroticism and other domains, especially conscientiousness ($r = .68$). This, coupled with the unexpected positive correlation between predicted neuroticism and predicted openness, suggests that tweet-based predictions of neuroticism were of questionable validity.

III. STUDY 2: PREDICTING PERSONALITY FROM FOLLOWED ACCOUNTS

The aim of Study 2 was to assess computerized personality judgments from outgoing network ties on Twitter (i.e., the accounts that users follow). In the first of Study 2's aims (*Aim 2a*), I examine the extent to which these outgoing ties, or *followed accounts*, predict self-reported personality using a cross-validated machine learning approach, testing out combinations of unsupervised and supervised machine learning techniques. As with Aim 1a, I compare models in terms of predictive accuracy and interpretability, ultimately seeking a model that can predict self-reports from followed accounts that are theoretically relevant to the construct that they are predicting. In *Aim 2b*, I examine how number of tweets and number of followed accounts relate to followed-account-based accuracy. Together, these analyses provide insight into the extent to which individuals' personalities are reflected in the accounts they follow on Twitter, and whether it depends on how they engage with the platform.

Methods

Samples & Procedure. Study 2 was conducted on all of the eligible participants from the NIMH and NSF samples that completed Big Six questionnaires and for whom we were able to successfully download followed-account lists, which resulted in a total sample of $N_{combined}$ = 1,277 participants.

Analytic Procedure. Aim 2a consisted of predicting personality from followed accounts, analogously to Aim 1a, using a procedure designed to reduce

overfitting and data leakage in estimating predictive accuracy. This consisted of a multi-stage process detailed next.

Data Partitioning. Like Study 1, we first split the final sample ($N_{combined}$ = 1,277) into a training and holdout (testing) set using the Caret package in R (Kuhn et al., 2019). The training and holdout samples consisted of roughly 80% ($n_{training}$ = 1023) and 20% ($n_{holdout}$ = 254) of the data respectively. All feature selection, data reduction, model training, estimation, and selection were determined in the training data. The final models, trained and selected within the training data, were tested on the holdout sample to get an unbiased estimate of out-of-sample accuracy.

Preparing & Pre-processing Followed Accounts. The followed accounts data were structured as a user-account matrix, where each row was an individual user, each column was a distinct account followed by some user(s) in the sample, and cells are filled in with 1's or 0's indicating whether (1) or not (0) each distinct user follows each distinct account. The total sample of 1,277 users followed 513,634 distinct accounts, which exceeded what is computationally feasible or efficient. Moreover, many of these accounts were followed by so few users as to be of little use in predictive modeling. At the extreme, uniquely followed accounts are effectively zero-variance predictors and therefore useless for most modeling and data reduction techniques. As such, the first step of our model training consisted of minimal feature selection, pruning followed accounts from the data that had few followers in our data

analogously to (Kosinski et al., 2013) approach to Facebook likes. The optimal threshold for feature selection in this data is not yet known, so we tried three values, eliminating friends followed by fewer than 3, 4, and 5 of the participants in our data; the minimum of 3 was chosen through extensive exploratory data analysis in similar data sets.

Within the training data, removing followed accounts with fewer than 3, 4, or 5 followers reduced the 513,634 distinct followed accounts to 21,436 accounts, 12,884 accounts, and 8,923 accounts respectively. Thus, the most precipitous drop occurred when going from no threshold to a threshold of 3 followers; each subsequent increase of the threshold cut the number of distinct accounts almost in half. The impact this filtering decision had on predictive accuracy is discussed below.

Modeling approaches For followed accounts, we compared four different modeling approaches: Relaxed LASSO, Random Forests, Supervised Principal Components Analysis (Supervised PCA), and two-step Principal Components Regression (PCR) with ridge regularization. Each is described in greater detail below.

Mirroring Youyou et al. (2015)'s approach to predicting personality from Facebook likes, we trained models predicting each personality variable with a variant of LASSO regression on the raw user-friend matrix, treating each distinct followed account as a predictor variable. Classic LASSO is a penalized regression model like ridge that minimizes the sum of absolute (instead of squared) beta weights (i.e., the

L1 penalty,$\lambda * \Sigma_{j=1}^{j=\beta_j} |B_j|$, where λ is a scaling parameter that determines the weight of the penalty). However, classic LASSO is known to perform poorly in contexts like these, with many noisy predictors (Meinshausen, 2007). Meinhausen (2007) developed relaxed LASSO to overcome this issue, by separating LASSO's variable/feature selection function from its regularization (shrinkage) function. Essentially, it runs two LASSO regressions in sequence; the first performs variable selection, selecting k predictors (where k is $\leq$ total number of predictors j) based on scaling hyperparameter λ, and the second performs a (LASSO) regularized regression with the remaining k variables, shrinking the parameter estimates for the reduced variable set based on scaling hyperparameter ϕ. Relaxed LASSO, like classic LASSO, can be difficult to interpret when features are correlated, which may or may not be the case with Twitter friends in our data.

The second approach was the Random Forests algorithm on the raw user-friend matrix, which was chosen due to its ability to build effective models with sparse and noisy predictors (Kuhn & Johnson, 2013). Details on Random Forests can be seen above in the methods section of Study 1.

The third approach was Supervised Principal Components Analysis (sPCA), which first conducts feature selection by eliminating features that are below some minimum (bi-variate) correlation with the outcome variable, and then performs a Principal Components Regression (PCR) with the remaining feature variables; both the minimum correlation threshold and number of components to extract are

traditionally determined via cross-validation (Bair, Hastie, Paul, & Tibshirani, 2006). Interpretation tends to be relatively straightforward, even with correlated predictors, which is why it was selected as a candidate for the present aims.

Finally, mirroring Kosinski et al. (2013), we conducted a two-step PCR with ridge regularization, first conducting an unsupervised sparse PCA on the user-friend matrix and using the resulting (orthogonal) components as predictors in a Ridge regression; we extracted the number of components that corresponds to 70% of the variance in the original (filtered) user-account matrices. The analysis section of Study 1 provides further detail on Ridge regression.

Model training and selection. All models were trained using the training data, and each model's training performance was indexed via root mean squared error (RMSE) and multiple correlation (R) from 10-fold cross-validation. Like Study 1, we aimed to maximize predictive accuracy and interpretability as much as possible.

Model evaluation. As with Study 1, I selected the candidate models based on the training data, completed an interim registration of model selection (available: https://osf.io/x7tnp/?view_only=e16eb14eec714ac285610543b84cc2e1), and then tested the selected models' accuracy using the (heldout) test data. To guard against overfitting, I selected one candidate model per outcome variable, while also testing the out-of-sample accuracy for the non-selected models as exploratory analyses, distinguishing selected from non-selected models (which can be verified in our

registration).

Aim 2b: moderator analyses. After selecting the model and evaluating it on the holdout set, we used the followed-account-based predicted personality scores for all 1277 participants in a series of OLS moderated multiple regressions. In these analyses, actual self-reported personality scores were regressed on followed-account-based scores, number of tweets (followed accounts), and their interaction, with a significant interaction indicating an effect of number of tweets (followed accounts) on followed-account-based predictive accuracy. Each of the Big Six personality domains were examined separately, resulting in 12 total moderator analyses.

Results

Aim 2a: Predictive Accuracy. Below I describe our results from model training, which models we selected for the holdout dataset, and how accurate the selected and non-selected models were in the holdout dataset. Of the 12 combinations of minimum-follower thresholds and modeling approaches, one combination failed to converge entirely: supervised PCA using followed accounts with at least 3 followers in the data. Thus, the model training and selection results below concern just the 11 other combinations.

Model Training & Selection. First, I examined the accuracy with which each combination of minimum followers' filter and modeling approach could predict

self-reported Big Six domains, focusing on the average R and RMSE for predicting the holdout-folds in the 10-fold cross-validation procedure. Figure 12 shows the average R (Panel A) and RMSE (Panel B) for each combination of minimum-followers-filter threshold (y-axes) and modeling approach (color); each dot represents the average R and RMSE for each set of hyperparameters and the bar represents the average (of average Rs or RMSEs) across hyperparameter specifications. Big Six domains are shown in separate panels, indicated with the first letter of the domain name. Note that some specifications of Relaxed LASSO are omitted from the RMSE plot because they were an order of magnitude greater and beyond the limits set on the x-axis.

Figure 12 demonstrates that personality can be predicted from followed accounts on Twitter with at least some degree of accuracy using different combinations of feature selection rules, modeling algorithms, and hyperparameter specifications. Moreover, it is apparent in Figure 12 that Random Forests achieved the greatest accuracy (highest R and lowest RMSE) and was relatively robust across hyperparameter specifications (indicated by the tightly clustered dots). Indeed, the worst hyperparameter specifications for Random Forests often outperformed the best specifications by the other algorithms.

Figure 13 shows these same metrics for the best hyperparameter specification per modeling approach and minimum-follower-filter threshold, where it shows that the best Random Forests always outperforms the best alternatives, and that the

minimum-follower-filter threshold made very little impact. Together, our quantitative criteria unequivocally support Random Forests and further suggest that, within Random Forests, the minimum-follower filter and hyperparameter specifications made little difference.

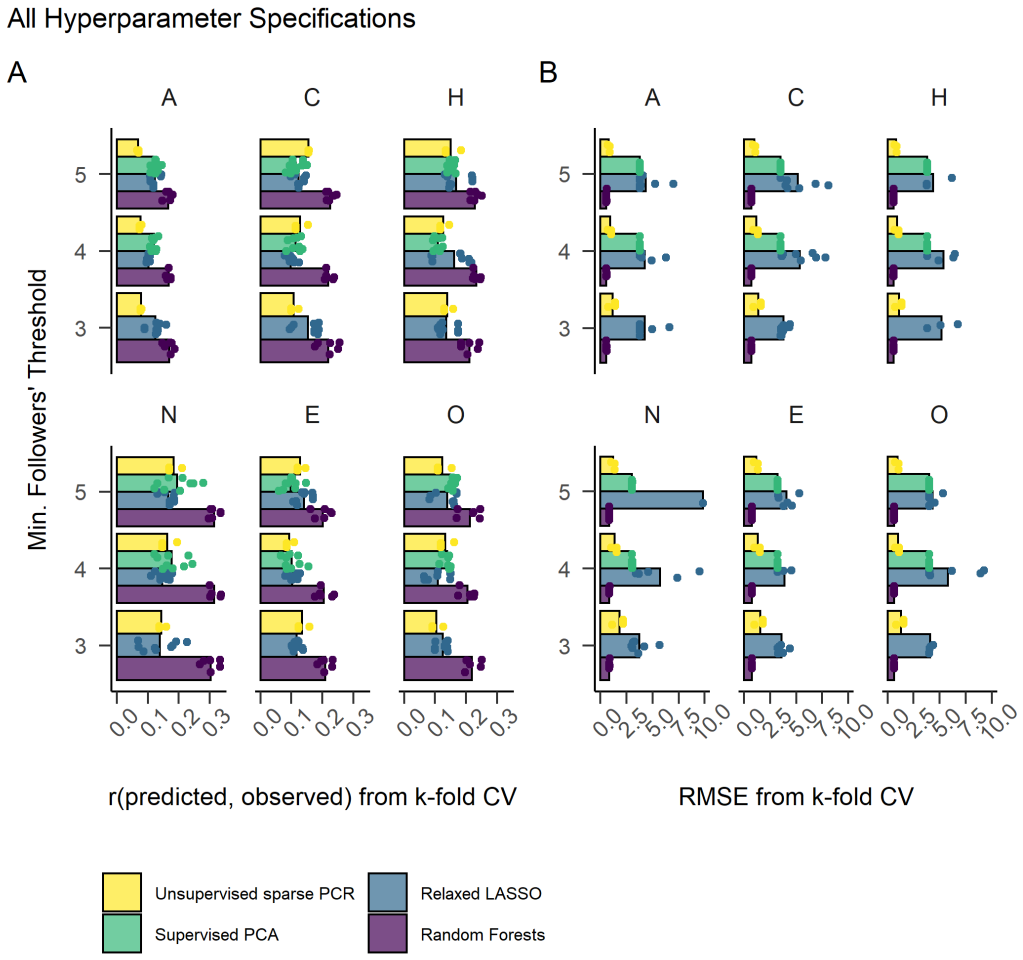

Figure 12. K-Fold CV Accuracy for Predicting Personality from Followed Accounts (All Model Specifications).

Interpretability. I next considered the interpretability of the models as a criteria for model selection. Interpretability did not strongly differentiate the trained models;

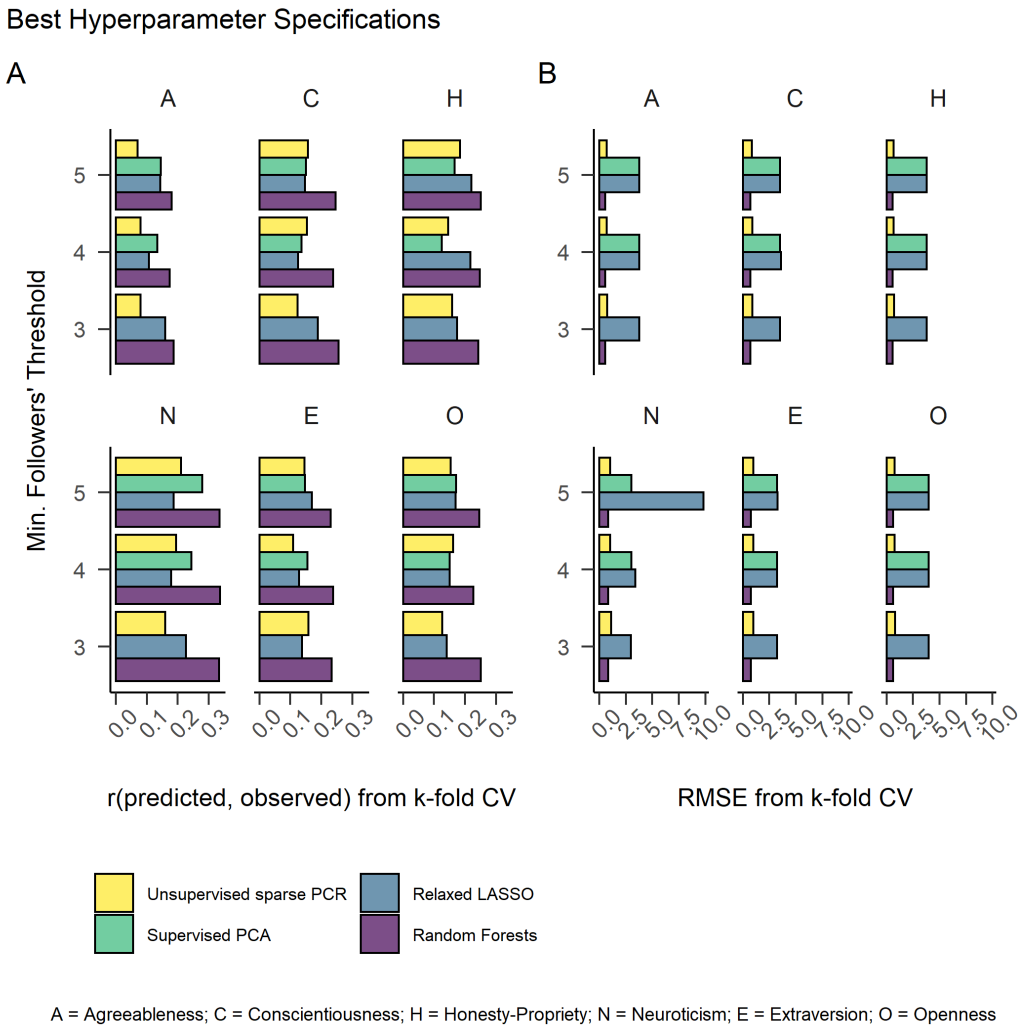

Figure 13. K-Fold CV Accuracy for Predicting Personality from Followed Accounts (Best Model Specifications).

each trained model had importance scores or model coefficients for some followed accounts that seemed theoretically relevant to the predicted domain, and some followed accounts with a less straightforward theoretical connection. Moreover, there was a great deal of overlap in which followed accounts had high model coefficients or importance scores, further highlighting the lack of differentiation according to interpretability. Domains differed with respect to interpretability, but even domains

with less interpretable models (e.g., agreeableness) had some important accounts which seemed theoretically relevant. Reporting the model coefficient and importance values for each subset of followed accounts, predicting each domain, with each modeling approach would be well-beyond the scope of this report; it would literally include hundreds of thousands of values. As such, I focus here just on the best fitting Random Forests models given their superiority in the more differentiating quantitative criteria. Tables 7 – 12 show the 15 accounts highest in importance scores from these models alongside the zero-order correlation with the corresponding personality domain.

The 15 most important accounts for predicting Agreeableness are shown in Table 7. Agreeableness was perhaps the least straightforward, but the account with the highest importance score – the founder of celebrity news and gossip site TMZ (a negative zero-order correlate) – makes some theoretical sense given the often antagonistic nature of tabloid outlets like TMZ. Otherwise, it contained a mix of brands (dove chocolate, playstation), celebrities (e.g., Nikolaj Coster-Waldau from HBO's Game of Thrones), and other accounts.

The 15 most important accounts for predicting conscientiousness are shown in Table 8; these tended to be negative zero-order correlates that relate to entertainment (video games, podcasts) and also included subversive humor accounts (e.g., "notofeminism"), potentially suggesting that lower conscientiousness is expressed by

Table 7
15 Most Important Accounts Predicting Agreeableness

followed account	importance	r
harveylevintmz	100.00	-0.12
jhony4942	88.14	-0.11
ossoff	82.89	-0.03
vizmedia	82.51	0.09
terrydpowell	81.62	-0.11
nikolajcw	79.45	0.09
jaguars	78.28	-0.02
dovechocolate	77.46	0.03
fancynews24	74.88	-0.13
pierrebouvier	74.78	-0.05
playstation	74.21	0.07
threadless	74.11	0.10
momspark	74.06	0.07
netaporter	73.42	0.09
lootably	72.96	-0.06

Note. Importance scores obtained with permutation method from the Random Forests with the highest R (and second lowest RMSE). r corresponds to zero-order correlation between following that account and self-reported agreeableness.

using twitter for entertainment (rather than work or news) and perhaps especially more subversive entertainment.

The 15 most important accounts for predicting honesty are shown in Table 9. Honesty, like agreeableness, was harder to interpret. Some highly important accounts were associated with more wholesome video games, including the official Pokemon account and the creator of the game "Stardew Valley" ("concernedape"), potentially reflecting a preference for more wholesome media content.

Table 8
15 Most Important Accounts Predicting Conscientiousness

followed account	importance	r
bts_twt	100.00	-0.14
thezonecast	60.46	-0.14
tobyfox	55.56	-0.14
wweuniverse	53.02	-0.02
travismcelroy	47.48	-0.13
notofeminism	40.71	-0.13
suethetrex	35.60	-0.06
cia	33.44	0.10
griffinmcelroy	32.77	-0.11
shitduosays	32.10	-0.05
gselevator	31.50	0.08
louisepentland	30.30	-0.11
zachanner	30.18	-0.10
amazon	29.20	0.09
usainbolt	28.75	0.10

Note. Importance scores obtained with permutation method from the Random Forests with the highest R (and lowest RMSE). r corresponds to zero-order correlation between following that account and self-reported conscientiousness.

The 15 most important accounts for predicting neuroticism are shown in Table 10; these seemed indirectly related to neuroticism and included several artists known for emotionally-evocative music (Taylor Swift, Kid Cudi, Lana Del Rey), American activist/whistle blower Chelsea Manning, and ESPN (negative zero-order correlate).

The 15 most important accounts for predicting extraversion are shown in

Table 9
15 Most Important Accounts Predicting Honesty

followed account	importance	r
fancynews24	100.00	-0.13
benlandis	74.83	0.07
hughlaurie	73.56	0.13
concernedape	71.29	0.12
badastronomer	71.12	0.12
businessinsider	69.17	-0.09
pokemon	66.01	0.10
ladygaga	65.23	-0.06
sirpatstew	64.17	0.11
thetweetofgod	60.74	0.06
kanyewest	58.99	-0.11
thesims	58.90	0.08
chaseiyons	58.75	-0.10
zachlowe_nba	56.91	-0.04
iownjd	56.88	0.08

Note. Importance scores obtained with permutation method from the Random Forests with the second highest R (and lowest RMSE). r corresponds to zero-order correlation between following that account and self-reported honesty.

Table 11; these included vinecreators (a no-longer-active stream of content from the no-longer-active platform Vine), vinecreators' successor account called twittervideo, all-female Korean pop group Loona, postsecret (a site for sharing secrets), an account for a developer that releases content for an anime-inspired rhythm-game, ESPN, Khloe Kardashian, and subversive humor account dril, all of which may suggest that extraversion vs. introversion may be reflected by following culutral content that is more mainstream (ESPN, Khloe Kardashian) vs. niche (anime, k-pop, etc.).

Table 10
15 Most Important Accounts Predicting Neuroticism

followed account	importance	r
justinmcelroy	100.00	0.17
taylornation13	87.17	0.10
thezonecast	74.03	0.15
xychelsea	72.20	0.14
espn	70.06	-0.15
colourpopco	67.26	0.13
lanadelrey	66.20	0.14
kidcudi	63.34	0.11
griffinmcelroy	63.06	0.15
nickiminaj	60.14	0.08
travismcelroy	58.58	0.15
gilliana	56.26	0.14
notofeminism	51.88	0.13
lin_manuel	50.35	0.14
vinecreators	50.33	-0.06

Note. Importance scores obtained with permutation method from the Random Forests with the highest R (and lowest RMSE). r corresponds to zero-order correlation between following that account and self-reported neuroticism.

Finally, the 15 most important accounts for predicting Openness are shown in Table 12. These were the most straightforward to interpret, with important accounts that include celebrity-scientist Neil Degrasse Tyson, comedian Patton Oswalt, the Daila LLama, musical artists (k-pop band Loona), and the online craft market ETSY, all of which seem to reflect the intellectual and artistic interests characteristic of high Openness.

Selected Models. Given their superior quantitative performance and sufficient

Table 11
15 Most Important Accounts Predicting Extraversion

followed account	importance	r
vinecreators	100.00	0.10
postsecret	93.60	-0.08
twittervideo	90.19	0.07
bbcworld	88.21	0.10
loonatheworld	79.90	-0.09
lastweektonight	79.30	0.02
id_536649400	74.52	0.11
taylornation13	71.90	-0.07
espn	58.82	0.13
rayfirefist	52.14	-0.09
translaterealdt	49.45	0.03
iamjohnoliver	49.14	0.02
askaaronlee	47.67	-0.07
kourtneykardash	47.58	0.07
dril	45.56	-0.08

Note. Importance scores obtained with permutation method from the Random Forests with the highest R (and lowest RMSE). r corresponds to zero-order correlation between following that account and self-reported extraversion.

interpretability, we selected random forests as our approach, choosing the minimum followers threshold and hyperparameter specifications based on training accuracy. Selected models are shown in Table 13. The highest R and lowest RMSE were the same model specification for conscientiousness, neuroticism, extraversion, and openness, so we selected these specifications. For agreeableness, the model with the lowest RMSE differed from the model with the highest R, though each had almost identical R and RMSE values (see Tables 14 and 15); the difference in R was greater

Table 12
15 Most Important Accounts Predicting Openness

followed account	importance	r
neiltyson	100.00	0.09
pattonoswalt	80.57	0.14
fancynews24	61.77	-0.09
dalailama	55.43	0.11
loonatheworld	54.84	-0.04
officialjaden	53.79	0.11
actuallynph	50.28	0.14
jcrasnick	48.23	-0.12
thefakeespn	44.19	-0.12
mirandalambert	44.14	-0.08
andyrichter	40.39	0.13
etsy	39.60	0.10
cashapp	39.06	0.09
zaynmalik	38.89	-0.01
gameofthrones	38.77	0.09

Note. Importance scores obtained with permutation method from the Random Forests with the highest R (and lowest RMSE). r corresponds to zero-order correlation between following that account and self-reported Openness.

than the difference in RMSE so we selected the model with the highest R. The same was true for Honesty, where the model with the highest R differed from the model with the lowest RMSE, but the difference in each (R and RMSE) was negligible. We thus selected the model with the lowest RMSE since its minimum-follower filter (4) and its hyperparameters were similar to selected models for neuroticism and extraversion. It's worth noting that we found moderate accuracy in the training set for all six domains, but it was lowest for agreeableness, highest for neuroticism, and

Table 13
Table of Selected Followed-Account-Based Models, their Specifications, and their Training Accuracy

domain	Modeling approach	Filter	mtry	MNS	r	RMSE
agreeableness	Random Forests	3	207	5	0.19	0.58
conscientiousness	Random Forests	3	207	5	0.26	0.70
honesty	Random Forests	4	160	5	0.25	0.56
neuroticism	Random Forests	4	160	5	0.34	0.85
extraversion	Random Forests	4	160	5	0.24	0.77
openness	Random Forests	3	207	5	0.25	0.61

Note. Filter refers to the minimum followers threshold used to filter out followed accounts. mtry and MNS are hyperparameter specifications. mtry corresponds to how many predictors the algorithm samples to build each tree in the forest. MNS stands for minimum node size and corresponds to the minimum number of observation in each 'node', meaning it won't create a split in the data for fewer observations than MNS.

roughly the same for the other four domains.

Model Evaluation. I next evaluated the models by assessing their accuracy in predicting self-reported personality scores in the holdout data. Correlations between predicted scores derived from the trained models and observed scores for the holdout data are shown for selected (triangles) and non-selected (circles) in Figure 14. You can see in Figure 14 that the model selection procedure tended to lead to choosing the model with the highest or nearly highest out-of-sample accuracy. Importantly, the accuracy estimates from non-selected models should be taken with a grain of salt; all of these estimates are subject to fluctuation and taking the non-selected models accuracy at face value undermines the principal behind using a separate evaluation set, namely, estimating accuracy removed from a further (biasing)

Table 14
Model Specifications with Highest R Predicting Personality from Followed Accounts

domain	Modeling approach	Filter	mtry	MNS	r	RMSE
agreeableness	Random Forests	3	207	5	0.19	0.58
conscientiousness	Random Forests	3	207	5	0.26	0.70
honesty	Random Forests	5	133	5	0.25	0.56
neuroticism	Random Forests	4	160	5	0.34	0.85
extraversion	Random Forests	4	160	5	0.24	0.77
openness	Random Forests	3	207	5	0.25	0.61

Note. Filter refers to the minimum followers threshold used to filter out followed accounts. mtry and MNS are hyperparameter specifications. mtry corresponds to how many predictors the algorithm samples to build each tree in the forest. MNS stands for minimum node size and corresponds to the minimum number of observation in each 'node', meaning it won't create a split in the data for fewer observations than MNS.

Table 15
Model Specifications with Lowest RMSE Predicting Personality from Followed Accounts

domain	Modeling approach	Filter	mtry	MNS	r	RMSE
agreeableness	Random Forests	5	2	5	0.16	0.58
conscientiousness	Random Forests	3	207	5	0.26	0.70
honesty	Random Forests	4	160	5	0.25	0.56
neuroticism	Random Forests	4	160	5	0.34	0.85
extraversion	Random Forests	4	160	5	0.24	0.77
openness	Random Forests	3	207	5	0.25	0.61

Note. Filter refers to the minimum followers threshold used to filter out followed accounts. mtry and MNS are hyperparameter specifications. mtry corresponds to how many predictors the algorithm samples to build each tree in the forest. MNS stands for minimum node size and corresponds to the minimum number of observation in each 'node', meaning it won't create a split in the data for fewer observations than MNS.

selection effect. Figure 15 shows the estimates for selected models compared to predictive accuracy predicting personality from Facebook-like ties from Kosinski and colleagues'(2013) study. As seen in Figure 15, followed accounts predict openness with considerable accuracy, neuroticism, extraversion, and honesty with moderate accuracy, and conscientiousness and agreeableness with little accuracy. Followed accounts predict openness and neuroticism about as well as Facebook likes, they predict extraversion with just slightly less accuracy than Facebook likes, and agreeableness and conscientiousness with considerably less accuracy than Facebook likes.

The accuracy achieved by the models in predicting each of the Big Six suggests that, with the possible exceptions of conscientiousness and agreeableness, personality is reflected in the accounts people choose to follow. However, it is an open question whether the models are picking up on distinctive information about each domain (e.g., how openness in particular is reflected in followed accounts) or some more general information relevant across domains (e.g., how general positivity is reflected in followed accounts). Mirroring Study 1, I assessed these possibilities first by examining the intercorrelations among all followed-account-based predicted and observed scores, which are shown in Table 16. As with tweet-based predictions, followed-account-based predictions were more strongly intercorrelated than (observed) self-reported scores, though the difference was less pronounced than with tweet-based predictions. Unlike with tweet-based predicted scores, the pattern of correlations among followed-account-based predicted scores looked quite different than

intercorrelations among observed scale scores, with predicted conscientiousness, for example, showing virtually no correlation with predicted agreeableness and honesty-propriety. Likewise, predicted neuroticism correlated positively with predicted openness, agreeableness, and honesty, which are either uncorrelated or negatively correlated among observed scores. The structure of followed-account-based predicted scores was thus less differentiated, like tweet-based predictions, but also showed a relatively distinct pattern of intercorrelations, unlike tweet-based scores.

The high intercorrelations between followed-account-based predicted scores could suggest that followed-accounts are not differentiating between cues for different Big Six domains, a possibility we examine more directly by regressing each observed domain on all of the predicted scores simultaneously. If personality domains are distinctly reflected in followed accounts, we should see a significant slope for the matching predicted score and non-significant (near-zero) slopes for the non-matching predicted scores. These results are shown in Figure 16, where it is apparent that models picked up on distinctive information for all of the Big Six except for agreeableness and conscientiousness, for which there was only a small degree of accuracy to begin with (see Figure 15). This suggests that followed accounts distinctly reflect specific personality domains when they achieve any appreciable accuracy.

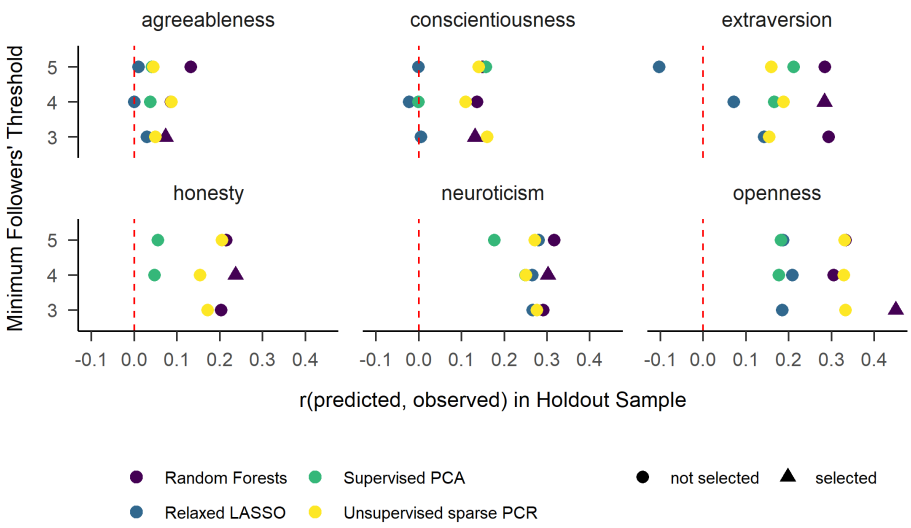

Figure 14. Out-of-sample Accuracy (R) for Selected and Non-Selected Followed-Account-Based Predictive Models

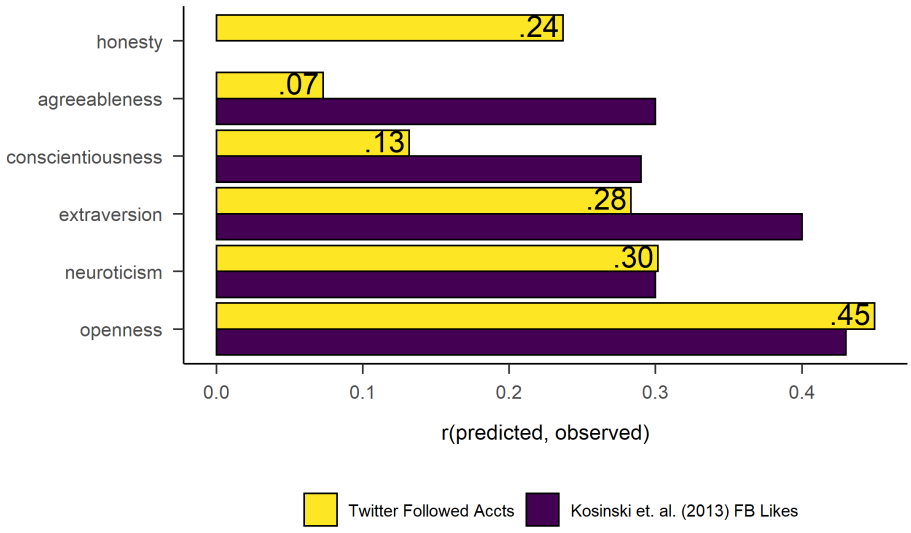

Figure 15. Out-of-sample Accuracy (R) of Followed-Account-Based Predictions Compared to Facebook Likes

Table 16

Correlations Between Followed-Account-Based Predictions and Observed Big Six Scores

Variable	1	2	3	4	5	6	7	8	9	10	11
1. Obs. A											
2. Obs. C	.26**										
	[.14, .37]										
3. Obs. H	.42**	.42**									
	[.31, .51]	[.32, .52]									
4. Obs. N	-.19**	-.43**	-.09								
	[-.30, -.07]	[-.53, -.33]	[-.21, .03]								
5. Obs. E	.07	.18**	-.30**	-.41**							
	[-.05, .19]	[.05, .29]	[-.41, -.19]	[-.50, -.30]							
6. Obs. O	.23**	.14*	.08	.01	.17**						
	[.11, .35]	[.02, .26]	[-.04, .21]	[-.11, .13]	[.05, .28]						
7. Pred. A	.07	.06	.13*	.12	-.08	.08					
	[-.05, .19]	[-.06, .18]	[.01, .25]	[-.00, .24]	[-.20, .04]	[-.05, .20]					

Table 16 continued

Variable	1	2	3	4	5	6	7	8	9	10	11
8. Pred. C	-.00	.13*	.05	-.20**	.10	-.28**	-.05				
	[-.13, .12]	[.01, .25]	[-.07, .17]	[-.32, -.08]	[-.02, .22]	[-.39, -.16]	[-.17, .08]				
9. Pred. H	.09	.09	.24**	.18**	-.22**	.13*	.61**	-.04			
	[-.04, .21]	[-.03, .21]	[.12, .35]	[.06, .30]	[-.33, -.10]	[.01, .25]	[.53, .68]	[-.16, .08]			
10. Pred. N	.06	-.14*	.01	.30**	-.15*	.25**	.35**	-.66**	.39**		
	[-.06, .18]	[-.26, -.02]	[-.11, .14]	[.19, .41]	[-.26, -.02]	[.13, .36]	[.23, .45]	[-.72, -.59]	[.28, .49]		
11. Pred. E	-.05	.05	-.16**	-.23**	.28**	-.12*	-.22**	.55**	-.43**	-.62**	
	[-.17, .08]	[-.07, .17]	[-.28, -.04]	[-.34, -.11]	[.17, .39]	[-.24, -.00]	[-.34, -.10]	[.46, .63]	[-.52, -.32]	[-.69, -.54]	
12. Pred. O	.14*	-.03	.05	.11	-.01	.45**	.20**	-.18**	.21**	.28**	-.03
	[.02, .26]	[-.15, .09]	[-.07, .17]	[-.02, .23]	[-.14, .11]	[.35, .54]	[.08, .32]	[-.30, -.06]	[.09, .32]	[.16, .39]	[-.15, .09]

Note. Pred. are tweet-based predictions and Obs. are (observed) self-reports. *p < .05; **p < .01; ***p < .001; 95 percent CIs are enclosed in brackets.

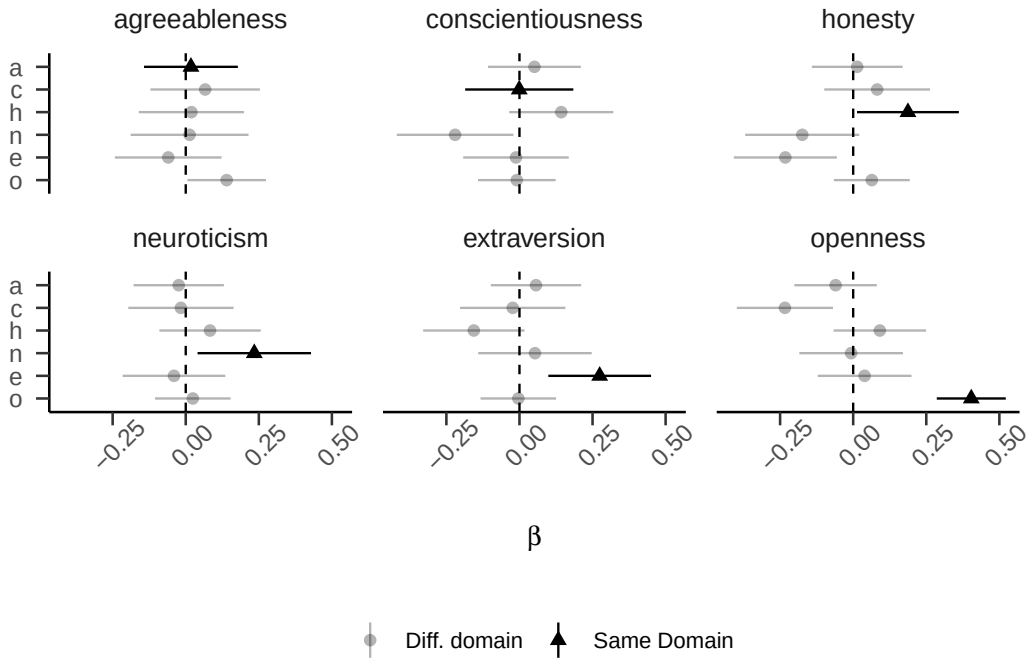

Figure 16. Results from Regressing Observed Big Six from All Followed-Account-Based Scores Simultaneously

Aim 2b: Does activity moderate followed-account-based accuracy?

I next examined the extent to which predictive accuracy was moderated by how often individuals tweet and how many accounts they follow by regressing self-reported Big Six scores on followed-account-based predicted scores, number of tweets (followed accounts), and the interaction term. The standardized results from these models are shown in Table 17, which shows that number of followed accounts moderates accuracy for agreeableness and honesty, and number of tweets moderate accuracy for agreeableness. However, these moderation effects were quite small, as seen in Figures 17 and 18, which show moderator results for agreeableness and honesty-propriety respectively. Indeed, the significant moderation in the left-hand

panel of Figure 18 looks hardly distinguishable from the non-significant moderation on the right-hand side. Thus, followed accounts are similarly accurate for twitter users across different rates of tweeting and following accounts.

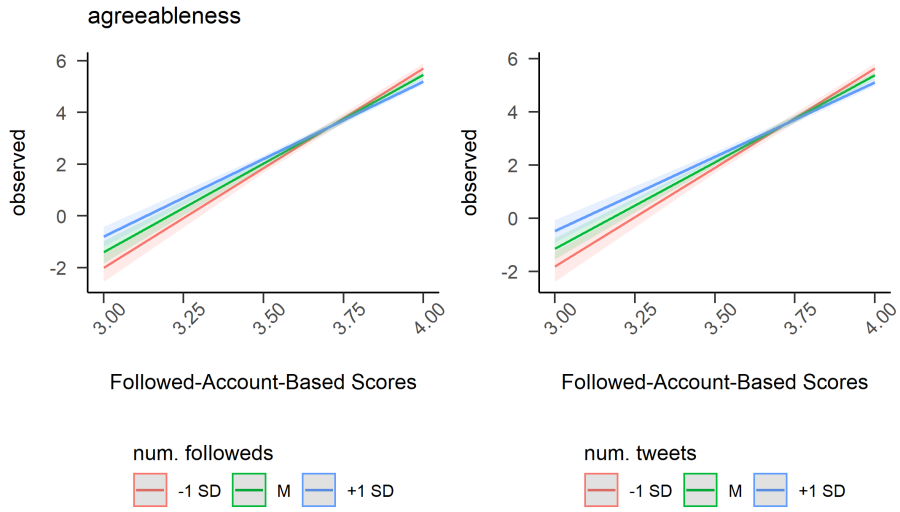

Figure 17. Followed-Account-Based Predictive Accuracy Moderated by Activity for Agreeableness

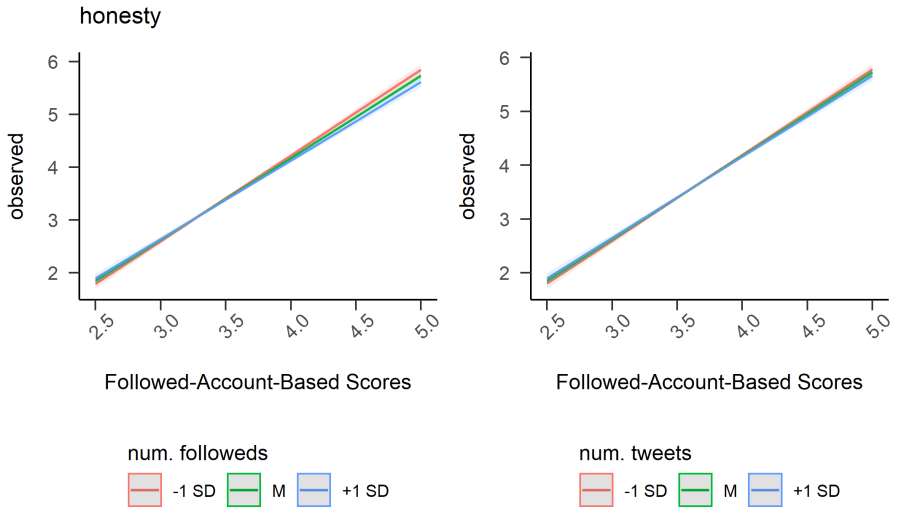

Figure 18. Followed-Account-Based Predictive Accuracy Moderated by Activity for Honesty-Propriety

Table 17
Followed-Account-Based Predictive Accuracy Moderated by Activity

domain	moderator	term	estimate	SE	t	p	CI LL	CI UL
agreeableness	followeds	Intercept	0.00	0.02	0.13	.900	-0.04	0.05
agreeableness	followeds	predicted	0.62	0.03	23.98	< .001	0.57	0.67
agreeableness	followeds	num. followed	-0.07	0.02	-3.09	.002	-0.12	-0.03
agreeableness	followeds	predicted * num. followed	-0.08	0.01	-6.32	< .001	-0.10	-0.05
conscientiousness	followeds	Intercept	0.01	0.01	0.47	.638	-0.02	0.04
conscientiousness	followeds	predicted	0.86	0.01	58.77	< .001	0.83	0.89
conscientiousness	followeds	num. followed	-0.02	0.02	-1.23	.219	-0.06	0.01
conscientiousness	followeds	predicted * num. followed	-0.02	0.02	-1.13	.259	-0.05	0.01
honesty	followeds	Intercept	0.00	0.01	0.02	.982	-0.03	0.03
honesty	followeds	predicted	0.87	0.01	58.25	< .001	0.84	0.90
honesty	followeds	num. followed	-0.05	0.02	-3.04	.002	-0.08	-0.02
honesty	followeds	predicted * num. followed	-0.04	0.02	-2.30	.022	-0.07	-0.01

Table 17 continued

domain	moderator	term	estimate	SE	t	p	CI LL	CI UL
neuroticism	followeds	Intercept	0.01	0.01	0.51	.613	-0.02	0.04
neuroticism	followeds	predicted	0.86	0.01	59.10	< .001	0.83	0.89
neuroticism	followeds	num. followed	-0.01	0.02	-0.41	.682	-0.05	0.03
neuroticism	followeds	predicted * num. followed	-0.01	0.02	-0.57	.568	-0.05	0.03
extraversion	followeds	Intercept	0.01	0.01	0.54	.592	-0.02	0.03
extraversion	followeds	predicted	0.88	0.01	64.22	< .001	0.86	0.91
extraversion	followeds	num. followed	-0.01	0.01	-1.00	.319	-0.04	0.01
extraversion	followeds	predicted * num. followed	-0.02	0.01	-1.62	.105	-0.05	0.00
openness	followeds	Intercept	0.01	0.01	0.44	.659	-0.02	0.03
openness	followeds	predicted	0.89	0.01	66.67	< .001	0.86	0.91
openness	followeds	num. followed	0.01	0.01	0.41	.680	-0.02	0.03
openness	followeds	predicted * num. followed	-0.01	0.01	-1.36	.173	-0.03	0.01

Table 17 continued

domain	moderator	term	estimate	SE	t	p	CI LL	CI UL
agreeableness	tweets	Intercept	0.01	0.02	0.35	.726	-0.04	0.05
agreeableness	tweets	predicted	0.59	0.02	24.09	< .001	0.54	0.64
agreeableness	tweets	num. of tweets	-0.05	0.02	-2.26	.024	-0.10	-0.01
agreeableness	tweets	predicted * num. of tweets	-0.08	0.02	-4.80	< .001	-0.12	-0.05
conscientiousness	tweets	Intercept	0.01	0.01	0.38	.701	-0.02	0.03
conscientiousness	tweets	predicted	0.86	0.01	58.49	< .001	0.83	0.89
conscientiousness	tweets	num. of tweets	-0.01	0.02	-0.60	.549	-0.04	0.02
conscientiousness	tweets	predicted * num. of tweets	-0.02	0.01	-1.89	.059	-0.04	0.00
honesty	tweets	Intercept	0.00	0.01	0.06	.952	-0.03	0.03
honesty	tweets	predicted	0.87	0.01	57.91	< .001	0.84	0.90
honesty	tweets	num. of tweets	-0.01	0.02	-0.61	.544	-0.04	0.02
honesty	tweets	predicted * num. of tweets	-0.02	0.02	-1.42	.156	-0.06	0.01

Table 17 continued

domain	moderator	term	estimate	SE	t	p	CI LL	CI UL
neuroticism	tweets	Intercept	0.01	0.01	0.48	.631	-0.02	0.04
neuroticism	tweets	predicted	0.86	0.01	58.76	< .001	0.83	0.89
neuroticism	tweets	num. of tweets	-0.01	0.02	-0.65	.519	-0.05	0.02
neuroticism	tweets	predicted * num. of tweets	-0.01	0.02	-0.34	.731	-0.03	0.02
extraversion	tweets	Intercept	0.01	0.01	0.57	.569	-0.02	0.03
extraversion	tweets	predicted	0.88	0.01	63.96	< .001	0.86	0.91
extraversion	tweets	num. of tweets	0.02	0.01	1.52	.129	-0.01	0.05
extraversion	tweets	predicted * num. of tweets	0.00	0.01	-0.06	.950	-0.02	0.02
openness	tweets	Intercept	0.01	0.01	0.43	.664	-0.02	0.03
openness	tweets	predicted	0.89	0.01	66.71	< .001	0.86	0.91
openness	tweets	num. of tweets	0.01	0.01	1.00	.318	-0.01	0.04
openness	tweets	predicted * num. of tweets	0.01	0.02	0.83	.405	-0.02	0.05

Table 17 continued

domain	moderator	term	estimate	SE	t	p	CI LL	CI UL

Note. num. of tweets and num. of followed accounts were grand-mean-centered. CI LL and CI UL are the lower and upper bound of the 95 percent CI.

Discussion

The results of Study 2 suggest that personality is indeed reflected in the accounts that people follow on Twitter, though there was considerable variability in the extent of accuracy across domains. Moreover, the different results appear to converge on several key findings. First, Openness is the most predictable from followed accounts. Models achieved considerable accuracy during training and evaluation (with the holdout data), the selected model appeared to use theoretically relevant followed accounts in its predictions, and the follow-up analyses demonstrated that these accounts appeared to distinctively reflect openness. Neuroticism, honesty, and extraversion were similarly, though slightly less, predictable and interpretable from followed accounts, and each appeared to be distinctly reflected in followed accounts. Agreeableness and conscientiousness were at the other extreme, with relatively poor performance in training, even poorer performance in evaluation, and with potentially the least theoretically-consistent model parameters. With the exception of agreeableness and conscientiousness, followed-account-based predictions were similarly accurate to accuracy obtained from predicting personality from Facebook likes (Kosinski et al., 2013), which is especially impressive given the slightly more conservative design of the present study - namely, the use of a holdout sample for model evaluation. Finally, followed-account-based accuracy was virtually unaffected by how much people tweet or how many accounts they follow, suggesting that this approach is relatively robust to differences in activity and use of Twitter.

Followed accounts are thus a relatively robust predictor of personality, with the notable exceptions of agreeableness and conscientiousness.

In some ways, it is unsurprising that followed accounts seem to reflect openness more than the other personality domains. Indeed, following accounts on twitter is the primary way people can curate their timeline - what they see when they log into twitter - and it thus makes sense that this appears to be most related to openness, the personality domain most centrally concerned with aesthetic and intellectual interests. Likewise, extraversion was fairly predictable, and the model appeared to achieve this via picking up on more mainstream (high extraversion) vs. niche (low extraversion) cultural interests, a finding consistent with work Park et al. (2015)'s work on predicting personality from Facebook status updates and with some of the open-vocab results from Study 1. Thus, one reason for the heterogeneity of predictive accuracy across domains could be the extent to which personality domains are expressed in interests, and agreeableness and (to a lesser extent) conscientiousness may simply have few systematic relations to the kinds of interests people can seek out on Twitter.

The intercorrelations among followed-account-based predictions were quite a bit stronger than among (observed) self-report scales, suggesting that followed-account-based predictions may be picking up on some broader, less specific personality information. However, unlike tweet-based scores, these did not appear to map as cleanly onto the Big Two (Digman, 1997; Saucier & Srivastava, 2015), which is somewhat puzzling. One possible explanation for this is that the structure could be

obscured by the seemingly poor predictions for conscientiousness and agreeableness, though these were in the same ballpark as the lowest estimates (agreeableness and extraversion) for tweet-based predictive accuracy. Another possibility is that the true structure of followed-account based predictions does not correspond to the Big Two, either reflecting deeper psychological truths (e.g., personality-relevant interests may not have the same correlation structure as personality adjectives) or for methodological reasons limited to twitter (e.g., twitter's recommendation algorithm could introduce bias and noise).

IV. STUDY 3: PERCEIVING PERSONALITY IN PROFILES

Study 3 focuses on judgments made by human perceivers from users' profiles and has three specific aims. First, I examine the extent to which people reach consensus and accuracy in their judgments of targets' personalities after viewing their (targets') Twitter profiles, providing insight into the extent to which profiles convey consistent and accurate information about target users' personalities (*Aim 3a*). Second, I examine the extent to which consensus and accuracy are affected by targets' self-presentation goals and the density of the target users' follower networks, speaking to the process of (in)accurate personality judgment on Twitter (*Aim 3b*). Third, I examine the intra- and interpersonal consequences for accuracy and idealization by examining their impact on targets' well-being and likability (*Aim 3c*). Together, these aims elucidate the social functions of personality expression and interpersonal perception online.

Methods

Samples & Procedure. Study 3 used two samples of participants. The *target* sample consisted of $n_{targets} = 100$ participants from the NSF sample that provided self-reports of their personality, self-presentation goals, and access to their twitter data. Target participants were 27.22 years old on average; race and gender breakdowns are shown in Table 18. In addition to the data collection described above, we collected additional data for our target sample. We obtained screencaps of each of

these 100 participants' Twitter profiles, which served as the stimuli for our sample of perceivers. We also downloaded each of their (targets') full follower list and their followers' followed accounts list; this data was used to calculate follower-network density.

The *perceiver* sample consisted of an initial sample of 308 participants drawn from the UO Human Subjects Pool. Data first underwent a blinded screen wherein another PhD student in the lab screened a masked dataset (i.e., where the link between targets, perceivers, and ratings were broken) for random responding, leading to the removal of 10 participants and a final sample of $n_{perceivers} = 298$. Perceiver participants were 19.67 years old on average; race and gender breakdowns are shown in Table 19. Data collection was approved by the University of Oregon Institutional Review Board (Protocol # 10122017.011) and was conducted in a manner consistent with the ethical treatment of human subjects. Perceivers were shown a random block of five profile screencaps from the target sample and instructed to rate target participants "based only on the information included in the profile" and to "give [their] best answer, even if it is just a guess." Participants rated the targets standing on the Big Five using the 10-item Big Five Inventory (Rammstedt & John, 2007) and a single item for honesty. Relevant to Aim 3c, perceivers rated the extent to which they think the target is likable. After they completed their ratings of the five targets, they were thanked and compensated for their time with course credit. The questionnaire includes several other ratings not examined here (intelligent,

Table 18
Target Race and Gender

	female	male	Not Reported	other
American Indian or Alaska Native	1	0	0	0
Asian	8	8	0	0
Black or African American	3	7	0	0
Not Reported	0	0	1	0
Other	0	2	0	0
White	25	44	0	1

Note. Demographic questions were based on NIH enrollment reporting categories.

self-esteem, trustworthy, funny, lonely, assertive, modest, arrogant, and physically attractive, perceived race, perceived gender, and perceived socio-economic status).

Measures. For this study, we measured self-reported Big Six personality domains and well-being, reports of how targets wish to be seen on Twitter, perceived Big Six personality domains, and calculated targets' follower-network density using Twitter API data.

Self-reported Big Six. Target participants completed the self-reported Big Six measure described in the overview section (prior to Study 1). To summarize, the Big Five were measured with the 60-item BFI-2 (Soto & John, 2017b) and 8 items from the Questionnaire Big Six family of measures for honesty-propriety (Thalmayer & Saucier, 2014). Internal consistency was adequate, with alphas ranging from a low of .64 for honesty-propriety and .92 for neuroticism.

Self-reported well-being. Target participants completed the single-item satisfaction with life measure of well-being (Cheung & Lucas, 2014).

Table 19
Perceiver Race and Gender

	female	male	other
American Indian or Alaska Native	4	5	0
Asian	26	18	0
Black or African American	9	5	0
Native Hawaiian or Other Pacific Islander	1	3	0
Not Reported	1	0	0
Other	7	4	1
White	137	78	1

Note. Demographic questions were based on NIH enrollment reporting categories.

Self-presentational Big Six. Target participants in the NSF sample provided self-reports of how they present themselves on Twitter. We asked participants to indicate "what impression [they] would like to make on people who see [their] Twitter profile" using the 15 item extra short BFI-2 (BFI-2-XS; Soto & John, 2017a) and three items to measure honesty-propriety. Alphas were much lower for these scales - as is typical for short measures - and ranged from 0.18 for honesty and 0.71 for neuroticism.

Perceiver-rated Big Six. Perceiver participants rated targets using the 10-item Big Five Inventory (Rammstedt & John, 2007) and a single item for honesty. Alphas ranged from 0.44 for neuroticism and 0.68 for extraversion.

Follower-network Density Targets' Follower-network density was calculated by taking each targets' network of followers (i.e., all users that follow the target), downloading those followers' followed account list, and then scoring each for density using the igraph library (Csardi & Nepusz, 2006). Each targets' score thus

represents the proportion of edges (relative to the total number of possible edges) among their follower-network.

Analyses

Aims 3a and 3b concern the extent of consensus, accuracy, idealization, and moderators of these effects in profile-based perceptions. All of these analyses will consist of a series of cross-classified random effects models (Bryk & Raudenbush, 2002). In this design, ratings are cross-classified by perceivers and targets, which are nested in blocks. We will examine consensus and accuracy for each trait separately, by conducting a series of mixed effects models (Bryk & Raudenbush, 2002). We'll start with an intercept only model from which we can estimate consensus, and subsequently add self-reports (for accuracy), self-presentation reports (for idealization), and follower-network density and its interaction with self-reports to examine whether density accuracy. Specific details, including equations, are shown in the results section as relevant.

Aim 3c concerns the extent to which accurate or idealized perceptions affect targets' self-reported well-being and perceived (i.e., perceiver-rated) likability using a technique called response surface analysis (RSA; Barranti et al., 2017). RSA consists of running a polynomial regression predicting an outcome from two predictors, their quadratic effect, and their interaction. This equation is used to define the response surface, the shape of which can be used to test several different questions about whether and how matches or mismatches between predictors relate to the outcome.

This approach is considered the most comprehensive method for examining the consequences of accuracy in interpersonal perception (see Barranti et al., 2017). Since target well-being is a single- (target-) level variable, response surfaces for well-being will be defined using single-level regressions. Since likability is a target-perceiver dyadic variable, response surfaces for likability will be defined using cross-classified mixed effects models and use multi-level RSA (Nestler, Humberg, & Schönbrodt, 2019).

RSA simultaneously estimates five parameters, each of which has a meaningful interpretation. First, *the slope of the line of congruence* (a_1) captures the extent to which matching at high values is associated with different outcomes than matching at low levels. Second, *the curvature of the line of congruence* (a_2) captures the extent to which matching at extreme values is associated with different outcomes than matching at less extreme values. Third, *the slope along the line of incongruence* (a_3) captures whether one mismatch is better or worse than the other. Fourth, *the curvature of the line of incongruence* (a_4) captures the extent to which matches or mismatches are better. Finally, Humberg, Nestler, and Back (2019) suggest testing that the first principal axis (also called the ridge) of the surface is positioned at the line of congruence by testing a_5, which provides a strict test of congruence hypotheses (i.e., that matching leads to the highest value for the outcome).

Results

I start by examining consensus, accuracy, and idealization, then examine whether density moderates accuracy, and finally examine the consequences for accuracy and idealization on well-being and perceived likability.

Consensus. Consensus was estimated using an intercept only model (per domain). At level 1, we regressed scale scores for each rating of target i by perceiever j in block k on a random intercept. Random effects for target and perceiver were included at level 2 and random effects for block was included at level 3. This is shown in Equation (1) below.

$$
\begin{aligned}
&Level1: && (1)\\
&\quad Y_{ijk} = \pi_{0ijk} + e_{ijk}\\
&Level2:\\
&\quad \pi_{0ijk} = \beta_{00k} + r_{0ik} + r_{0jk}\\
&Level3:\\
&\quad \beta_{00k} = \gamma_{000} + u_{00k}
\end{aligned}
$$

This decomposed each rating into the grand mean (γ_{00}), variance explained by the target (Var(r_{0ik}) or σ^2_{target}), variance explained by the perceiver (Var(r_{0jk}) or

$\sigma^2_{perceiver}$), and residual variance (Var(e_{ijk}) or σ^2_{resid}; Kenny, 1994).

Consensus was estimated using these baseline models, by computing the target Intraclass Correlation Coefficient (ICC_{target}). The ICC_{target} is defined as the target variance over the total variance (see Kenny, 1994) as shown in Equation (2) below:

$$ICC_{Target} = \frac{\sigma^2_{target}}{\sigma^2_{target} + \sigma^2_{perceiver} + \sigma^2_{block} + \sigma^2_{resid}} \tag{2}$$

The ICC_{target} measures the percentage of the variance in ratings explained by the target being rated or the percent agreement in ratings from different perceivers rating the same target. It is also equivalent to the expected correlation between ratings made by two randomly sampled perceivers, and is thus a straightforward metric of single-judge (rather than average) agreement. ICC_{target} and bootstrapped 95% Confidence Intervals for each Big Six domain are shown in Figure 19. You can see in Figure 19 that perceivers reach consensus about all of the Big Six after viewing targets' profiles. Consensus was substantial for openness and extraversion, moderately large for agreeableness, conscientiousness, and neuroticism, and then low but distinguishable from chance guessing for honesty. These results suggest that perceivers do agree about targets' personalities based on twitter profiles, but they do not speak to the accuracy of these judgments.

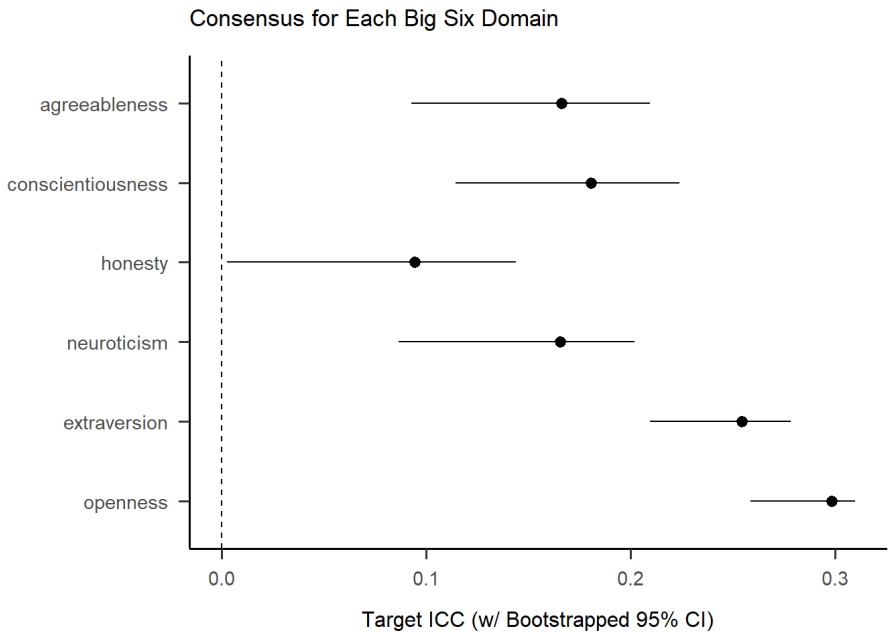

Figure 19. Plot of ICC_{target} for Each Big Six Domain

Accuracy. Accuracy was estimated by adding target self-reports (SR_{ik}) as a level-1 predictor in the mixed effects models, allowing the accuracy slope to vary randomly over targets, perceivers, and blocks as shown in Equation (3) below:

$$
\begin{aligned}
&Level1: && (3)\\
&\quad Y_{ijk} = \pi_{0ijk} + \pi_{1ijk} SR_{ik} + e_{ijk}\\
&Level2:\\
&\quad \pi_{0ijk} = \beta_{00k} + r_{0ik} + r_{0jk}\\
&\quad \pi_{1ijk} = \beta_{10k} + r_{1ik} + r_{1jk}\\
&Level3:\\
&\quad \beta_{00k} = \gamma_{000} + u_{00k}\\
&\quad \beta_{10k} = \gamma_{100} + u_{10k}
\end{aligned}
$$

The accuracy slope's intercept,γ_{100}, corresponds to the average accuracy across targets and perceivers. The fixed effects for this model are shown in Table 20 and random effects are shown in Table 21. Accuracy was relatively low across the board, though the CIs generally range from no accuracy to moderate accuracy. Only agreeableness has a CI which excludes 0, meaning it is the only domain for which accuracy is distinguishable from chance guessing. The rest were in a similar ballpark, with the exceptions of the somewhat lower estimates for conscientiousness and

Table 20
Accuracy of Profile-Based Perceptions

domain	term	γ_{100}	SE	t	df	p	CI LL	CI UL
agreeableness	accuracy	0.19	0.07	2.53	38.15	.016	0.04	0.34
conscientiousness	accuracy	0.04	0.06	0.67	26.44	.508	-0.08	0.17
extraversion	accuracy	0.13	0.08	1.69	31.18	.100	-0.02	0.27
honesty	accuracy	0.06	0.06	1.06	19.12	.300	-0.05	0.18
neuroticism	accuracy	0.09	0.05	1.82	91.81	.072	0.00	0.18
openness	accuracy	0.13	0.07	1.89	46.23	.065	-0.01	0.26

Note. Effect sizes are unstandardized. CI LL and CI UL correspond to the lower and upper limtis of the 95 percent CI respectively.

honesty. Moreover, the target- and perceiver- variance in accuracy slopes tended to be quite low, with the possible exception of target-level variance in accuracy for agreeableness ($Var(u_{1ik} = .11$), suggesting only small individual differences in accuracy across targets and perceivers. The results are thus consistent with a small degree of accuracy which is indistinguishable from chance guessing in many cases, and which varies little across targets and perceivers.

Table 21
Random Effects for Accuracy Models

domain	term	effect	estimate
agreeableness	var(u_{0jk})	intercept | perceiver:block	0.59
agreeableness	var(u_{1jk})	accuracy slope | perceiver:block	0.05
agreeableness	cov(u_{0jk}, u_{1jk})	intercept | perceiver:block, accuracy slope | perceiver:block	-0.17
agreeableness	var(u_{0ik})	intercept | target:block	1.85
agreeableness	var(u_{1ik})	accuracy slope | target:block	0.11
agreeableness	cov(u_{0ik}, u_{1ik})	intercept | target:block, accuracy slope | target:block	-0.45
agreeableness	var(u_{00k})	intercept | block	0.00
agreeableness	var(u_{10k})	accuracy slope | block	0.00
agreeableness	cov(u_{00k}, u_{10k})	intercept | block, accuracy slope | block	0.00
agreeableness	var(e_{ijk})	residual	0.47
coscientiousness	var(u_{0jk})	intercept | perceiver:block	0.11
coscientiousness	var(u_{1jk})	accuracy slope | perceiver:block	0.01

Table 21 continued

domain	term	effect	estimate
coscientiousness	cov(u_{0jk}, u_{1jk})	intercept | perceiver:block, accuracy slope | perceiver:block	-0.04
coscientiousness	var(u_{0ik})	intercept | target:block	0.20
coscientiousness	var(u_{1ik})	accuracy slope | target:block	0.00
coscientiousness	cov(u_{0ik}, u_{1ik})	intercept | target:block, accuracy slope | target:block	-0.01
coscientiousness	var(u_{00k})	intercept | block	0.05
coscientiousness	var(u_{10k})	accuracy slope | block	0.01
coscientiousness	cov(u_{00k}, u_{10k})	intercept | block, accuracy slope | block	-0.02
coscientiousness	var(e_{ijk})	residual	0.51
honest	var(u_{0jk})	intercept | perceiver:block	1.05
honest	var(u_{1jk})	accuracy slope | perceiver:block	0.07
honest	cov(u_{0jk}, u_{1jk})	intercept | perceiver:block, accuracy slope | perceiver:block	-0.26
honest	var(u_{0ik})	intercept | target:block	0.21

Table 21 continued

domain	term	effect	estimate
honest	var(u_{1ik})	accuracy slope | target:block	0.01
honest	cov(u_{0ik}, u_{1ik})	intercept | target:block, accuracy slope | target:block	-0.05
honest	var(u_{00k})	intercept | block	0.11
honest	var(u_{10k})	accuracy slope | block	0.01
honest	cov(u_{00k}, u_{10k})	intercept | block, accuracy slope | block	-0.04
honest	var(e_{ijk})	residual	0.55
neuroticism	var(u_{0jk})	intercept | perceiver:block	0.02
neuroticism	var(u_{1jk})	accuracy slope | perceiver:block	0.00
neuroticism	cov(u_{0jk}, u_{1jk})	intercept | perceiver:block, accuracy slope | perceiver:block	0.00
neuroticism	var(u_{0ik})	intercept | target:block	0.11
neuroticism	var(u_{1ik})	accuracy slope | target:block	0.00
neuroticism	cov(u_{0ik}, u_{1ik})	intercept | target:block, accuracy slope | target:block	0.00

Table 21 continued

domain	term	effect	estimate
neuroticism	var(u_{00k})	intercept | block	0.00
neuroticism	var(u_{10k})	accuracy slope | block	0.00
neuroticism	cov(u_{00k}, u_{10k})	intercept | block, accuracy slope | block	0.00
neuroticism	var(e_{ijk})	residual	0.45
extraversion	var(u_{0jk})	intercept | perceiver:block	0.45
extraversion	var(u_{1jk})	accuracy slope | perceiver:block	0.02
extraversion	cov(u_{0jk}, u_{1jk})	intercept | perceiver:block, accuracy slope | perceiver:block	-0.10
extraversion	var(u_{0ik})	intercept | target:block	0.15
extraversion	var(u_{1ik})	accuracy slope | target:block	0.00
extraversion	cov(u_{0ik}, u_{1ik})	intercept | target:block, accuracy slope | target:block	0.02
extraversion	var(u_{00k})	intercept | block	0.20
extraversion	var(u_{10k})	accuracy slope | block	0.01

Table 21 continued

domain	term	effect	estimate
extraversion	cov(u_{00k}, u_{10k})	intercept | block, accuracy slope | block	-0.05
extraversion	var(e_{ijk})	residual	0.71
openness	var(u_{0jk})	intercept | perceiver:block	0.02
openness	var(u_{1jk})	accuracy slope | perceiver:block	0.00
openness	cov(u_{0jk}, u_{1jk})	intercept | perceiver:block, accuracy slope | perceiver:block	0.01
openness	var(u_{0ik})	intercept | target:block	0.48
openness	var(u_{1ik})	accuracy slope | target:block	0.02
openness	cov(u_{0ik}, u_{1ik})	intercept | target:block, accuracy slope | target:block	-0.09
openness	var(u_{00k})	intercept | block	0.00
openness	var(u_{10k})	accuracy slope | block	0.00
openness	cov(u_{00k}, u_{10k})	intercept | block, accuracy slope | block	0.00
openness	var(e_{ijk})	residual	0.49

Table 21 continued

domain	term	effect	estimate

Note. perceiver:block refers to perceivers nested in blocks; target:block refers to targets nested in blocks.

Accuracy vs. Idealization. I next examined the extent to which targets' self-presentation goals affect the accuracy of twitter-profile-based perceptions. To do so, I added target self-presentation (SP_{ik}; i.e., how they wish they'd be seen on twitter) to the model, allowing its effect to vary randomly across targets, perceivers, and block, as shown in Equation (4) shown below:

$$
\begin{aligned}
&Level1: && (4)\\
&\quad Y_{ijk} = \pi_{0ijk} + \pi_{1ijk} SR_{ik} + \beta_{2ijk} SP_{ik} + e_{ijk}\\
&Level2:\\
&\quad \pi_{0ijk} = \beta_{00k} + r_{0ik} + r_{0jk}\\
&\quad \pi_{1ijk} = \beta_{10k} + r_{1ik} + r_{1jk}\\
&\quad \pi_{2ijk} = \beta_{20k} + r_{2ik} + r_{2jk}\\
&Level3:\\
&\quad \beta_{00k} = \gamma_{000} + u_{00k}\\
&\quad \beta_{10k} = \gamma_{100} + u_{10k}\\
&\quad \beta_{20k} = \gamma_{200} + u_{20k}
\end{aligned}
$$

Evidence for self-idealization corresponds to the magnitude of the self-presentation slope,$\gamma_{\ 200}$, analogous to (Back et al., 2010). If profiles communicate how people are, not how they wish to be seen, then adding self-presentation to the

model should result in virtually no change to the accuracy slope (γ_{100}) and near-zero estimate for the self-presentation slope (γ_{200}). At the other extreme, if profiles communicate how people wish to be seen, we should see the accuracy slope reduce to near zero and the self-presentation slope to be greater than zero. The results of these models can be seen in Figure 20, which shows accuracy (circles) and idealization (triangles) for each of the Big Six. Table 22 shows the random effects around these estimates. Although most of these slopes did not cross the threshold for significance, perceptions were more closer to targets' ideal personality for conscientiousness and honesty, similarly influenced by both real and ideal personality for agreeableness and extraversion, and more influenced by targets' real personality for neuroticism and openness. Random effects were generally small, suggesting small systematic variability in these effects across targets and perceivers. The results thus suggest profile-based perceptions are influenced by both what targets say they're like and how they'd ideally be seen on Twitter, with the relative contribution of each differing across domains.

Accuracy X Density. I examined the extent to which the density of targets' follower network affects accuracy by including density and the interaction between density (d_{ik}) and self-reported personality domains as predictors in a mixed effects model, creating Equation (5).

Table 22
Random Effects for Accuracy vs. Idealization Models

domain	term	effect	estimate
agreeableness	var(u_{0jk})	intercept | perceiver:block	0.65
agreeableness	var(u_{1jk})	accuracy slope | perceiver:block	0.08
agreeableness	var(u_{2jk})	idealization slope | perceiver:block	0.04
agreeableness	cov(u_{0jk}, u_{1jk})	intercept | perceiver:block, accuracy slope | perceiver:block	-0.15
agreeableness	cov(u_{0jk}, u_{2jk})	intercept | perceiver:block, idealization slope | perceiver:block	-0.04
agreeableness	cov(u_{1jk}, u_{2jk})	accuracy slope | perceiver:block, idealization slope | perceiver:block	-0.03
agreeableness	var(u_{0ik})	intercept | target:block	2.02
agreeableness	var(u_{1ik})	accuracy slope | target:block	0.09
agreeableness	var(u_{2ik})	idealization slope | target:block	0.01
agreeableness	cov(u_{0ik}, u_{1ik})	intercept | target:block, accuracy slope | target:block	-0.39
agreeableness	cov(u_{0ik}, u_{2ik})	intercept | target:block, idealization slope | target:block	-0.11
agreeableness	cov(u_{1ik}, u_{2ik})	accuracy slope | target:block, idealization slope | target:block	0.01

Table 22 continued

domain	term	effect	estimate
agreeableness	var(u_{00k})	intercept | block	0.06
agreeableness	var(u_{10k})	accuracy slope | block	0.08
agreeableness	var(u_{20k})	idealization slope | block	0.04
agreeableness	cov(u_{00k}, u_{10k})	intercept | block, accuracy slope | block	-0.07
agreeableness	cov(u_{00k}, u_{20k})	intercept | block, idealization slope | block	0.05
agreeableness	cov(u_{10k}, u_{20k})	accuracy slope | block, idealization slope | block	-0.06
agreeableness	var(e_{ijk})	residual	0.46
coscientiousness	var(u_{0jk})	intercept | perceiver:block	0.05
coscientiousness	var(u_{1jk})	accuracy slope | perceiver:block	0.03
coscientiousness	var(u_{2jk})	idealization slope | perceiver:block	0.01
coscientiousness	cov(u_{0jk}, u_{1jk})	intercept | perceiver:block, accuracy slope | perceiver:block	-0.04
coscientiousness	cov(u_{0jk}, u_{2jk})	intercept | perceiver:block, idealization slope | perceiver:block	0.02

Table 22 continued

domain	term	effect	estimate
coscientiousness	$\text{cov}(u_{1jk}, u_{2jk})$	accuracy slope | perceiver:block, idealization slope | perceiver:block	-0.02
coscientiousness	$\text{var}(u_{0ik})$	intercept | target:block	0.50
coscientiousness	$\text{var}(u_{1ik})$	accuracy slope | target:block	0.00
coscientiousness	$\text{var}(u_{2ik})$	idealization slope | target:block	0.03
coscientiousness	$\text{cov}(u_{0ik}, u_{1ik})$	intercept | target:block, accuracy slope | target:block	0.00
coscientiousness	$\text{cov}(u_{0ik}, u_{2ik})$	intercept | target:block, idealization slope | target:block	-0.11
coscientiousness	$\text{cov}(u_{1ik}, u_{2ik})$	accuracy slope | target:block, idealization slope | target:block	0.00
coscientiousness	$\text{var}(u_{00k})$	intercept | block	0.08
coscientiousness	$\text{var}(u_{10k})$	accuracy slope | block	0.00
coscientiousness	$\text{var}(u_{20k})$	idealization slope | block	0.00
coscientiousness	$\text{cov}(u_{00k}, u_{10k})$	intercept | block, accuracy slope | block	-0.02
coscientiousness	$\text{cov}(u_{00k}, u_{20k})$	intercept | block, idealization slope | block	-0.01

Table 22 continued

domain	term	effect	estimate
coscientiousness	cov(u_{10k}, u_{20k})	accuracy slope | block, idealization slope | block	0.00
coscientiousness	var(e_{ijk})	residual	0.50
honest	var(u_{0jk})	intercept | perceiver:block	0.66
honest	var(u_{1jk})	accuracy slope | perceiver:block	0.19
honest	var(u_{2jk})	idealization slope | perceiver:block	0.05
honest	cov(u_{0jk}, u_{1jk})	intercept | perceiver:block, accuracy slope | perceiver:block	-0.33
honest	cov(u_{0jk}, u_{2jk})	intercept | perceiver:block, idealization slope | perceiver:block	0.18
honest	cov(u_{1jk}, u_{2jk})	accuracy slope | perceiver:block, idealization slope | perceiver:block	-0.10
honest	var(u_{0ik})	intercept | target:block	0.72
honest	var(u_{1ik})	accuracy slope | target:block	0.01
honest	var(u_{2ik})	idealization slope | target:block	0.01
honest	cov(u_{0ik}, u_{1ik})	intercept | target:block, accuracy slope | target:block	-0.09

Table 22 continued

domain	term	effect	estimate
honest	$\text{cov}(u_{0ik}, u_{2ik})$	intercept | target:block, idealization slope | target:block	-0.10
honest	$\text{cov}(u_{1ik}, u_{2ik})$	accuracy slope | target:block, idealization slope | target:block	0.01
honest	$\text{var}(u_{00k})$	intercept | block	0.03
honest	$\text{var}(u_{10k})$	accuracy slope | block	0.06
honest	$\text{var}(u_{20k})$	idealization slope | block	0.09
honest	$\text{cov}(u_{00k}, u_{10k})$	intercept | block, accuracy slope | block	0.04
honest	$\text{cov}(u_{00k}, u_{20k})$	intercept | block, idealization slope | block	-0.05
honest	$\text{cov}(u_{10k}, u_{20k})$	accuracy slope | block, idealization slope | block	-0.07
honest	$\text{var}(e_{ijk})$	residual	0.53
neuroticism	$\text{var}(u_{0jk})$	intercept | perceiver:block	0.06
neuroticism	$\text{var}(u_{1jk})$	accuracy slope | perceiver:block	0.00
neuroticism	$\text{var}(u_{2jk})$	idealization slope | perceiver:block	0.01

Table 22 continued

domain	term	effect	estimate
neuroticism	cov(u_{0jk}, u_{1jk})	intercept | perceiver:block, accuracy slope | perceiver:block	0.00
neuroticism	cov(u_{0jk}, u_{2jk})	intercept | perceiver:block, idealization slope | perceiver:block	-0.02
neuroticism	cov(u_{1jk}, u_{2jk})	accuracy slope | perceiver:block, idealization slope | perceiver:block	0.00
neuroticism	var(u_{0ik})	intercept | target:block	0.25
neuroticism	var(u_{1ik})	accuracy slope | target:block	0.01
neuroticism	var(u_{2ik})	idealization slope | target:block	0.03
neuroticism	cov(u_{0ik}, u_{1ik})	intercept | target:block, accuracy slope | target:block	-0.03
neuroticism	cov(u_{0ik}, u_{2ik})	intercept | target:block, idealization slope | target:block	-0.03
neuroticism	cov(u_{1ik}, u_{2ik})	accuracy slope | target:block, idealization slope | target:block	0.00
neuroticism	var(u_{00k})	intercept | block	0.07
neuroticism	var(u_{10k})	accuracy slope | block	0.00
neuroticism	var(u_{20k})	idealization slope | block	0.01

Table 22 continued

domain	term	effect	estimate
neuroticism	$\text{cov}(u_{00k}, u_{10k})$	intercept | block, accuracy slope | block	0.00
neuroticism	$\text{cov}(u_{00k}, u_{20k})$	intercept | block, idealization slope | block	-0.03
neuroticism	$\text{cov}(u_{10k}, u_{20k})$	accuracy slope | block, idealization slope | block	0.00
neuroticism	$\text{var}(e_{ijk})$	residual	0.44
extraversion	$\text{var}(u_{0jk})$	intercept | perceiver:block	0.36
extraversion	$\text{var}(u_{1jk})$	accuracy slope | perceiver:block	0.04
extraversion	$\text{var}(u_{2jk})$	idealization slope | perceiver:block	0.03
extraversion	$\text{cov}(u_{0jk}, u_{1jk})$	intercept | perceiver:block, accuracy slope | perceiver:block	-0.09
extraversion	$\text{cov}(u_{0jk}, u_{2jk})$	intercept | perceiver:block, idealization slope | perceiver:block	0.01
extraversion	$\text{cov}(u_{1jk}, u_{2jk})$	accuracy slope | perceiver:block, idealization slope | perceiver:block	-0.02
extraversion	$\text{var}(u_{0ik})$	intercept | target:block	0.22
extraversion	$\text{var}(u_{1ik})$	accuracy slope | target:block	0.04

Table 22 continued

domain	term	effect	estimate
extraversion	var(u_{2ik})	idealization slope | target:block	0.07
extraversion	cov(u_{0ik}, u_{1ik})	intercept | target:block, accuracy slope | target:block	0.09
extraversion	cov(u_{0ik}, u_{2ik})	intercept | target:block, idealization slope | target:block	-0.12
extraversion	cov(u_{1ik}, u_{2ik})	accuracy slope | target:block, idealization slope | target:block	-0.05
extraversion	var(u_{00k})	intercept | block	0.22
extraversion	var(u_{10k})	accuracy slope | block	0.05
extraversion	var(u_{20k})	idealization slope | block	0.01
extraversion	cov(u_{00k}, u_{10k})	intercept | block, accuracy slope | block	-0.11
extraversion	cov(u_{00k}, u_{20k})	intercept | block, idealization slope | block	0.05
extraversion	cov(u_{10k}, u_{20k})	accuracy slope | block, idealization slope | block	-0.02
extraversion	var(e_{ijk})	residual	0.70
openness	var(u_{0jk})	intercept | perceiver:block	0.12

Table 22 continued

domain	term	effect	estimate
openness	var(u_{1jk})	accuracy slope | perceiver:block	0.01
openness	var(u_{2jk})	idealization slope | perceiver:block	0.01
openness	cov(u_{0jk}, u_{1jk})	intercept | perceiver:block, accuracy slope | perceiver:block	-0.04
openness	cov(u_{0jk}, u_{2jk})	intercept | perceiver:block, idealization slope | perceiver:block	0.01
openness	cov(u_{1jk}, u_{2jk})	accuracy slope | perceiver:block, idealization slope | perceiver:block	0.00
openness	var(u_{0ik})	intercept | target:block	0.19
openness	var(u_{1ik})	accuracy slope | target:block	0.03
openness	var(u_{2ik})	idealization slope | target:block	0.01
openness	cov(u_{0ik}, u_{1ik})	intercept | target:block, accuracy slope | target:block	-0.04
openness	cov(u_{0ik}, u_{2ik})	intercept | target:block, idealization slope | target:block	0.02
openness	cov(u_{1ik}, u_{2ik})	accuracy slope | target:block, idealization slope | target:block	-0.02
openness	var(u_{00k})	intercept | block	0.06

Table 22 continued

domain	term	effect	estimate
openness	var(u_{10k})	accuracy slope | block	0.00
openness	var(u_{20k})	idealization slope | block	0.00
openness	cov(u_{00k}, u_{10k})	intercept | block, accuracy slope | block	0.00
openness	cov(u_{00k}, u_{20k})	intercept | block, idealization slope | block	-0.01
openness	cov(u_{10k}, u_{20k})	accuracy slope | block, idealization slope | block	0.00
openness	var(e_{ijk})	residual	0.48

Note. perceiver:block refers to perceivers nested in blocks; target:block refers to targets nested in blocks.

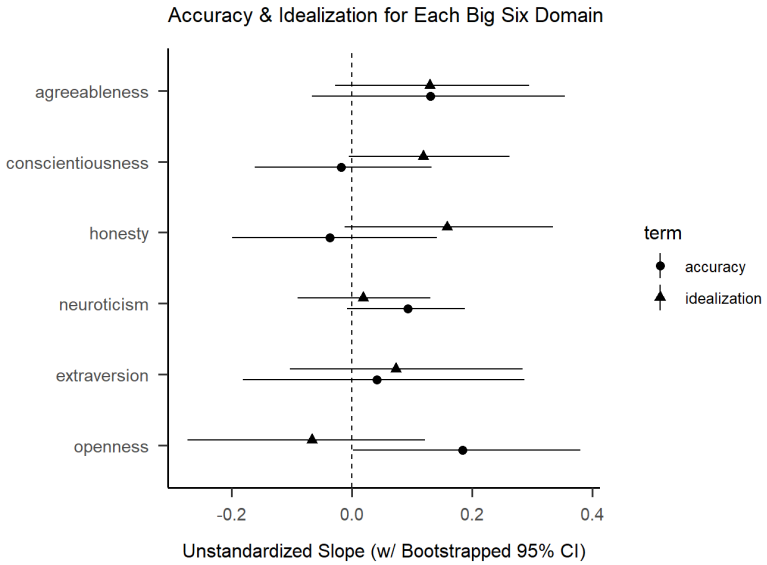

Figure 20. Accuracy vs. Idealization in Perceptions Based on Twitter Profile

$$Level1: \tag{5}$$

$$Y_{ijk} = \pi_{0ijk} + \beta_{1ijk}SR_{ik} + \beta_{2ijk}d_{ik} + \beta_{3ijk}SR_{ik} * d_{ik} + e_{ijk}$$

$$Level2:$$

$$\pi_{0ijk} = \beta_{00k} + r_{0ik} + r_{0jk}$$

$$\pi_{1ijk} = \beta_{10k} + r_{1ik} + r_{1jk}$$

$$\pi_{2ijk} = \beta_{20k} + r_{2ik} + r_{2jk}$$

$$\pi_{3ijk} = \beta_{30k} + r_{3ik} + r_{3jk}$$

$$Level3:$$

$$\beta_{00k} = \gamma_{000} + u_{00k}$$

$$\beta_{10k} = \gamma_{100} + u_{10k}$$

$$\beta_{20k} = \gamma_{200} + u_{20k}$$

$$\beta_{30k} = \gamma_{300} + u_{30k}$$

The interaction term,γ_{300}, is the critical test of the hypothesized effect of

density on accuracy. The fixed effects from these models are shown in Table 23 and random effects are shown in Table 24. The interaction term was not significant for any of the Big Six and most of the CIs ranged from large negative to large positive values. We thus found no evidence in favor of density moderating accuracy, though the CIs are large enough as to be consistent with a moderate positive, negative effect, or no effect.

Consequences for Accuracy & Idealization on Targets' Well-Being. To examine the consequences that being perceived accurately (ideally) on Twitter has on well-being, I ran a series of response surface analyses, predicting targets' self-reported well-being from their self-reported personality (self-presentation goals) and average perceiver-rated personality, separately for each Big Six domain. For idealization, we controlled for self-reports to mirror the idealization effects shown previously.

The surface parameters for accuracy are shown in Table 25 and surface plots are shown in Figure 21. Surface parameters were small and indistinguishable from zero for Agreeableness and Openness. Conscientiousness was characterized by large a1 and a4 values, though the latter's CI did overlap with zero, which together suggest well-being is higher for targets that are higher (vs. lower) in self-reported and perceived conscientiousness (a1), but that accuracy (matching self- and perceived-conscientiousness) is generally associated with lower well-being (a4).

Table 23
Results from Density X Accuracy Models

domain	effect	term	estimate	SE	t	df	p	CI LL	CI UL
agreeableness	accuracy	γ_{100}	0.18	0.07	2.53	32.68	.016	0.04	0.34
agreeableness	density	γ_{200}	1.04	10.44	0.10	7.21	.923	-17.76	24.76
agreeableness	accuracy * density	γ_{300}	14.25	15.40	0.93	5.71	.392	-18.29	47.36
conscientiousness	accuracy	γ_{100}	0.01	0.07	0.11	17.95	.917	-0.14	0.14
conscientiousness	density	γ_{200}	-3.35	9.80	-0.34	6.17	.744	-26.31	30.28
conscientiousness	accuracy * density	γ_{300}	-17.05	23.99	-0.71	3.97	.517	-87.80	47.85
extraversion	accuracy	γ_{100}	0.12	0.08	1.51	0.18	.708	-0.05	0.31
extraversion	density	γ_{200}	-5.15	26.56	-0.19	0.03	.975	-68.50	62.08
extraversion	accuracy * density	γ_{300}	2.87	36.93	0.08	0.02	.988	-90.08	92.34
honesty	accuracy	γ_{100}	0.08	0.06	1.38	15.05	.187	-0.05	0.21
honesty	density	γ_{200}	-18.29	12.22	-1.50	0.87	.400	-48.19	8.10
honesty	accuracy * density	γ_{300}	28.05	27.46	1.02	0.76	.533	-47.28	99.09

Table 23 continued

domain	effect	term	estimate	SE	t	df	p	CI LL	CI UL
neuroticism	accuracy	γ_{100}	0.08	0.05	1.63	95.40	.106	-0.03	0.18
neuroticism	density	γ_{200}	15.42	8.58	1.80	3.81	.150	-4.57	36.42
neuroticism	accuracy * density	γ_{300}	-14.99	8.92	-1.68	2.96	.193	-36.83	6.46
openness	accuracy	γ_{100}	0.12	0.06	1.95	65.08	.056	-0.01	0.25
openness	density	γ_{200}	1.17	7.34	0.16	2.16	.887	-20.56	23.92
openness	accuracy * density	γ_{300}	16.51	14.83	1.11	1.99	.382	-18.35	53.47

Note. Effect sizes are unstandardized. CI LL and CI UL correspond to the lower and upper limtis of the 95 percent CI respectively.

Table 24
Random Effects for Density X Accuracy Models

domain	term	effect	estimate
agreeableness	var(u_{0jk})	intercept | perceiver:block	0.02
agreeableness	var(u_{1jk})	accuracy | perceiver:block	0.05
agreeableness	var(u_{2jk})	density | perceiver:block	33.23
agreeableness	var(u_{3jk})	accuracy * density | perceiver:block	135.52
agreeableness	cov(u_{0jk}, u_{1jk})	intercept | perceiver:block, accuracy | perceiver:block	0.02
agreeableness	cov(u_{0jk}, u_{2jk})	intercept | perceiver:block, density | perceiver:block	0.02
agreeableness	cov(u_{0jk}, u_{3jk})	intercept | perceiver:block, accuracy * density | perceiver:block	-0.01
agreeableness	cov(u_{1jk}, u_{2jk})	accuracy | perceiver:block, density | perceiver:block	-1.01
agreeableness	cov(u_{1jk}, u_{3jk})	accuracy | perceiver:block, accuracy * density | perceiver:block	2.05
agreeableness	cov(u_{2jk}, u_{3jk})	density | perceiver:block, accuracy * density | perceiver:block	-67.10
agreeableness	var(u_{0ik})	intercept | target:block	0.05
agreeableness	var(u_{1ik})	accuracy | target:block	0.10

Table 24 continued

domain	term	effect	estimate
agreeableness	var(u_{2ik})	density | target:block	684.42
agreeableness	var(u_{3ik})	accuracy * density | target:block	2,699.84
agreeableness	cov(u_{0ik}, u_{1ik})	intercept | target:block, accuracy | target:block	-0.01
agreeableness	cov(u_{0ik}, u_{2ik})	intercept | target:block, density | target:block	-4.40
agreeableness	cov(u_{0ik}, u_{3ik})	intercept | target:block, accuracy * density | target:block	4.03
agreeableness	cov(u_{1ik}, u_{2ik})	accuracy | target:block, density | target:block	6.11
agreeableness	cov(u_{1ik}, u_{3ik})	accuracy | target:block, accuracy * density | target:block	-15.67
agreeableness	cov(u_{2ik}, u_{3ik})	density | target:block, accuracy * density | target:block	-1,200.52
agreeableness	var(u_{00k})	intercept | block	0.00
agreeableness	var(e_{ijk})	residual	0.47
coscientiousness	var(u_{0jk})	intercept | perceiver:block	0.03
coscientiousness	var(u_{1jk})	accuracy | perceiver:block	0.01

Table 24 continued

domain	term	effect	estimate
coscientiousness	var(u_{2jk})	density | perceiver:block	263.06
coscientiousness	var(u_{3jk})	accuracy * density | perceiver:block	1,315.21
coscientiousness	cov(u_{0jk}, u_{1jk})	intercept | perceiver:block, accuracy | perceiver:block	0.01
coscientiousness	cov(u_{0jk}, u_{2jk})	intercept | perceiver:block, density | perceiver:block	-0.07
coscientiousness	cov(u_{0jk}, u_{3jk})	intercept | perceiver:block, accuracy * density | perceiver:block	0.96
coscientiousness	cov(u_{1jk}, u_{2jk})	accuracy | perceiver:block, density | perceiver:block	1.33
coscientiousness	cov(u_{1jk}, u_{3jk})	accuracy | perceiver:block, accuracy * density | perceiver:block	-2.60
coscientiousness	cov(u_{2jk}, u_{3jk})	density | perceiver:block, accuracy * density | perceiver:block	-582.57
coscientiousness	var(u_{0ik})	intercept | target:block	0.13
coscientiousness	var(u_{1ik})	accuracy | target:block	0.00
coscientiousness	var(u_{2ik})	density | target:block	6.85
coscientiousness	var(u_{3ik})	accuracy * density | target:block	2,172.48

Table 24 continued

domain	term	effect	estimate
coscientiousness	cov(u_{0ik}, u_{1ik})	intercept | target:block, accuracy | target:block	-0.02
coscientiousness	cov(u_{0ik}, u_{2ik})	intercept | target:block, density | target:block	0.65
coscientiousness	cov(u_{0ik}, u_{3ik})	intercept | target:block, accuracy * density | target:block	-16.47
coscientiousness	cov(u_{1ik}, u_{2ik})	accuracy | target:block, density | target:block	-0.13
coscientiousness	cov(u_{1ik}, u_{3ik})	accuracy | target:block, accuracy * density | target:block	2.67
coscientiousness	cov(u_{2ik}, u_{3ik})	density | target:block, accuracy * density | target:block	-97.10
coscientiousness	var(u_{00k})	intercept | block	0.01
coscientiousness	var(u_{10k})	accuracy | block	0.01
coscientiousness	var(u_{20k})	density | block	5.85
coscientiousness	var(u_{30k})	accuracy * density | block	23.51
coscientiousness	cov(u_{00k}, u_{10k})	intercept | block, accuracy | block	0.01
coscientiousness	cov(u_{00k}, u_{20k})	intercept | block, density | block	0.25

Table 24 continued

domain	term	effect	estimate
coscientiousness	cov(u_{00k}, u_{30k})	intercept | block, accuracy * density | block	-0.50
coscientiousness	cov(u_{10k}, u_{20k})	accuracy | block, density | block	0.24
coscientiousness	cov(u_{10k}, u_{30k})	accuracy | block, accuracy * density | block	-0.49
coscientiousness	cov(u_{20k}, u_{30k})	density | block, accuracy * density | block	-11.63
coscientiousness	var(e_{ijk})	residual	0.50
honest	var(u_{0jk})	intercept | perceiver:block	0.10
honest	var(u_{1jk})	accuracy | perceiver:block	0.07
honest	var(u_{2jk})	density | perceiver:block	6.78
honest	var(u_{3jk})	accuracy * density | perceiver:block	12.12
honest	cov(u_{0jk}, u_{1jk})	intercept | perceiver:block, accuracy | perceiver:block	0.01
honest	cov(u_{0jk}, u_{2jk})	intercept | perceiver:block, density | perceiver:block	0.83
honest	cov(u_{0jk}, u_{3jk})	intercept | perceiver:block, accuracy * density | perceiver:block	0.66

Table 24 continued

domain	term	effect	estimate
honest	$\text{cov}(u_{1jk}, u_{2jk})$	accuracy | perceiver:block, density | perceiver:block	0.10
honest	$\text{cov}(u_{1jk}, u_{3jk})$	accuracy | perceiver:block, accuracy * density | perceiver:block	0.69
honest	$\text{cov}(u_{2jk}, u_{3jk})$	density | perceiver:block, accuracy * density | perceiver:block	5.43
honest	$\text{var}(u_{0ik})$	intercept | target:block	0.02
honest	$\text{var}(u_{1ik})$	accuracy | target:block	0.03
honest	$\text{var}(u_{2ik})$	density | target:block	238.03
honest	$\text{var}(u_{3ik})$	accuracy * density | target:block	12.84
honest	$\text{cov}(u_{0ik}, u_{1ik})$	intercept | target:block, accuracy | target:block	0.01
honest	$\text{cov}(u_{0ik}, u_{2ik})$	intercept | target:block, density | target:block	2.23
honest	$\text{cov}(u_{0ik}, u_{3ik})$	intercept | target:block, accuracy * density | target:block	0.39
honest	$\text{cov}(u_{1ik}, u_{2ik})$	accuracy | target:block, density | target:block	0.65
honest	$\text{cov}(u_{1ik}, u_{3ik})$	accuracy | target:block, accuracy * density | target:block	0.16

Table 24 continued

domain	term	effect	estimate
honest	$\text{cov}(u_{2ik}, u_{3ik})$	density | target:block, accuracy * density | target:block	23.44
honest	$\text{var}(u_{00k})$	intercept | block	0.00
honest	$\text{var}(e_{ijk})$	residual	0.55
neuroticism	$\text{var}(u_{0jk})$	intercept | perceiver:block	0.04
neuroticism	$\text{var}(u_{1jk})$	accuracy | perceiver:block	0.00
neuroticism	$\text{var}(u_{2jk})$	density | perceiver:block	459.91
neuroticism	$\text{var}(u_{3jk})$	accuracy * density | perceiver:block	481.88
neuroticism	$\text{cov}(u_{0jk}, u_{1jk})$	intercept | perceiver:block, accuracy | perceiver:block	0.01
neuroticism	$\text{cov}(u_{0jk}, u_{2jk})$	intercept | perceiver:block, density | perceiver:block	-0.74
neuroticism	$\text{cov}(u_{0jk}, u_{3jk})$	intercept | perceiver:block, accuracy * density | perceiver:block	1.34
neuroticism	$\text{cov}(u_{1jk}, u_{2jk})$	accuracy | perceiver:block, density | perceiver:block	-0.12
neuroticism	$\text{cov}(u_{1jk}, u_{3jk})$	accuracy | perceiver:block, accuracy * density | perceiver:block	0.22

Table 24 continued

domain	term	effect	estimate
neuroticism	cov(u_{2jk}, u_{3jk})	density | perceiver:block, accuracy * density | perceiver:block	-465.71
neuroticism	var(u_{0ik})	intercept | target:block	0.10
neuroticism	var(u_{1ik})	accuracy | target:block	0.00
neuroticism	var(u_{2ik})	density | target:block	315.16
neuroticism	var(u_{3ik})	accuracy * density | target:block	30.96
neuroticism	cov(u_{0ik}, u_{1ik})	intercept | target:block, accuracy | target:block	0.01
neuroticism	cov(u_{0ik}, u_{2ik})	intercept | target:block, density | target:block	-5.48
neuroticism	cov(u_{0ik}, u_{3ik})	intercept | target:block, accuracy * density | target:block	1.71
neuroticism	cov(u_{1ik}, u_{2ik})	accuracy | target:block, density | target:block	-0.33
neuroticism	cov(u_{1ik}, u_{3ik})	accuracy | target:block, accuracy * density | target:block	0.10
neuroticism	cov(u_{2ik}, u_{3ik})	density | target:block, accuracy * density | target:block	-98.51
neuroticism	var(u_{00k})	intercept | block	0.01

Table 24 continued

domain	term	effect	estimate
neuroticism	var(u_{10k})	accuracy | block	0.00
neuroticism	var(u_{20k})	density | block	8.94
neuroticism	var(u_{30k})	accuracy * density | block	55.85
neuroticism	cov(u_{00k}, u_{10k})	intercept | block, accuracy | block	0.00
neuroticism	cov(u_{00k}, u_{20k})	intercept | block, density | block	-0.26
neuroticism	cov(u_{00k}, u_{30k})	intercept | block, accuracy * density | block	-0.66
neuroticism	cov(u_{10k}, u_{20k})	accuracy | block, density | block	0.04
neuroticism	cov(u_{10k}, u_{30k})	accuracy | block, accuracy * density | block	0.10
neuroticism	cov(u_{20k}, u_{30k})	density | block, accuracy * density | block	22.34
neuroticism	var(e_{ijk})	residual	0.44
extraversion	var(u_{0jk})	intercept | perceiver:block	0.05
extraversion	var(u_{1jk})	accuracy | perceiver:block	0.02

Table 24 continued

domain	term	effect	estimate
extraversion	var(u_{2jk})	density | perceiver:block	41.37
extraversion	var(u_{3jk})	accuracy * density | perceiver:block	250.47
extraversion	cov(u_{0jk}, u_{1jk})	intercept | perceiver:block, accuracy | perceiver:block	-0.03
extraversion	cov(u_{0jk}, u_{2jk})	intercept | perceiver:block, density | perceiver:block	1.22
extraversion	cov(u_{0jk}, u_{3jk})	intercept | perceiver:block, accuracy * density | perceiver:block	3.40
extraversion	cov(u_{1jk}, u_{2jk})	accuracy | perceiver:block, density | perceiver:block	-0.83
extraversion	cov(u_{1jk}, u_{3jk})	accuracy | perceiver:block, accuracy * density | perceiver:block	-2.31
extraversion	cov(u_{2jk}, u_{3jk})	density | perceiver:block, accuracy * density | perceiver:block	97.08
extraversion	var(u_{0ik})	intercept | target:block	0.27
extraversion	var(u_{1ik})	accuracy | target:block	0.00
extraversion	var(u_{2ik})	density | target:block	133.91
extraversion	var(u_{3ik})	accuracy * density | target:block	39.15

Table 24 continued

domain	term	effect	estimate
extraversion	cov(u_{0ik}, u_{1ik})	intercept | target:block, accuracy | target:block	0.03
extraversion	cov(u_{0ik}, u_{2ik})	intercept | target:block, density | target:block	5.68
extraversion	cov(u_{0ik}, u_{3ik})	intercept | target:block, accuracy * density | target:block	2.57
extraversion	cov(u_{1ik}, u_{2ik})	accuracy | target:block, density | target:block	0.56
extraversion	cov(u_{1ik}, u_{3ik})	accuracy | target:block, accuracy * density | target:block	0.26
extraversion	cov(u_{2ik}, u_{3ik})	density | target:block, accuracy * density | target:block	67.21
extraversion	var(u_{00k})	intercept | block	0.01
extraversion	var(u_{10k})	accuracy | block	0.01
extraversion	var(u_{20k})	density | block	133.77
extraversion	var(u_{30k})	accuracy * density | block	2.89
extraversion	cov(u_{00k}, u_{10k})	intercept | block, accuracy | block	-0.01
extraversion	cov(u_{00k}, u_{20k})	intercept | block, density | block	1.13

Table 24 continued

domain	term	effect	estimate
extraversion	cov(u_{00k}, u_{30k})	intercept | block, accuracy * density | block	-0.08
extraversion	cov(u_{10k}, u_{20k})	accuracy | block, density | block	-1.19
extraversion	cov(u_{10k}, u_{30k})	accuracy | block, accuracy * density | block	0.09
extraversion	cov(u_{20k}, u_{30k})	density | block, accuracy * density | block	-9.99
extraversion	var(e_{ijk})	residual	0.71
openness	var(u_{0jk})	intercept | perceiver:block	0.11
openness	var(u_{1jk})	accuracy | perceiver:block	0.00
openness	var(u_{2jk})	density | perceiver:block	242.56
openness	var(u_{3jk})	accuracy * density | perceiver:block	1,418.78
openness	cov(u_{0jk}, u_{1jk})	intercept | perceiver:block, accuracy | perceiver:block	0.02
openness	cov(u_{0jk}, u_{2jk})	intercept | perceiver:block, density | perceiver:block	1.29
openness	cov(u_{0jk}, u_{3jk})	intercept | perceiver:block, accuracy * density | perceiver:block	0.39

Table 24 continued

domain	term	effect	estimate
openness	$\text{cov}(u_{1jk}, u_{2jk})$	accuracy | perceiver:block, density | perceiver:block	0.19
openness	$\text{cov}(u_{1jk}, u_{3jk})$	accuracy | perceiver:block, accuracy * density | perceiver:block	0.06
openness	$\text{cov}(u_{2jk}, u_{3jk})$	density | perceiver:block, accuracy * density | perceiver:block	-563.33
openness	$\text{var}(u_{0ik})$	intercept | target:block	0.13
openness	$\text{var}(u_{1ik})$	accuracy | target:block	0.00
openness	$\text{var}(u_{2ik})$	density | target:block	178.43
openness	$\text{var}(u_{3ik})$	accuracy * density | target:block	140.94
openness	$\text{cov}(u_{0ik}, u_{1ik})$	intercept | target:block, accuracy | target:block	0.00
openness	$\text{cov}(u_{0ik}, u_{2ik})$	intercept | target:block, density | target:block	-4.83
openness	$\text{cov}(u_{0ik}, u_{3ik})$	intercept | target:block, accuracy * density | target:block	-4.16
openness	$\text{cov}(u_{1ik}, u_{2ik})$	accuracy | target:block, density | target:block	0.07
openness	$\text{cov}(u_{1ik}, u_{3ik})$	accuracy | target:block, accuracy * density | target:block	0.06

Table 24 continued

domain	term	effect	estimate
openness	cov(u_{2ik}, u_{3ik})	density | target:block, accuracy * density | target:block	153.16
openness	var(u_{00k})	intercept | block	0.00
openness	var(u_{10k})	accuracy | block	0.00
openness	var(u_{20k})	density | block	10.70
openness	var(u_{30k})	accuracy * density | block	145.86
openness	cov(u_{00k}, u_{10k})	intercept | block, accuracy | block	0.00
openness	cov(u_{00k}, u_{20k})	intercept | block, density | block	0.00
openness	cov(u_{00k}, u_{30k})	intercept | block, accuracy * density | block	0.00
openness	cov(u_{10k}, u_{20k})	accuracy | block, density | block	0.00
openness	cov(u_{10k}, u_{30k})	accuracy | block, accuracy * density | block	-0.01
openness	cov(u_{20k}, u_{30k})	density | block, accuracy * density | block	-39.25
openness	var(e_{ijk})	residual	0.48

Table 24 continued

domain	term	effect	estimate

Note. perceiver:block refers to perceivers nested in blocks; target:block refers to targets nested in blocks.

Honesty was characterized by a large, positive a4, suggesting that well-being is higher the more self-reports and perceptions of targets' honesty depart from one another; a5 was large and significant, suggesting that a strict (in)congruence hypothesis is not, however, met. Neuroticism had a large negative a1, a large positive a2, and a large negative a3, suggesting well-being is higher when both self-reported and perceived neuroticism are lower (rather than higher; a1), that accuracy is associated with greater well-being at the scale extremes (vs. middle; a2), and that well-being is higher when perceived neuroticism is higher than self-reported neuroticism. Together with the graph in Figure 21, it is apparent that well-being is lowest for people who are high in neuroticism but come across as low in neuroticism; virtually every other combination is similarly high in well-being. Extraversion was characterized by large, positive a1 and a3 values, suggesting that well-being is higher when both self-reported and perceived extraversion are higher (rather than lower; a1) and that well-being is higher when self-reported extraversion is greater than perceived extraversion. Together with Figure 21, these results suggest a strong main effect of self-reported extraversion, with a small benefit for being (accurately) perceived as higher in extraversion.

Turning to idealization, the surface parameters for these effects are shown in Table 26 and surface plots are shown in Figure 22, where it is apparent that these effects were generally small and indistinguishable from zero, with honesty being the major exception. Honesty was characterized by a large, positive a2 value and a large negative a4 value, suggesting that well-being is associated with idealization at more extreme values (a2) and well-being increases as idealization increases (a4).

Table 25
Surface Parameters for Accuracy & Self-Reported Well-Being RSA

	Surface Parameter	estimate	CI LL	CI UL	p
agreeableness	a1	0.26	-0.50	1.02	.499
agreeableness	a2	-0.20	-1.85	1.45	.812
agreeableness	a3	0.59	-0.28	1.47	.185
agreeableness	a4	-1.17	-2.81	0.48	.165
agreeableness	a5	-0.32	-1.77	1.13	.666
conscientiousness	a1	0.70	0.17	1.22	.010
conscientiousness	a2	-0.68	-1.58	0.22	.138
conscientiousness	a3	0.27	-0.26	0.80	.321
conscientiousness	a4	1.11	-0.06	2.27	.062
conscientiousness	a5	0.21	-0.56	0.98	.590
honesty	a1	0.48	-0.49	1.45	.332
honesty	a2	1.16	-0.22	2.54	.099
honesty	a3	-0.29	-1.28	0.69	.557
honesty	a4	2.80	0.66	4.93	.010
honesty	a5	-2.09	-3.50	-0.68	.004
neuroticism	a1	-0.75	-1.19	-0.31	.001
neuroticism	a2	1.02	0.02	2.02	.045
neuroticism	a3	-0.95	-1.49	-0.40	.001
neuroticism	a4	-0.17	-1.59	1.26	.820

Table 25 continued

	Surface Parameter	estimate	CI LL	CI UL	p
neuroticism	a5	-0.24	-1.23	0.75	.633
extraversion	a1	0.82	0.47	1.17	$< .001$
extraversion	a2	-0.04	-0.57	0.48	.871
extraversion	a3	0.85	0.40	1.31	$< .001$
extraversion	a4	-0.14	-0.97	0.68	.731
extraversion	a5	-0.20	-0.72	0.33	.460
openness	a1	0.09	-0.51	0.68	.777
openness	a2	-0.19	-0.87	0.49	.584
openness	a3	0.05	-0.68	0.78	.889
openness	a4	0.12	-1.49	1.73	.883
openness	a5	-0.37	-1.20	0.45	.376

Note. CI LL and CI UL are the lower and upper limits of the 95 percent CI.

Consequences for Accuracy & Idealization on Targets' Likability.

To examine the consequences that being perceived accurately (ideally) on Twitter has on likability, I ran a series of multi-level response surface analyses, predicting each perceiver *i's* rating of target *j's* likability from target *j's* self-reported personality (self-presentation goals) and perceiver *i's* rating of target *j's* personality, separately for each Big Six domain. For idealization, we controlled for self-reports to mirror the idealization effects shown previously.

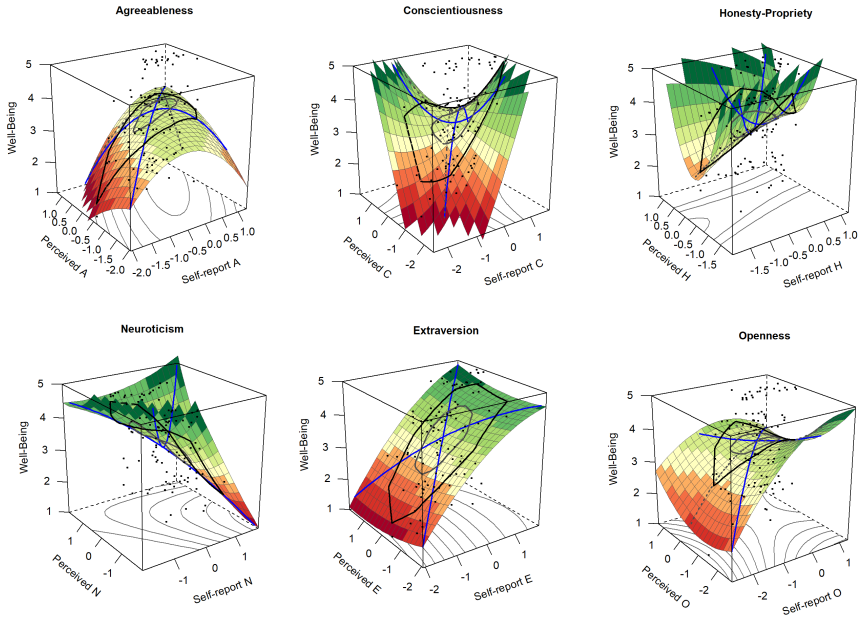

Figure 21. Accuracy and Well-Being Response Surface Plots

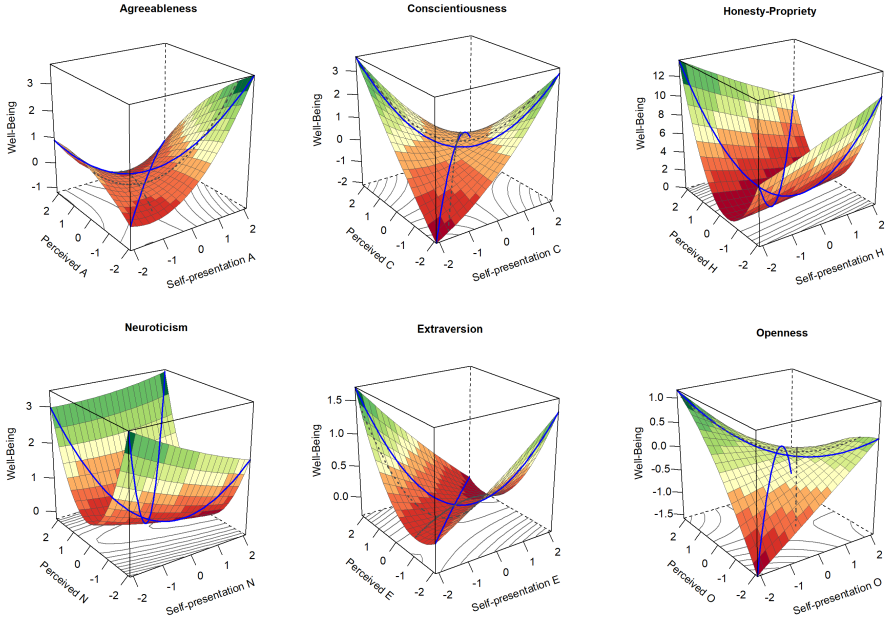

Figure 22. Idealization and Well-Being Response Surface Plots

Table 26
Surface Parameters for Idealization & Self-Reported Well-Being RSA

	Surface Parameter	estimate	CI LL	CI UL	p
agreeableness	a1	0.01	-0.68	0.70	.982
agreeableness	a2	-0.07	-1.47	1.33	.920
agreeableness	a3	0.67	-0.39	1.72	.214
agreeableness	a4	0.13	-1.69	1.95	.891
agreeableness	a5	0.62	-0.28	1.52	.178
conscientiousness	a1	0.20	-0.31	0.72	.443
conscientiousness	a2	-0.38	-1.17	0.40	.338
conscientiousness	a3	-0.05	-0.75	0.65	.886
conscientiousness	a4	-0.48	-1.52	0.57	.372
conscientiousness	a5	0.75	0.31	1.18	.001
honesty	a1	0.31	-0.61	1.22	.507
honesty	a2	1.50	0.16	2.84	.029
honesty	a3	-0.58	-1.60	0.44	.266
honesty	a4	-2.68	-4.73	-0.64	.010
honesty	a5	0.72	-0.49	1.93	.243
neuroticism	a1	-0.01	-0.62	0.59	.961
neuroticism	a2	0.77	-0.28	1.82	.151
neuroticism	a3	-0.25	-0.86	0.36	.414
neuroticism	a4	-0.50	-1.95	0.94	.496

Table 26 continued

	Surface Parameter	estimate	CI LL	CI UL	p
neuroticism	a2	0.77	-0.28	1.82	.151
neuroticism	a3	-0.25	-0.86	0.36	.414
neuroticism	a4	-0.50	-1.95	0.94	.496
neuroticism	a5	-0.08	-0.47	0.30	.676
extraversion	a1	-0.08	-0.49	0.34	.723
extraversion	a2	-0.01	-0.66	0.64	.987
extraversion	a3	-0.04	-0.52	0.43	.854
extraversion	a4	-0.33	-1.15	0.49	.430
extraversion	a5	0.20	-0.25	0.65	.383
openness	a1	0.08	-0.78	0.94	.860
openness	a2	-0.32	-1.54	0.89	.604
openness	a3	-0.17	-0.93	0.58	.652
openness	a4	-0.18	-1.60	1.25	.808
openness	a5	0.25	-0.64	1.14	.579

Note. CI LL and CI UL are the lower and upper limits of the 95 percent CI.

The surface parameters for accuracy are shown in Table 27 and the corresponding surface plots are shown in Figure 23. With the exception of openness, the pattern of results is the same across the Big Six, with a positive a1 and negative

a3 (directions are reversed for neuroticism), suggesting that perceivers like targets that are on the more desirable end of the personality domain according to self- and perceiver-reports (a1), and like targets more that they mis-perceive as being on the more desirable end (a3). This pattern of results is effectively a main effect of perceived personality, suggesting that perceivers liked targets more if they perceived them more desirably, whether that was accurate (a1) or not (a3).

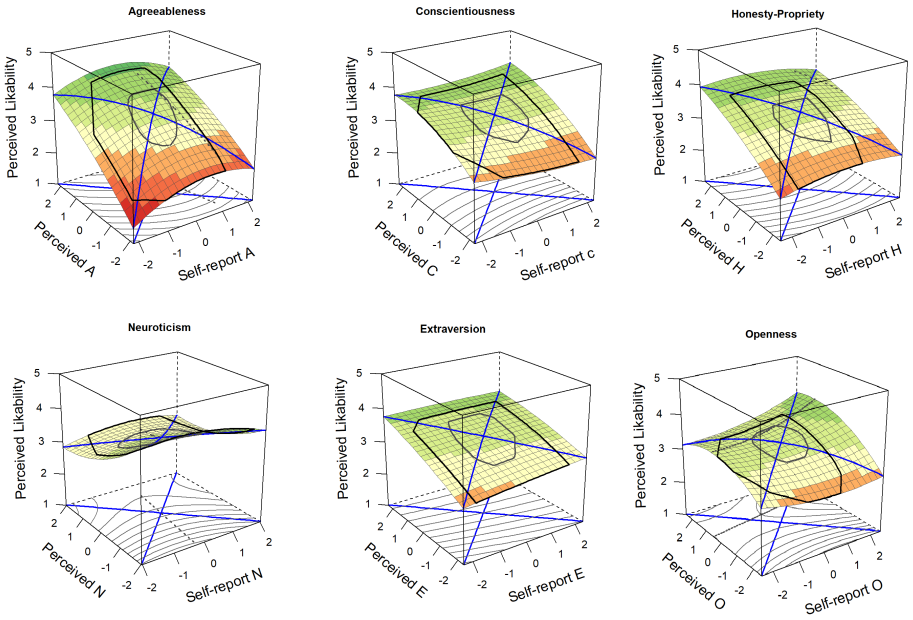

Figure 23. Accuracy and Likability Surface Plots

Turning to idealization, surface parameters for idealization are shown in Table 28 and the corresponding surface plots are shown in Figure 23. We saw virtually the same pattern of results for idealization that we did for accuracy, where all but openness have positive a1 and negative a3 values (reversed for neuroticism). This, as with accuracy, suggests a main effect whereby more positive perceptions are associated with liking

Table 27
Surface Parameters for Accuracy and Likability MLRSA

	Surface Parameter	estimate	CI LL	CI UL	p
agreeableness	a1	0.56	0.44	0.68	$< .001$
agreeableness	a2	-0.16	-0.34	0.02	.073
agreeableness	a3	-0.41	-0.55	-0.28	$< .001$
agreeableness	a4	-0.11	-0.32	0.09	.281
agreeableness	a5	-0.06	-0.22	0.09	.440
conscientiousness	a1	0.28	0.17	0.40	$< .001$
conscientiousness	a2	0.00	-0.14	0.13	.961
conscientiousness	a3	-0.32	-0.45	-0.19	$< .001$
conscientiousness	a4	-0.07	-0.22	0.09	.401
conscientiousness	a5	0.07	-0.04	0.18	.197
honesty	a1	0.31	0.15	0.47	$< .001$
honesty	a2	-0.07	-0.27	0.12	.449
honesty	a3	-0.33	-0.50	-0.17	$< .001$
honesty	a4	-0.04	-0.26	0.17	.682
honesty	a5	0.00	-0.18	0.17	.963
neuroticism	a1	-0.29	-0.41	-0.18	$< .001$
neuroticism	a2	0.06	-0.09	0.21	.459

Table 27 continued

	Surface Parameter	estimate	CI LL	CI UL	p
neuroticism	a3	0.23	0.09	0.36	.002
neuroticism	a4	-0.01	-0.18	0.16	.911
neuroticism	a5	-0.13	-0.27	0.00	.050
extraversion	a1	0.25	0.15	0.35	$< .001$
extraversion	a2	-0.04	-0.15	0.07	.509
extraversion	a3	-0.18	-0.30	-0.05	.005
extraversion	a4	0.00	-0.16	0.15	.954
extraversion	a5	0.01	-0.09	0.12	.827
openness	a1	0.22	0.08	0.36	.002
openness	a2	0.00	-0.17	0.16	.956
openness	a3	-0.10	-0.26	0.06	.215
openness	a4	-0.08	-0.28	0.11	.387
openness	a5	0.11	-0.04	0.26	.152

Note. CI LL and CI UL are the lower and upper limits of the 95 percent CI.

whether they match targets' ideal (a1) or not (a3). Openness again breaks from this pattern, but in this case is characterized by a moderate positive a1 and negative a4. This suggests that idealization is associated with liking when targets are self-presenting as higher in openness (a1) and that liking is higher the more perceptions match targets' self-presentation goals (a4).

Table 28
Surface Parameters for Idealization and Likability MLRSA

	Surface Parameter	estimate	CI LL	CI UL	p
agreeableness	a1	0.53	0.40	0.66	$< .001$
agreeableness	a2	0.00	-0.16	0.15	.976
agreeableness	a3	-0.41	-0.56	-0.26	$< .001$
agreeableness	a4	0.07	-0.10	0.24	.418
agreeableness	a5	0.11	-0.02	0.24	.108
conscientiousness	a1	0.30	0.17	0.43	$< .001$
conscientiousness	a2	0.02	-0.11	0.16	.731
conscientiousness	a3	-0.31	-0.45	-0.18	$< .001$
conscientiousness	a4	-0.11	-0.24	0.02	.096
conscientiousness	a5	0.07	-0.04	0.18	.197
honesty	a1	0.33	0.17	0.49	$< .001$
honesty	a2	-0.11	-0.31	0.08	.262
honesty	a3	-0.31	-0.47	-0.14	$< .001$
honesty	a4	-0.03	-0.25	0.19	.784
honesty	a5	0.00	-0.17	0.18	.975
neuroticism	a1	-0.25	-0.38	-0.11	$< .001$
neuroticism	a2	0.06	-0.07	0.19	.385
neuroticism	a3	0.26	0.12	0.40	$< .001$

Table 28 continued

	Surface Parameter	estimate	CI LL	CI UL	p
neuroticism	a4	0.05	-0.11	0.21	.517
neuroticism	a5	-0.12	-0.24	-0.01	.041
extraversion	a1	0.15	0.04	0.27	.009
extraversion	a2	-0.05	-0.15	0.05	.348
extraversion	a3	-0.26	-0.39	-0.13	$< .001$
extraversion	a4	-0.07	-0.19	0.06	.295
extraversion	a5	-0.03	-0.12	0.07	.593
openness	a1	0.29	0.11	0.48	.002
openness	a2	0.02	-0.16	0.20	.820
openness	a3	-0.03	-0.22	0.16	.776
openness	a4	-0.21	-0.41	-0.02	.033
openness	a5	0.09	-0.07	0.25	.248

Note. CI LL and CI UL are the lower and upper limits of the 95 percent CI.

Discussion

Study 3 was aimed at examining the extent to which twitter profiles communicate a consistent, accurate, and/or idealized impression of individuals' personalities, and additionally provide insight into social functions of how people present themselves on Twitter. Findings indicate an appreciable degree of consensus, suggesting that

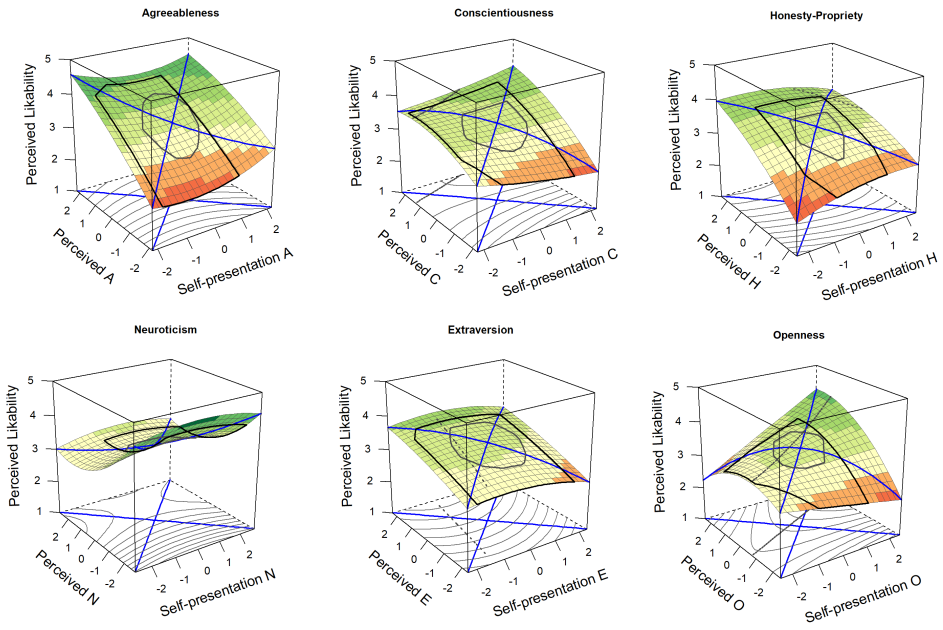

Figure 24. Idealization and Likability Surface Plots

perceivers largely agree what targets are like based on their profiles. However, these impressions reach only a small degree of accuracy across the board, which is indistinguishable from chance guessing for most domains. The lack of accuracy is *not* explained by idealization, as perceptions were not overwhelmingly influenced by how targets want to be seen. Furthermore, accuracy appears to be una˙ected by follower-network density, and showed very little systematic variation across targets and perceivers more generally. Finally, we found that profile-based perceptions relate to well-being and likability in some more and less straightforward ways. These findings have implications for how people present themselves on Twitter and online environments more generally and why they do so. These findings have implications for how p eople present themselves

on Twitter and online environments more generally and why they do so. Our findings are somewhat opposed to the work by (Back et al., 2010), which found that perceptions based on Facebook profiles are more similar to targets' real (rather than ideal) personalities. Our findings are somewhat murkier, suggesting that perceptions are more similar to what targets are really like for some domains (openness, neuroticism), more like how targets want to be seen for others (conscientiousness, honesty), and a roughly even mix of the two for others (agreeableness and extraversion). At first glance, there might be little theoretical sense to these results - they don't track with overall evaluativeness (i.e., openness is perceived more accurately, honesty more ideally) for example. However, some of these results may be better understood by considering them in conjunction with the results of the RSAs.

The RSAs for neuroticism suggested that well-being is higher when people are accurately perceived as higher in neuroticism. This might be why accuracy is greater than idealization for neuroticism - presenting an idealized front might have an intrapersonal cost in the way of lowered well-being that most aren't willing to pay. Indeed, coupled with the RSA suggesting a negative main effect of perceived neuroticism on likability, these results point to an interesting tension between their intrapersonal needs to enhance well-being by expressing more negative a˙ect and their interpersonal needs to enhance liking by expressing less negative a˙ect. More generally, our results highlight that being perceived more positively has interpersonal benefits, but occasionally has intrapersonal costs. Interestingly, idealization for openness appears to be associated with greater liking, which makes it puzzling why we see greater accuracy than idealization.

intrapersonal needs to enhance well-being by expressing more negative affect and their interpersonal needs to enhance liking by expressing less negative affect. More generally, our results highlight that being perceived more positively has interpersonal benefits, b ut o c casionally h as i ntrapersonal c o sts. I nterestingly, i dealization for openness appears to be associated with greater liking, which makes it puzzling why we see greater accuracy than idealization. However, there appears to be such a clear signal of openness on Twitter - based both on consensus and accuracy in this Study as well as predictive accuracy in Studies 1 and 2 - that it might be too challenging to fake, even if participants are motivated to do so. For conscientiousness and honesty-propriety, response surface results suggested that accuracy was associated with lower well-being, whereas idealized perceptions of honesty were associated with greater well-being. It is not clear why well-being is associated with inaccuracy for conscientiousness and honesty, but it is interesting that these are the domains that show the most idealization and least accuracy.

V. GENERAL DISCUSSION

To what extent are our personalities reflected in and ultimately recoverable from digital footprints? Perhaps unsurprisingly, the answer to this question appears to vary across personality domains, types of digital footprints (i.e., language, network ties, profiles), and whether one is using machine learning algorithms or human judges. Indeed, openness was among the more accurately predicted or inferred across studies, predictions made from followed accounts were generally more accurate than those made from tweets, and machine learning algorithms tended to reach greater accuracy than human judges. It is of course possible that differences in accuracy within and across studies reflect specific features of the technologies and methods used presently and that future work using different technologies or with a different design might find different results. At the same time, they may reflect something deeper and more enduring, possibly reflecting differences in how personality is manifest in the behaviors afforded by online social networks like Twitter and the extent to which the records of those behaviors can be used to infer personality for basic or applied purposes.

Accuracy and its Implications for Personality Expression Online

Accuracy varied considerably both within and across studies, with accuracy estimates ranging from an r of .45 for predicting openness from followed accounts to a low of .04 for human judges' perceptions of conscientiousness from profiles. Though all of these estimates are far from perfectly accurate (i.e., from an r of 1), some met

or exceeded their benchmarks (e.g., Openness and neuroticism from followed accounts; see Figure 15), others were close but lower (e.g., conscientiousness from tweets; see Figure 10), and others were substantially lower (e.g., agreeableness from tweets or followed accounts; see Figures 15 and 10). Broadening out, the higher-end estimates of accuracy are approximately as high as meta-analytic estimates of the accuracy achieved by family members and close friends (r's from approximately .3 for judgments of agreeableness by close friends to .5 for judgments of extraversion by family members; see Table 5 from Connelly & Ones, 2010) and the lower-end estimates are approximately as low as meta-analytic estimates of accuracy achieved by strangers from a variety of information sources (r's from approximately .1 for strangers judging neuroticism to .2 for strangers judging extraversion; see Table 5 from Connelly & Ones, 2010). Thus, while no prediction or judgment exhibited perfect accuracy (r of 1) or inaccuracy (r of 0), they tended to range from approximately as (in)accurate as a stranger to about as accurate as a close friend or family member.

Accuracy tended to be highest for openness across studies, which is consistent with prior work examining Facebook (Back et al., 2010; Kosinski et al., 2013; Park et al., 2015). This is especially interesting considering that openness is typically considered a less observable and more evaluative trait, two features of personality domains which are thought to depress consensus and accuracy (Connelly & Ones, 2010; John & Robins, 1993; Vazire, 2010). One possibility is that openness is much

more observable in online spaces than it is offline, and that this is part of why it can be predicted by machine learning algorithms and inferred by people with relatively greater accuracy. Indeed, consider what online social networks like Twitter afford to their users: they provide people a place to consume and express their interests. On Twitter, this might include tweeting about an essay one finds interesting, following favorite artists or public intellectuals associated with one's interests, or mentioning those interests in the bio field of one's profiles. In this way, individual differences in openness may be more relevant to the behaviors that Twitter affords, resulting in relatively more information about users' degree of openness in the digital records of those behaviors.

On the other hand, accuracy tended to be lower for agreeableness across studies, though it was inferred by human judges from profiles in Study 3 with accuracy approximately between meta-analytic estimates of judgments made by strangers and work colleague (Connelly & Ones, 2010). Agreeableness, like openness, is generally considered lower in observability and higher in evaluativeness, but differs from openness in being highly interpersonal in nature. Indeed, individuals higher in agreeableness tend to treat others more kindly and be more trusting of others, and lower scorers tend to be more critical of others, less considerate, and more rude. This may make agreeableness a poor match for the approaches taken here, especially in Studies 1 and 2, which likely had trouble with the nuances of how agreeableness is expressed behaviorally online. For example, swear words may have been useful

indicators of agreeableness when a user was cussing someone out, but those same words may have also been used as words of support or encouragement in another conversational context. This could be one reason human judges were somewhat accurate in inferences of agreeableness from profiles despite the very poor performance of the machine learning algorithms in Studies 1 and 2 - human judges may have been able to pick up on some of these nuances, even in the relatively small amount of information contained in users' profiles.

The specificity, or lack thereof, with which the machine learning algorithms predicted different personality domains likewise speaks to how behavior is manifest online. Although some predictions appeared specific to the corresponding domain - like openness in both Studies 1 and 2 - many did not, and the correlations between predicted scores were higher than would be expected based on the observed (self-reported) scores. The lack of specificity is not altogether surprising given that personality-behavior relations are thought to be complex many-to-many, rather than one-to-one, mappings (Wood, Gardner, & Harms, 2015), and so it is unlikely that cues relate to one and only one personality domain. However, it is interesting to note that the correlations between predicted scores roughly approximated the Big Two (Digman, 1997; Saucier & Srivastava, 2015) for tweet-based predictions, possibly suggesting that tweets may be better captured by this broader structural model of personality than the Big Six. This could be examined in future work by training models to predict the broader Big Two and comparing the accuracy with which they

can predict those broad factors to the accuracy of the relatively narrower Big Six.

More generally, the present work highlights many of the diculties inherent in predicting personality from digital footprints. Indeed, behaviors are multiply determined and can have vastly different psychological meanings in different social contexts and for different people. The fact that we can apply relatively blunt tools to these noisy records and infer any aspects of someone's personality with some degree of accuracy – let alone at accuracy similar to judgments made by a close friend or family member – is somewhat surprising and promising. With time and refinement, such techniques might even become useful for answering basic scientific q uestions or for applied purposes. Presently, however, many of the predictions and judgments examined here have yet to pass even the most basic requirement of predictive accuracy. Moreover, even those that do require further validity research before they can be interpreted in scientific r esearch o r application.

Implications for identity claims and behavioral residue. To what extent are differences in accuracy within and between studies due to differences in the contribution of identity claims and behavioral residue? Although it is dicult to say, it is worth speculating about the possibility that this could underpin some of the present findings. T heoretically, p rofile-based ju dgments sh ould re ly mo st on identity claims, followed-account-based predictions should rely most on behavioral residue, and tweet-based predictions should rely on cues generated by both processes. Some prior work suggests that predictions made with identity claims are more accurate

than those made with behavioral residue, suggesting that tweet-based predictions should be more accurate than followed-account-based predictions (Gladstone et al., 2019). At the outset of this dissertation, I proposed a slightly more nuanced possibility whereby identity claims are less accurate for more evaluative traits due to being more subject to (potentially inaccurate) self-presentation efforts. Our findings are at odds with both possibilities. If anything, followed-account-based predictions were generally more accurate, suggesting that predictions made from behavioral residue might be more accurate than predictions made from identity claims. With respect to the second possibility, no systematic relation between accuracy and evaluativeness emerged within or across studies. Accuracy for the highly evaluative domain of openness was greater than other less evaluative domains (e.g., extraversion) across studies, and accuracy was often similarly high (and similarly low) for domains that differ in terms of evaluativeness.

One possible explanation is that we were wrong about which cues contain more identity claims or behavioral residue, but it seems dicult to explain why followed accounts would contain more identity claims than tweets. Alternatively, it's possible that we were correct about which cues contain more or less behavioral residue and identity claims, but failed to consider how these cue types interact with features of the judgment procedure. That is, it could be that machine learning algorithms primarily achieve accuracy through using behavioral residue rather than identity claims and human judges primarily achieve accuracy through identity claims rather

than behavioral residue, and that the machine-behavioral-residue and human-identity-claim combinations thus look more similar to one another than the alternative. This is an interesting possibility that future research should evaluate in a design better suited to examine it directly (e.g., a fully-crossed design between cue category and judgment procedure).

It is worth considering the possibility that the relative presence of identity claims and behavioral residue may have less straightforward implications for accuracy than previously thought. Indeed, careful consideration of how people navigate complex and sometimes competing social motives in online spaces may provide clearer insight into how cues that are more or less subject to self-presentation affect the accuracy of different judgment procedures.

The social functions of self-presentation and personality expression on twitter. One goal of this project, particularly Study 3, was to examine the extent to which people present an idealized front on Twitter and why they might decide to present themselves more accurately or ideally. Evidence here was mixed, with human-based perceptions showing evidence of both accuracy and idealization, and computerized predictions seeming relatively robust to differences in evaluativeness across domains. However, we did find evidence that perceivers like others more if they perceive them more positively, which could provide the motivation to manage impressions highlighted central to self-presentation according to Leary and Kowalski (1990). However, this motivation to be seen positively and reap the interpersonal

reward of greater likability might be at odds with the intrapersonal gains in well-being associated with expressing one's self accurately and being self-verified (Swann et al., 1989). We only found evidence of this tension for neuroticism, and even found that being perceived less accurately was beneficial for conscientiousness and honesty, a pattern of results which may explain why we saw more idealization for some domains (conscientiousness and honesty) more accuracy for others (neuroticism and openness), and a roughly even split for others (extraversion and agreeableness). More generally, this, more than evaluativeness per se, might clarify the findings across studies, where some highly evaluative traits were predicted accurately (openness) and other less evaluative traits (extraversion) were difficult to predict accurately. Put differently, accuracy may be less affected by evaluativeness per se and more affected by the interplay of the the more externally-motivated evaluativeness and more internally-motivated desires to express one's less desirable characteristics. This would be broadly consistent with the work by Swann et al. (1989) on the interplay between self-verification and self-enhancement motives in interpersonal behavior.

We found little to no evidence that the density of targets' network of followers moderated profile-based accuracy, a hypothesis that stemmed from considering the constraining role audiences are thought to have on targets' behavior in the identity negotiation process (Back et al., 2010; Boyd, 2007; Hogan, 2010; Swann, 1987). One plausible explanation for this is that follower-network density is a poor proxy of the true constraining factors, either the *lowest common denominator* audience that

Hogan (2010) considered or the presence of offline friends in one's network considered important in Back and colleagues' (2010) *extended real-life hypothesis*. This seems plausible and future work could more directly measure the features these theories consider important, such as the presence of people that would take issue with unrealistic self-presentation (for the lowest common denominator approach) or how many of their twitter followers they know offline (for the extended real-life hypothesis). It is also worth considering the possibility that people with a public account, a prerequisite for being included in this study, have in mind the possible audience (which is basically anyone with access to Twitter) rather than the likely audience (one's followers). This too could be examined more directly in future work by asking participants whom they have in mind when they post on Twitter.

The Utility of Dictionaries and Implications for Selecting and Extracting Psychologically Meaningful Features from Noisy Data

One of the single most surprising results across studies was the relatively high accuracy achieved by models using dictionary-based scores of tweets. Indeed, models trained with the 77 dictionary scores, including the 68 LIWC categories (Tausczik & Pennebaker, 2010), sentiment, and the eight specific affect categories, were able to predict personality domains nearly as well as those trained with much more exhaustive and much more advanced sets of linguistic features. Moreover, the importance scores suggest that the LIWC scores were especially important for predicting personality (relative to the sentiment and affect dictionaries), which is even

more surprising given that they were developed for a very different context and kind of text (personal essays). This highlights a potentially important implication of this work, namely, the utility of domain-specific expertise in creating tools useful for extracting meaningful features from otherwise noisy digital footprints.

How were dictionary-based models able to achieve similar performance to models trained with a far greater number of predictors or with far more advanced features? One possibility is that dictionaries like LIWC are an effective filter and that this filtering capacity is especially useful when working with highly noisy data like tweets. On the technical side, this could increase accuracy by offloading feature selection from the machine learning algorithm, removing that one (potentially substantial) source of error and variability from algorithm training. More theoretically interesting, it may be that the expertise that went into the development of LIWC - both the psychological expertise that went into its initial design and development and its refinement with psychologically-informed empirical studies - make it especially useful for predicting psychological constructs like personality. More generally, the present work suggests that a little bit of domain-expertise in the design of a tool can go a fairly long way.

One interesting implication of this finding is that it highlights the promise of continued work refining tools for predicting psychological constructs from digital footprints. Dictionaries like LIWC can be refined to work even better in a domain like Twitter, by including linguistic features that are unique to twitter but relevant to

LIWC categories (e.g., adding more twitter-slang to LIWC). Perhaps even more promising, expertise-driven scoring algorithms could be developed for features where non exist, such as followed accounts, by combining discovery-oriented work like the present with careful theorizing and focused experimentation. Tools like this will likely take much more work and time to be useful for basic or applied science, but the present work provides a jumping off point for such efforts.

From Predictive Accuracy to Construct Validation

Predictive accuracy was quite high in some cases, sometimes matching or exceeding benchmarks. This begs the question of whether those models are presently useful in basic or applied scientific pursuits. To use a concrete example, predictions of openness from followed accounts were moderately accurate by most standards, slightly exceeds the closest benchmark (Facebook likes), and is even *slightly higher than* meta-analytic estimates of accuracy achieved by family members (*r's* of .45 vs. 43; Connelly & Ones, 2010). Should we start using this model as a measure of openness in psychological research? This is tempting as it would open up new possibilities, allowing us to obtain "openness" scores from millions of people passively (i.e., without them having to actively fill out a questionnaire), cheaply, and rapidly. However tempting this possibility is, it would almost certainly be jumping the gun. Indeed, it would be quite misguided to focus solely on predictive accuracy in evaluating whether or not an algorithm is ready for use in basic or applied science, and even the most promising models from the present study should undergo even

more rigorous evaluation before inferences from them are used.

Models that demonstrate sufficient predictive accuracy will need to be subjected to a rigorous program of construct validation research (Cronbach & Meehl, 1955). Indeed, the present work (especially Studies 1 and 2) can be viewed as following the tradition of criterion validity; the self-reports are treated as a gold standard criterion that we're attempting to predict with a new technique. As recently pointed out, prediction is a worthwhile goal and might be one of the more realistic goals for a research area or program, like the present, which is still in its infancy (Yarkoni & Westfall, 2017). It is worthwhile and an important step, but it is merely a first step. Moving forward, it will become necessary to begin formulating and testing the relations – the so-called nomological network – considered important for the constructs we think these predicted scores are capturing. This might include showing similar longitudinal stability and change as is reported in work with the relatively well-validated self- and observer-reported personality measures (Roberts & DelVecchio, 2000; Roberts, Walton, & Viechtbauer, 2006), similar levels of agreement with peer-reports at di˙erent levels of acquaintanceship as seen with typical self-reports (Connelly & Ones, 2010), and showing a lack of bias across groups similar to measurement invariance research (Stark, Chernyshenko, & Drasgow, 2006). Work like this will take considerable effort, but is ultimately necessary to move inferences from digital footprint from a passing curiosity to a tool useful for scientific inquiry and intervention.

Conclusion

The increasing digitization of our social world presents new opportunities for people to interact, to produce and consume content they find interesting, and to express their thoughts, feelings, and identities. Likewise, it presents new opportunities for studying these interpersonal processes. Our findings indicate that human perceivers and machine learning algorithms can infer personality with some degree of accuracy using different cues available to them, simultaneously speaking to how personality is expressed and perceived online. Moreover, the convergence across very different kinds of cues and very different kinds of "judges" suggests that some personality domains are more related to behavior on twitter, and behavior in online environments more generally. These findings thus provide an incremental increase in understanding how personality is expressed, perceived, and ultimately recoverable from digital footprints, and the consequences these processes have for individuals' well-being and social standing. While promising, the findings also emphasize the long road ahead before inferences from digital footprints could be used for either basic or applied purposes.

REFERENCES CITED

Anderson, C., Keltner, D., & John, O. P. (2003). Emotional convergence between people over time. *Journal of Personality and Social Psychology*, *84*(5), 1054–1068. https://doi.org/10.1037/0022-3514.84.5.1054

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook Profiles Reflect Actual Personality, Not Self-Idealization. *Psychological Science*, *21*(3), 372–374. https://doi.org/10.1177/0956797609360756

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, *101*(473), 119–137. https://doi.org/10.1198/016214505000000628

Barranti, M., Carlson, E. N., & Cote, S. (2017). How to Test Questions about Similarity in Personality and Social Psychology Research: Description and Empirical Demonstration of Response Surface Analysis. *Social Psychological and Personality Science*, (2001), 806–817. https://doi.org/10.1177/1948550617698204

Bates, D., & Maechler, M. (2019). *Matrix: Sparse and dense matrix classes and methods*. Retrieved from https://CRAN.R-project.org/package=Matrix

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). Quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77. https://doi.org/10.1145/2133806.2133826

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Bolker, B., & Robinson, D. (2020). *Broom.mixed: Tidying methods for mixed models*. Retrieved from https://CRAN.R-project.org/package=broom.mixed

Boyd, D. (2007). Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life. *MacArthur Foundation Series on Digital Learning - Youth, Identity, and Digital Media*, *7641*(41), 1–26. https://doi.org/10.1162/dmal.9780262524834.119

Bryk, A., & Raudenbush, S. (2002). *Hierarchical linear modeling. Applications and data analyses methods. Newburry park*. CA: Sage.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*(3), 193–217. https://doi.org/10.1037/h0047470

Burke, M., Kraut, R., & Marlow, C. (2011). Social capital on Facebook: Differentiating uses and users. *Conference on Human Factors in Computing Systems - Proceedings*, 571–580. https://doi.org/10.1145/1978942.1979023

Chan, C.-h., Chan, G. C., Leeper, T. J., & Becker, J. (2018). *Rio: A swiss-army knife for data file i/o*.

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2019). *Shiny: Web application framework for r*. Retrieved from https://CRAN.R-project.org/package=shiny

Cheung, F., & Lucas, R. E. (2014). Assessing the validity of single-item life satisfaction measures: results from three large samples. *Quality of Life Research*, *23*(10), 2809–2818. https://doi.org/10.1007/s11136-014-0726-4

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, *136*(6), 1092–1122. https://doi.org/10.1037/a0021212

Coppersmith, G. A., Harman, C. T., & Dredze, M. H. (2014). Measuring Post Traumatic Stress Disorder in Twitter. *In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM).*, *2* (1), 23–45.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *129*(1), 3–9. https://doi.org/10.1037/h0040957

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems*, 1695. Retrieved from http://igraph.org

De Choudhury, M., Counts, S., & Horvitz, E. (2013a). Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the sigchi conference on human factors in computing systems - chi '13* (p. 3267). New York, New York, USA: ACM Press. https://doi.org/10.1145/2470654.2466447

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013b). Predicting depression via social media. In *Seventh international aaai conference on weblogs and social media* (pp. 128–137). IEEE. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6124/6351%20http://ieeexplore.ieee.org/document/6302998/

DeYoung, C. G. (2015). Cybernetic Big Five Theory. *Journal of Research in Personality*, *56*, 33–58. https://doi.org/10.1016/j.jrp.2014.07.004

Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology*, *73*(6), 1246–1256. https://doi.org/10.1037//0022-3514.73.6.1246

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PLoS ONE*, *6*(12), e26752. https://doi.org/10.1371/journal.pone.0026752

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*(4), 652–670. https://doi.org/10.1037/0033-295X.102.4.652

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Gladstone, J. J., Matz, S. C., & Lemaire, A. (2019). Can Psychological Traits Be Inferred From Spending? Evidence From Transaction Data. *Psychological Science*, 095679761984943. https://doi.org/10.1177/0956797619849435

Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011). Predicting Personality from Twitter. In *2011 ieee third int'l conference on privacy, security, risk and trust and 2011 ieee third int'l conference on social computing* (pp. 149–156). IEEE. https://doi.org/10.1109/PASSAT/SocialCom.2011.33

Gosling, S. D., Ko, S. J., Mannarelli, T., & Morris, M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology*, *82*(3), 379–398. https://doi.org/10.1037/0022-3514.82.3.379

Grasz, J. (2016). Number of Employers Using Social Media to Screen Candidates Has Increased 500 Percent over the Last Decade. Retrieved from http://www.careerbuilder.com/share/aboutus/pressreleasesdetail.aspx?sd=5/14/2015%7B/&%7Did=pr893%7B/&%7Ded=12/31/2015

Henry, L., & Wickham, H. (2019). *Purrr: Functional programming tools*. Retrieved from https://CRAN.R-project.org/package=purrr

Hogan, B. (2010). The Presentation of Self in the Age of Social Media: Distinguishing Performances and Exhibitions Online. *Bulletin of Science, Technology & Society*, *30*(6), 377–386. https://doi.org/10.1177/0270467610385893

Human, L. J., Carlson, E. N., Geukes, K., Nestler, S., & Back, M. D. (2018). Do Accurate Personality Impressions Benefit Early Relationship Development?The Bidirectional Associations Between Accuracy and Liking. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000214

Humberg, S., Nestler, S., & Back, M. D. (2019). Response Surface Analysis in Personality and Social Psychology: Checklist and Clarifications for the Case of Congruence Hypotheses. *Social Psychological and Personality Science*, *10*(3), 409–419. https://doi.org/10.1177/1948550618757600

John, O. P., & Robins, R. W. (1993). Determinants of Interjudge Agreement on Personality Traits: The Big Five Domains, Observability, Evaluativeness, and the Unique Perspective of the Self. *Journal of Personality*, *61*(4), 521–551.

Kadushin, C. (2012). *Understanding social networks: Theories, concepts, and findings*. Oxford University Press.

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. Guilford.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods*, *21*(4), 493–506. https://doi.org/10.1037/met0000105

Kuhn, M., Jed Wing, C. from, Weston, S., Williams, A., Keefer, C., Engelhardt, A., ... Hunt., T. (2019). *Caret: Classification and regression training.* Retrieved from https://CRAN.R-project.org/package=caret

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). Springer.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82* (13), 1–26. https://doi.org/10.18637/jss.v082.i13

Leary, M. R. (2007). Motivational and Emotional Aspects of the Self. *Annual Review of Psychology*, *58*(1), 317–344. https://doi.org/10.1146/annurev.psych.58.110405.085658

Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological Bulletin*, *107*(1), 34–47. https://doi.org/10.1037/0033-2909.107.1.34

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, *90*(5), 862–877. https://doi.org/10.1037/0022-3514.90.5.862

Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics & Data Analysis*, *52*(1), 374–393. https://doi.org/10.1016/j.csda.2006.12.019

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2006). Distributed Representations of Words and Phrases and their Compositionality. *Neural Information Processing Systems*, *1*, 1–9. https://doi.org/10.1162/jmlr.2003.3.4-5.951

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in pre-training distributed word representations. Retrieved from http://arxiv.org/abs/1712.09405

Mohammad, S. M., & Kiritchenko, S. (2015). Using Hashtags to Capture Fine Emotion Categories from Tweets. *Computational Intelligence*, *31*(2), 301–326. https://doi.org/10.1111/coin.12024

Müller, K., & Wickham, H. (2019). *Tibble: Simple data frames*. Retrieved from https://CRAN.R-project.org/package=tibble

Nadeem, M. (2016). Identifying Depression on Twitter. *CoRR*, 1–9. Retrieved from https://arxiv.org/ftp/arxiv/papers/1607/1607.07384.pdf%7B/%%7D0Ahttp: //arxiv.org/abs/1607.07384%20http://arxiv.org/abs/1607.07384

Nestler, S., Humberg, S., & Schönbrodt, F. D. (2019). Response surface analysis with multilevel data: Illustration for the case of congruence hypotheses. *Psychological Methods*, *24*(3), 291–308. https://doi.org/10.1037/met0000199

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934–952. https://doi.org/10.1037/pspp0000020

Paulhus, D. L., & Trapnell, P. D. (2008). Self-Presentation of Personality: An Agency-Communion Framework. In *Handbook of personality psychology* (pp. 492–517).

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Empirical methods in natural language processing (emnlp)* (pp. 1532–1543). Retrieved from http://www.aclweb.org/anthology/D14-1162

Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). You are what you tweet: Personality expression and perception on Twitter. *Journal of Research in Personality*, *46*(6), 710–718. https://doi.org/10.1016/j.jrp.2012.08.008

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41*(1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

R Core Team. (2019). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., & Langer, E. J. (2017). Forecasting the onset and course of mental illness with Twitter data. *Scientific Reports*, *7*(1), 13006. https://doi.org/10.1038/s41598-017-12961-9

Roberts, B. W., & DelVecchio, W. F. (2000). The Rank-Order Consistency of Personality Traits From Childhood to Old Age: A Quantitative Review of Longitudinal Studies. https://doi.org/10.1037/0033-2909.126.1.3

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*(1), 1–25. https://doi.org/10.1037/0033-2909.132.1.1

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/

Sarkar, D. (2008). *Lattice: Multivariate data visualization with r.* New York: Springer. Retrieved from http://lmdvr.r-forge.r-project.org

Saucier, G., & Srivastava, S. (2015). What makes a good structural model of personality? Evaluating the big five and alternatives. In M. Mikulincer, P. R. Shaver, M. L. Cooper, & R. J. Larsen (Eds.), *Handbook of personality and social psychology. Vol. 3: Personality processes and individual differences* (pp. 283–305). Washington, DC.

Saucier, G., Thalmayer, A. G., Payne, D. L., Carlson, R., Sanogo, L., Ole-Kotikash, L., ... Zhou, X. (2014). A Basic Bivariate Structure of Personality Attributes Evident Across Nine Languages. *Journal of Personality*, *82*(1), 1–14. https://doi.org/10.1111/jopy.12028

Schaefer, D. R., Kornienko, O., & Fox, A. M. (2011). Misery Does Not Love Company. *American Sociological Review*, *76*(5), 764–785. https://doi.org/10.1177/0003122411420813

Schönbrodt, F. D., & Humberg, S. (2018). *RSA: An r package for response surface analysis (version 0.9.13)*. Retrieved from https://cran.r-project.org/package=RSA

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, *8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791

Soto, C. J., & John, O. P. (2017a). Short and extra-short forms of the Big Five Inventory-2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality*, *68*, 69–81. https://doi.org/10.1016/j.jrp.2017.02.004

Soto, C. J., & John, O. P. (2017b). 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), 117–143. https://doi.org/10.1037/pspp0000096

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292–1306. https://doi.org/10.1037/0021-9010.91.6.1292

Sumner, C., Byers, A., Boochever, R., & Park, G. J. (2012). Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In *2012 11th international conference on machine learning and applications* (pp. 386——393). IEEE. https://doi.org/10.1109/IRI.2012.6302998

Swann, W. B. (1987). Identity negotiation: Where two roads meet. *Journal of Personality and Social Psychology*, *53*(6), 1038–1051. https://doi.org/10.1037/0022-3514.53.6.1038

Swann, W. B., Pelham, B. W., & Krull, D. S. (1989). Agreeable fancy or disagreeable truth? Reconciling self-enhancement and self-verification. *Journal of Personality and Social Psychology*, *57*(5), 782–791. https://doi.org/10.1037/0022-3514.57.5.782

Swann, W. B., & Read, S. J. (1981). Self-verification processes: How we sustain our self-conceptions. *Journal of Experimental Social Psychology*, *17*(4), 351–372. https://doi.org/10.1016/0022-1031(81)90043-3

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*. https://doi.org/10.1177/0261927X09351676

Thalmayer, A. G., & Saucier, G. (2014). The questionnaire big six in 26 nations: Developing cross-culturally applicable big six, big five and big two inventories. *European Journal of Personality*, *28*(5), 482–496. https://doi.org/10.1002/per.1969

Tomasello, M. (2010). *Origins of human communication*. MIT press.

Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*(2), 281–300. https://doi.org/10.1037/a0017908

Watson, D., Beer, A., & McDade-Montez, E. (2014). The Role of Active Assortment in Spousal Similarity. *Journal of Personality*, *82*(2), 116–129. https://doi.org/10.1111/jopy.12039

Watson, D., Hubbard, B., & Wiese, D. (2000a). General Traits of Personality and Affectivity as Predictors of Satisfaction in Intimate Relationships: Evidence from Self- and Partner-Ratings. *Journal of Personality*, *68* (3), 413–449. https://doi.org/10.1111/1467-6494.00102

Watson, D., Hubbard, B., & Wiese, D. (2000b). Self-other agreement in personality and affectivity: The role of acquaintanceship, trait visibility, and assumed similarity. *Journal of Personality and Social Psychology*, *78*(3), 546–558. https://doi.org/10.1037/0022-3514.78.3.546

Watson, D., Klohnen, E. C., Casillas, A., Nus Simms, E., Haig, J., & Berry, D. S. (2004). Match Makers and Deal Breakers: Analyses of Assortative Mating in Newlywed Couples. *Journal of Personality*, *72*(5), 1029–1068. https://doi.org/10.1111/j.0022-3506.2004.00289.x

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2017). *Tidyverse: Easily install and load the 'tidyverse'*. Retrieved from https://CRAN.R-project.org/package=tidyverse

Wickham, H. (2019a). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from https://CRAN.R-project.org/package=forcats

Wickham, H. (2019b). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H., François, R., Henry, L., & Müller, K. (2019). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2019). *Tidyr: Easily tidy data with 'spread()' and 'gather()' functions*. Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., Hester, J., & Francois, R. (2018). *Readr: Read rectangular text data*. Retrieved from https://CRAN.R-project.org/package=readr

Wood, D., Gardner, M. H., & Harms, P. D. (2015). How functionalist and process approaches to behavior can explain trait covariation. *Psychological Review*, *122*(1), 84–111. https://doi.org/http://dx.doi.org/10.1037/a0038423

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112

Youyou, W., Stillwell, D., Schwartz, H. A., & Kosinski, M. (2017). Birds of a Feather Do Flock Together. *Psychological Science*, *28*(3), 276–284. https://doi.org/10.1177/0956797616678187