Journal of Personality and Social Psychology

0022-3514/14/$12.00 http://dx.doi.org/10.1037/pspp0000020

Automatic Personality Assessment Through Social Media Language

Gregory Park, H. Andrew Schwartz, Johannes C. Eichstaedt, and Margaret L. Kern
University of Pennsylvania

Michal Kosinski and David J. Stillwell
University of Cambridge

Lyle H. Ungar and Martin E. P. Seligman
University of Pennsylvania

Language use is a psychologically rich, stable individual difference with well-established correlations to personality. We describe a method for assessing personality using an open-vocabulary analysis of language from social media. We compiled the written language from 66,732 Facebook users and their questionnaire-based self-reported Big Five personality traits, and then we built a predictive model of personality based on their language. We used this model to predict the 5 personality factors in a separate sample of 4,824 Facebook users, examining (a) convergence with self-reports of personality at the domain- and facet-level; (b) discriminant validity between predictions of distinct traits; (c) agreement with informant reports of personality; (d) patterns of correlations with external criteria (e.g., number of friends, political attitudes, impulsiveness); and (e) test–retest reliability over 6-month intervals. Results indicated that language-based assessments can constitute valid personality measures: they agreed with self-reports and informant reports of personality, added incremental validity over informant reports, adequately discriminated between traits, exhibited patterns of correlations with external criteria similar to those found with self-reported personality, and were stable over 6-month intervals. Analysis of predictive language can provide rich portraits of the mental life associated with traits. This approach can complement and extend traditional methods, providing researchers with an additional measure that can quickly and cheaply assess large groups of participants with minimal burden.

Keywords: language, personality assessment, measurement, big data, social media

Supplemental materials: http://dx.doi.org/10.1037/pspp0000020.supp

Every day, millions of people express themselves by writing in social media (e.g., Facebook, Twitter, and blogs). Through simple text messages, people freely share their thoughts and emotions with their circle of friends, larger group of acquaintances, or even the entire online world. The written language accumulating in social media is a massive source of rich psychological data with unrealized scientific potential. If researchers can translate this language into novel measurement methods, they stand to substantially increase the scale and scope of psychological research. In this article, we describe and evaluate one such method: the automatic language-based assessment of personality using social media.

Gregory Park, Department of Psychology, University of Pennsylvania; H. Andrew Schwartz, Computer & Information Science, University of Pennsylvania; Johannes C. Eichstaedt and Margaret L. Kern, Department of Psychology, University of Pennsylvania; Michal Kosinski and David J. Stillwell, Psychometrics Centre, University of Cambridge; Lyle H. Ungar, Computer & Information Science, University of Pennsylvania; Martin E. P. Seligman, Department of Psychology, University of Pennsylvania.

Support for this publication was provided by the Robert Wood Johnson Foundation's Pioneer Portfolio, through the "Exploring Concepts of Positive Health" grant awarded to Martin Seligman, by a grant from the Templeton Religion Trust, and by the University of Pennsylvania Positive Psychology Center.

Correspondence concerning this article should be addressed to Gregory Park, Department of Psychology, University of Pennsylvania, 3701 Market Street, 2nd floor, Philadelphia, PA 19104. E-mail: gregpark@sas.upenn.edu

Language and Personality

Research on the diagnostic value of language has surged as computerized text analysis tools have become more accessible. Within the last decade, over 100 studies have linked language use to a wide range of psychological correlates (Tausczik & Pennebaker, 2010). Some of the earliest work found that word use was a stable individual difference with several modest but reliable correlations with self-reports of personality (Pennebaker & King, 1999). For example, individuals scoring higher on neuroticism used first-person singulars (e.g., *I*, *me*, *mine*) more frequently, whereas extraversion related to using more positive emotion words (e.g., *great*, *happy*, *amazing*). Many of these earliest findings have since been replicated across multiple studies (e.g., Schwartz et al., 2013b; Yarkoni, 2010). Several studies have used a similar approach—comparing word use with self-reports or behavioral assessments—and have yielded an impressive body of evidence linking language to personality (e.g., Cohen, Minor, Baillie, & Dahir, 2008; Fast & Funder, 2008; Hirsh & Peterson, 2009; Lee, Kim, Seo, & Chung, 2007; Mehl, Gosling, & Pennebaker, 2006;

Pennebaker, Mehl, & Niederhoffer, 2003; Tausczik & Pennebaker, 2010).

Social media has created an unprecedented amount of written language, vast amounts of which is publicly available. Twitter users alone write approximately 500 million messages every day (Reuters, 2013). Previously, researchers relied on either historical language samples, such as literature, scientific abstracts, and other publications, or prompted participants to write new text. Now, social media provides researchers with the natural language of millions of people with relative ease.

For personality researchers, the potential benefits of social media extend beyond massive sample sizes. First, social media language is written in natural social settings, and captures communication among friends and acquaintances. Essentially, social media offers an ongoing experiential sampling method that is naturally a part of many peoples' lives. Second, expensive prospective studies are less necessary, because the data can be retroactively accessed for research purposes. Third, social media users disclose information about themselves at unusually high rates; for many users, a frequent topic of discussion is themselves (Naaman, Boase, & Lai, 2010). Fourth, social media users typically present their true selves and not just idealized versions (Back et al., 2010). Thus, social media language potentially is a very rich source of personality data.

Closed Versus Open Approaches to Language Analysis

With few exceptions, psychological studies have used a *closed-vocabulary*, *word counting* approach to analyze language. This method starts with lists of words that are combined into categories (e.g., pronouns), based on theory, and then counts the relative frequency of these words within a body of text. This method's most popular implementation, the Linguistic Inquiry and Word Count (LIWC; Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007), automatically counts word frequencies for over 60 psychologically relevant categories, such as "function words" (e.g., *articles*, *pronouns*, *conjunctions*), "affective processes" (e.g., *happy*, *cried*, *nervous*), and "social processes" (e.g., *mate*, *friend*, *talk*). Because this approach starts with predefined categories of words, it has been described as "closed-vocabulary" (Schwartz et al., 2013b).

Closed-vocabulary methods have become particularly popular in recent analyses of social media language (Golbeck, Robles, & Turner, 2011; Holtgraves, 2011; Sumner, Byers, Boochever, & Park, 2012). Within computer science and related fields, several researchers have used closed-vocabulary analyses to study how well social media language can predict a user's personality. For example, Golbeck, Robles, and Turner (2011) used a closed-vocabulary approach to analyze the language written in the personal profiles and messages of Facebook users, who also completed personality measures. Relative uses of LIWC word categories (e.g., positive emotions, social processes) were then used as predictors in statistical models, where the outcomes were self-reports of personality. When applied to out-of-sample users, these models predicted users' personality traits better than chance, and the authors concluded that "users' Big Five personality traits can be predicted from the public information they share on Facebook" (Golbeck et al., 2011, p. 260). Similar predictive personality models have been built using closed-vocabulary language features of language from Twitter (e.g., Golbeck et al., 2011; Sumner et al., 2012).

In contrast, techniques from computational linguistics offer finer-grained, *open-vocabulary* methods for language analysis (e.g., Grimmer & Stewart, 2013; O'Connor, Bamman, & Smith, 2011; Schwartz et al., 2013b; Yarkoni, 2010). Open-vocabulary methods do not rely on a priori word or category judgments; rather, they extract a comprehensive collection of language features from the text being analyzed. In contrast to closed-vocabulary methods, open-vocabulary methods characterize a language sample by the relative use of (a) single, uncategorized words; (b) nonword symbols (e.g., emoticons, punctuation); (c) multiword phrases; and (d) clusters of semantically related words identified through unsupervised methods, or topics (Blei, Ng, & Jordan, 2003). Because these language features are not identified a priori, these methods can accommodate neologisms and unconventional language use. Compared with closed-vocabulary methods, open-vocabulary methods extract more numerous and richer features from a language sample. These methods can substantially improve predictions of personality.

Schwartz et al. (2013b) used both open-vocabulary and closed-vocabulary language features to predict the personality of 75,000 Facebook users. Models using open-vocabulary features significantly outperformed closed-vocabulary models, and the resulting predictions correlated with self-reports of personality in the range of $r = .31$ (for agreeableness and neuroticism) to $r = .41$ (for openness to experience) compared with $r = .21$ to .29 using closed-vocabulary features. These results supported earlier findings by Iacobelli, Gill, Nowson, and Oberlander (2011) who reported that open-vocabulary method significantly outperformed closed-vocabulary methods when predicting the personality of 3,000 bloggers.

If open-vocabulary language models can reliably predict individual differences in personality, can these models be the basis for a new mode of personality assessment? If so, this could lead to a class of fast, inexpensive *language-based assessments* (LBAs) that could be easily applied to existing social media samples. To date, researchers have evaluated predictive models of psychological characteristics on the basis of predictive accuracy alone, that is, how accurately a model can predict self-reports of personality (e.g., Golbeck et al., 2011; Iacobelli et al., 2011; Schwartz et al., 2013b; Sumner et al., 2012). Although predictive accuracy is a good indicator of this method's convergent validity, little is known about their broader validity and reliability. For example, do LBAs of personality adequately discriminate between distinct traits? Do they agree with other assessment methods? Are they capable of predicting relevant external criteria? Are LBAs sufficiently stable over time? These basic psychometric properties ought to be clearly demonstrated before researchers can comfortably use these methods.

The Present Study

In this study, we describe and evaluate our approach to LBAs of personality. Our method extends previous research in several ways. We used an unprecedented sample size to build our language model and used an open-vocabulary approach. Most prior research on personality and language used samples in the hundreds. We built our model on a sample of over 66,000 participants. Previous

research used a closed-vocabulary approach to language analysis. We used an open-vocabulary approach, which generated a rich set of several thousands of language features, including single words, multiword phrases, and clusters of semantically related words, or topics. These features were used as predictor variables in a regression model; when used with a variety of dimensionality-reduction methods (described below), more accurate predictions of personality from language use resulted than has occurred in any prior study. Finally, we used several procedures to extensively evaluate the validity of LBAs beyond simple predictive accuracy.

We start with a detailed description of our language processing methods and statistical modeling procedures. We applied these methods and built a predictive model of personality within a training sample of over 66,000 Facebook users, each of whom volunteered samples of their language and completed Big Five personality measures. We then evaluated this model within a separate validation sample of approximately 5,000 Facebook users. To avoid overfitting the regression models, we split our original sample into separate training and validation samples. In the validation sample, we used the prediction models—which were built over the training sample—to generate language-based predictions of users' personality traits. These predictions constitute our LBAs of personality.

Within the validation sample, we evaluated the validity and reliability of LBAs through a series of analyses. First, we compared LBAs with (a) self-report questionnaires of personality, (b) informant reports of personality, and (c) external criteria with theoretically expected correlations with personality. Second, we examined the language features that were most strongly correlated with predictions of each trait. Finally, we evaluated the stability of LBAs by comparing predictions over time, analogous to the traditional test–retest approach to reliability assessment.

Method

Participants

Participants were drawn from users of myPersonality, a third-party application (Kosinski & Stillwell, 2011) on the Facebook social network, which allowed users to take a series of psychological measures and share results with friends. The myPersonality application was installed by roughly 4.5 million users between 2007 and 2012. All users agreed to the anonymous use of their survey responses for research purposes.

Our analytic sample was a subset of myPersonality users (N = 71,556) who also allowed the application to access their *status messages* (i.e., brief posts on the user's main Facebook page). Unlike direct messages between specific users, status messages are *undirected* and displayed to a user's entire social network. Many users update their status message throughout the day, keeping their social network abreast of their current activities, moods, and thoughts. We limited the analytic sample to users who wrote at least 1,000 words across their status messages, provided their gender and age, and were younger than 65 years of age.

We captured every status message written by our study volunteers between January 2009 and November 2011, totaling over 15 million messages. Users wrote an average of 4,107 words across all status messages (median = 2,933; SD = 3,752). Mean user age was 23.4 (median = 20, SD = 8.9), and over half (62.6%) were female.

Personality Measures

All participants completed measures of personality traits as defined by the NEO-PI-R five factor model (Costa & McCrae, 1992): openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. Items came from the International Personality Item Pool (IPIP; Goldberg et al., 2006). Participants were free to complete measures of varying lengths, ranging from 20 to 100 items. A subset of users (n = 348) completed an additional 336-item IPIP proxy to the NEO-PI-R, designed to assess 30 facet-level personality traits.

Training and Validation Samples

To avoid overfitting, we split the analytic sample into a training sample (n = 66,732), which was used to build the regression models, and a validation sample (n = 4,824), which was used to evaluate predictions of the fitted regression models. In creating the training and validation samples, we balanced two goals. On one hand, we wanted to maximize the sample size for training the regression models, because predictive performance generally improves as more training observations are included. On the other hand, we wanted to retain a sufficiently large validation sample to ensure reliable evaluations. Additionally, because we were interested in evaluating LBAs against external measures, it was necessary that many users in the validation sample had completed these measures.

We estimated that a validation sample of roughly 5,000 users would provide very stable evaluations and leave the majority of the analytic sample for training. To create the validation sample, we first included all 348 users that completed the 336-item facet-level measure. Next, we oversampled users that completed external measures to ensure at least 500 users per measure in the validation sample. Finally, we randomly sampled users from the analytic sample until the sample size reached 5,000. Within the validation sample, 2,324 users completed the 20-item version of the IPIP measure, 1,943 completed the 100-item version, and 557 completed other variants ranging from 30 to 90 items. One-hundred and 76 users were missing data on the exact number of items completed, so these users were removed from the validation sample. This resulted in a final validation sample size of 4,824. The remaining 66,732 users were used as the training sample.

External Criteria

Many users in our sample also completed additional personality-related measures and volunteered Facebook profile information, and we used these as external criteria in our validity evaluations. Sample sizes below indicate the number of users in the validation sample with observations on each criterion. In each case, higher scores indicate more of that domain (e.g., greater life satisfaction, more self-monitoring).

Satisfaction With Life Scale. One-thousand and 82 users completed the Satisfaction With Life Scale, a five-item measure assessing life satisfaction (Diener, Emmons, Larsen, & Griffin, 1985).

Self-Monitoring Scale. Nine-hundred and 27 users completed the Self-Monitoring Scale, a 25-item scale assessing the degree to which one regulates self-presentation using situational cues (Snyder, 1974).

Orpheus Personality Questionnaire. Eight-hundred and 64 users completed two subscales of the Orpheus Personality Questionnaire (Rust & Golombok, 2009): fair-mindedness and self-disclosure. The fair-mindedness subscale assesses impartiality and fairness in decision-making; the self-disclosure subscale assesses self-disclosure and transparency in self-presentation.

Pennebaker Inventory of Limbic Languidness (PILL). We used responses to two items from the PILL, an inventory of respondents' experience of common physical symptoms and sensations (Pennebaker, 1982). Seven-hundred and 36 users indicated how many times they had recently visited a physician due to illness, and 733 users indicated the number of recent days that they had been sick. Due to skew, we log-transformed the number of physician visits and days sick after adding one to each observation.

Barratt Impulsiveness Scale (BIS-11). Five-hundred and 49 users completed the BIS-11, a 30-item scale assessing general and specific dimensions of impulsiveness (Patton, Stanford, & Barratt, 1995; Stanford et al., 2009). We used the full scale score.

Informant reports of personality. Seven-hundred and 45 users were rated by users' friends who had also installed the myPersonality application. Friends were given the option to rate the users' personality, using 10 items (two items per factor) from the 100-item measure.

Public profile information. In addition to self-reports of personality, the application also collected information from users' public Facebook profiles at the time of installation. We used this to determine the number of Facebook friends and political attitudes. The number of Facebook friends was available for 1,906 users. Due to skew, we log-transformed the number of friends before correlating with personality ratings. Seven-hundred and 56 users completed a field in their Facebook profile regarding political views. We considered those who identified themselves as *very conservative* ($n = 12$), *conservative* ($n = 201$), *moderate* ($n = 139$), *liberal* ($n = 339$), or *very liberal* ($n = 65$), and coded these responses from -2 (very conservative) to $+2$ (very liberal).

Language Model Creation: Training Sample

Our method of building a language model of personality consisted of three stages: feature extraction, dimensionality reduction, and regression modeling (see Figure 1).

Linguistic Feature Extraction. In the feature extraction stage, we transformed each user's collection of status messages into frequencies of hundreds of thousands of simpler language features. We extracted two types of language features: (a) words and phrases, and (b) topics.

Words and phrases. To extract words and phrases, we first split each of the users' messages into single words. Words were defined by an emoticon-aware tokenizer (Potts, 2011), which is sensitive to conventional words but also to nonword features like emoticons (e.g., *:-)*), punctuation (e.g., *!!!*), and nonconventional spellings and usages (e.g., *omg*, *wtf*). In addition to single words, we extracted phrases—two- and three-word sequences that occur at rates much higher than chance (e.g., *happy birthday, I love you*). We identified such phrases by calculating the pointwise mutual information (PMI) for each phrase, defined as:

$$pmi(phrase) = log(p(phrase)/\prod p(word))$$

where p(phrase) is the probability of the phrase based on its relative frequency, and Π p(word) is the product of the probabilities of each word in the phrase (Church & Hanks, 1990). The PMI criterion identifies *phrases* as co-occurrences of words that occurred more frequently than the individual probabilities of occurrence of the constituent words would suggest by chance. We kept all two- and three-word phrases with PMI values greater than $3 \times$ *size*, where *size* is the number of words in the phrase.

After identifying words and phrases, we counted the occurrence of each of these features within each user's language sample and

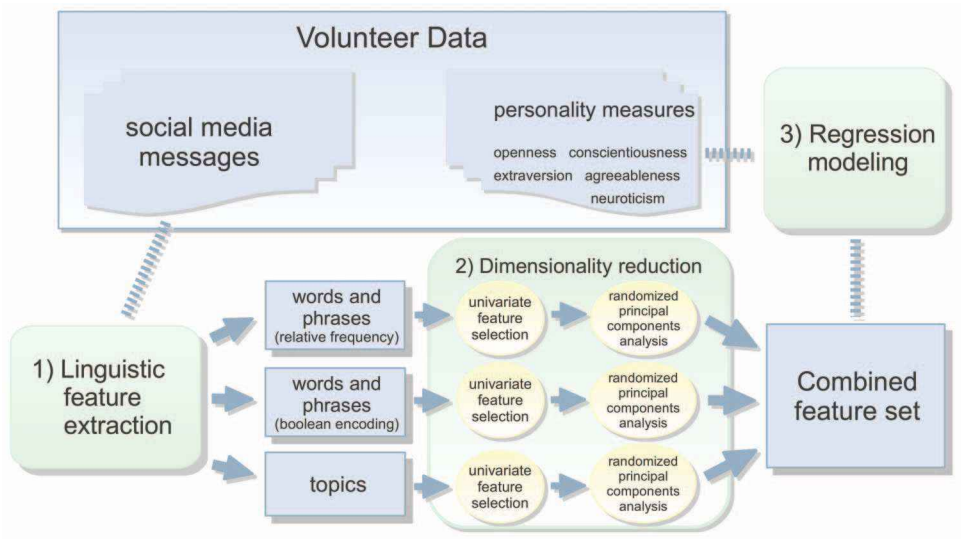


Figure 1. Process flow diagram illustrating the method of building language models of personality traits.

then normalized these counts based on each user's total word count. This created several million normalized values per user. Most of these features were never or only rarely used by the majority of users. To reduce the number of features, we kept the words and phrases that were used at least once by 1% of the sample. In addition, we created binary representations (0 or 1) of every language feature, indicating whether the word or phrase was ever used by each user. For some features, these more robust binary representations capture incremental variance and improve predictions.

Topics. Topics are clusters of semantically related words created through *latent Dirichlet allocation* (LDA; Blei et al., 2003; for an introduction to topic models, see Atkins et al., 2012). LDA assumes that a document (in this case, individual status messages) is a mixture of a fixed number of latent topics where each topic is a cluster of related words (this fixed number is specified in advance by the analyst). Through an iterative procedure, LDA identifies and refines the specified number of clusters of words. An example of a topic identified by LDA in our language sample includes the philosophically oriented words *human*, *beings*, *nature*, *spiritual*, *experience, compassion, sense, existence, reality,* and *universe*. These words tend to co-occur with each other in messages and are automatically identified by the LDA procedure. Note that this procedure is unaware of who wrote each message; it only uses distributions of words across all messages.

We fit an LDA model using an implementation provided in the Mallet package (MacCallum, 2002), setting the number of topics to 2,000. This produced 2,000 naturally occurring topics, each consisting of many words with relative weights. The topics are defined purely on the basis of the distribution of language use across statuses without consideration of personality or other outcome variables. We then calculated each individual's use of each topic, defined as the probability of using a topic:

$$p(topic,\ user) = \sum p(topic|word) \times p(word|user)$$

where $p(word|user)$ is the individual's normalized word use and $p(topic|word)$ is the probability of the topic given that word (i.e., part of the output of the fitted LDA model). For example, a person who mentions the words *human*, *spiritual*, and *reality* would have a higher probability of using the philosophical topic described above, as these three words are heavily weighted within that topic.

Dimensionality reduction. Through the feature extraction stage, our full set of language features consisted of three distinct feature sets: (a) normalized relative frequencies of words and phrases, (b) binary representations of words and phrases, and (c) topic usage. Across all three feature sets, this totaled 51,060 features (24,530 in each set of word and phrase features plus 2,000 topic features) across 66,764 users in the training set. A rough rule of thumb in predictive modeling is to use fewer features (predictors) than observations. However, in practice, predictive performance is highly dependent on aspects of the data (e.g., noise, collinearity) and the techniques used (e.g., some forms of regularization work very well when the number of features exceeds the number of observations). In our case, many of the features are highly correlated or irrelevant, so we used several techniques to reduce dimensionality and collinearity and to improve predictive performance. We processed each feature set separately, and then combined them into a single final predictor set.

Univariate feature selection. First, we identified and removed features in each set with very weak or no linear associations to the target trait (Guyon & Elisseeff, 2003). The criterion for removing features was based on a family wise error rate. We chose the value of this error rate through cross-validation within the training sample, in which we experimented with several different error rates in one randomly selected portion of the sample, and then assessed the resulting predictive performance in a held-out portion. Then, for our final model, we used the error rate that gave the best overall predictive performance. We calculated the p value corresponding to the Pearson correlation between each feature and the target trait, and features with p values above the final error rate were removed (Pedregosa et al., 2011).

Randomized principal components analysis. Next, we applied randomized principal components analysis (RPCA; Martinsson, Rokhlin, & Tygert, 2011) separately to the three reduced feature sets. RPCA provides the same regularization effect as principal components analysis, but it takes much less computing time to run than PCA by using random samples of features and observations for the singular value decomposition step. For example, our server takes approximately 2.5 hr to apply PCA to the training sample; RPCA completes in approximately 10 min. We kept a subset of the principal components from each feature set as predictors. Specifically, in each feature set, we kept k principal components, where k is equal to one-tenth of the total number of features prior to univariate feature selection. We chose $k = .10$ as the final number of features after experimenting with several values (.01, .02, .05, .10, and .2) through cross-validation in the training sample, taking the value (.10) that provided the best predictive performance. We then combined the RPCA-reduced feature sets for regression modeling.

To summarize, these two dimensionality reduction steps reduced the number of features used to predict each trait from 51,060 to 5,106. The initial feature size of 51,060, which was consistent across all five traits, combined three distinct feature sets: relative frequencies of words and phrases (24,530), binary representations of words and phrases (24,530), and topics (2,000). In univariate feature selection step, features were removed from each of the three feature sets, and the number of features removed varied by feature set and trait. For example, in the case of agreeableness, univariate feature selection kept 4,671 relative frequencies of words and phrases, 6,070 binary representations of words and phrases, and 1,420 topic usage features. In the case of conscientiousness, univariate feature selection kept 9,485 relative frequencies of words and phrases, 11,539 binary representations of words and phrases, and 1,680 topic usage features. In the RPCA step, these varying feature sets were all reduced to a fixed size: one-tenth of the original (i.e., preunivariate feature selection) size. Post-RPCA, the three feature sets for each trait contained 2,453 principal components from relative frequencies of words and phrases, 2,453 components from binary representations, and 200 components from topics. For each trait's regression model, these three reduced feature sets were combined to form a final set with 5,106 features.

Regression modeling. In the regression modeling stage, we regressed the combined feature set on users' personality measures. We fit five separate regression models (one for each Big Five trait), using a regularized form of regression known as ridge regression (Hoerl & Kennard, 1970). Ridge regression is similar to

linear regression, except it adds an additional penalty to the squared magnitude of the coefficients, biasing them toward zero. This additional bias reduces the variability of the estimated coefficients and improves predictive accuracy from the model, particularly in cases where there are many more predictors than observations and/or predictors are highly correlated.

Evaluation: Validation Sample

After fitting each model with the training sample, we generated predictions for each of the Big Five traits for all 4,824 users in the validation sample. We then conducted a series of analyses to evaluate validity and reliability.

Convergent and discriminant validity. Convergent validity was assessed by examining the correlations between LBAs and self-reports of each trait. Discriminant validity was assessed by comparing the magnitude of between-trait correlations (e.g., between extraversion and conscientiousness) within LBAs with those within self-reports. In addition, with a subset of users who completed a longer, 336-item IPIP facet-level personality measure, we examined patterns of convergence and discrimination between LBAs and self-reported personality at the facet-level.

Comparison with informant reports. We compared self-other agreement, or accuracy, of LBAs and informants for the 745 users with informant personality ratings, using correlations between self-reports of personality and LBAs (and informant reports) as our agreement metric. We first compared self-other agreement for single traits. Then, we found the average self-other agreement of each method by applying a Fisher *r*-to-*z* transformation to individual trait agreement, calculating the mean, and then transforming this mean back to the original *r* scale.

We also examined agreement between LBAs and informant reports of personality, which reflect the degree to which LBAs overlap with an external perspective. Because our language models were built using self-reports, we wanted to evaluate whether they agreed with an additional external judge in addition to self-reported personality.

Finally, we evaluated the incremental validity of LBAs over a single informant rating in two complementary ways: partial correlations and aggregate ratings. We calculated the partial correlations between LBAs and self-reported personality while controlling for informant reports. We created aggregate ratings by averaging the LBA and informant rating for each trait (after standardizing each rating across all users to weight each rating equally). Although we acknowledge that the simple average may not be the optimal weighting scheme for creating accurate ratings, we preferred an approach that was consistent with past work using aggregated ratings (e.g., Hofstee, 1994; Kolar, Funder, & Colvin, 1996; Vazire & Mehl, 2008). We then compared the agreement between these aggregate ratings and self-reports with the agreement between informant reports and self-reports. To test whether aggregate ratings were significantly more accurate than informants alone, we used a significance test for dependent correlations as described by Steiger (1980; as implemented by Revelle, 2014).

Correlations with external criteria. Two measures of the same construct should have similar patterns of correlations (in sign and magnitude) with external criteria, indicating that they map the same nomological network (Cronbach & Meehl, 1955). Therefore, we compared the patterns of correlations between (a) 14 external criteria and LBAs with those between (b) the same external criteria and self-reported personality.

We summarized patterns of correlations in three complementary ways: sign agreement, magnitudes of absolute correlations, and column-vector correlations. Sign agreement simply checks whether correlations between a criterion and both assessment methods agree in sign. For absolute correlations, we calculated the absolute correlations between each criterion and both assessment methods, and compared the relative magnitudes, testing whether the one assessment mode had significantly stronger correlations with the criterion, using a test for dependent correlations (Steiger, 1980). In addition, we summarized the absolute correlations from each assessment method by calculating the mean absolute correlations. Mean absolute correlations were calculated by transforming the absolute values of 14 correlations between each assessment and external criterion to Fisher *z*-scores, calculating the mean of these *z*-scores, and finally transforming this mean back to the original *r* scale. We then compared the magnitudes of the mean absolute correlation of each assessment method within each personality factor.

Lastly, for each personality factor, we calculated a column-vector correlation, or a correlation of correlations. We transformed correlations between each assessment type and external criteria to Fisher-*z* scores, then calculated the Pearson correlations (a) between the *z*-scores from external criteria and LBAs, and (b) between the *z*-scores from same external criteria and self-reports of personality. If two measures of the same construct have similar patterns of correlations with external criteria, then these correlations themselves should be highly correlated.

Analysis of distinctive language. Our modeling goal was predictive accuracy, and some modeling techniques, such as dimensionality reduction and ridge regression, obscure the associations between the original language and resulting trait predictions, creating a "black box" statistical model. Although lack of interpretability is not necessarily a threat to validity, we felt that a simple overview of each trait's most distinctive language features would be valuable to readers. On one hand, language features that are highly predictive of a trait should be reasonably consistent with expectations based on the patterns of thoughts, emotions, and behaviors that define each personality trait. For example, we may expect the language that predicts high extraversion to express some aspect of high sociability, enthusiasm, and/or positive affect. On the other hand, it is possible that some of the resulting predictive language may be unexpected or even run counter to theoretical expectations. In either case, we felt that a brief survey of the highly correlated language features would aid readers' understanding of our final language models.

After predicting trait values for users in the validation sample, we examined the correlations between trait predictions and relative frequencies of words, phrases, and topics. This resulted in several thousands of correlations, which we visualized in the form of word clouds. For each trait, we first selected the 100 most positively correlated and 100 most negatively correlated words and phrases. We then plotted each language feature, scaling the size of each word or phrase according to the absolute magnitude of corresponding correlation. We also colored each word to visually encode the relative frequency of each word across the entire validation sample to distinguish common and rare language.

We supplemented these word and phrase clouds with LDA language topic clouds. For each trait, we selected the six most positively and six most negatively correlated topics. Because words within topics are weighted relative to their prevalence within the topic, we scaled the size and color shade of each word according to its weight within the topic. Finally, we plotted the resulting topic clouds around the word and phrase clouds.

All correlations used to create these visualizations were also compiled as tables (see online supplement).

Test–retest reliability. To assess the reliability of the predictions, we approximated the traditional test–retest approach by generating multiple personality predictions for the same individuals using language from different time points, and then comparing within-person predictions over time. First, we split the validation sample's language into four 6-month subsets based on the timestamp of each message: Time 1 (July 2009 to December 2009), Time 2 (January 2010 to June 2010), Time 3 (July 2010 to December 2010), and Time 4 (January 2011 to June 2011). Within each subset, we identified users who had written at least 1,000 words within that 6-month interval. For users with at least 1,000 words in a given interval, we generated personality predictions using *only* the language from that interval. For example, 681 users from the validation sample wrote at least 1,000 words during both Time 1 and Time 2. For these users, we generated predictions within each interval and then calculated the correlations between predictions of the same traits (e.g., we correlated extraversion predictions from Time 1 with extraversion predictions from Time 2). We repeated this process across every pair of intervals, resulting in six test-retest correlations for each personality trait.

Across all possible comparisons, the shortest test–retest intervals were between consecutive 6-month intervals (e.g., Time 1 and Time 2, or Time 2 and Time 3, or Time 3 and Time 4); the longest test–retest interval was between Time 1 and Time 4, as the language samples from these two subsets were separated by at least 1 year (two 6-month intervals). Because users varied in their language use across intervals, sample sizes associated with these correlations also varied across intervals, ranging from $n = 331$ (users who wrote at least 1,000 words in both Time 1 and Time 4) to $n = 1{,}424$ (users who wrote at least 1,000 words in both Time 2 and Time 3).

Results

Convergent and Discriminant Validity

LBAs converged substantially with self-reports of Big Five personality traits. As shown in Table 1, mono-trait correlations (Pearson *r*s) between assessment methods were openness: $r = .43$; conscientiousness: $r = .37$; extraversion: $r = .42$; agreeableness: $r = .35$; and neuroticism: $r = .35$. The average convergent correlation was .38 across all 4,824 users. We repeated these analyses in subsets of users who completed 20- and 100-item self-report personality measures. To test for significant differences between two correlations in this and later comparisons, we used a z test for independent correlation coefficients (Preacher, 2002). Convergent validity was significantly higher in the 100-item subset (average $r = .41$) compared with the 20-item subset (average $r = .34$; $z = 2.65$, $p = .008$).

Table 1
Convergent Correlations (Pearson r) Between Language-Based Assessments and Self-Reports of Big Five Personality Traits

	Correlations with self-report questionnaires		
	All versions	20-item	100-item
Language-based assessment			
Openness	.43	.38	.46
Conscientiousness	.37	.34	.38
Extraversion	.42	.39	.41
Agreeableness	.35	.31	.40
Neuroticism	.35	.30	.39
M	.38	.34	.41

Note. *N*s = 4,824 (all versions), 2,324 (20-item), and 1,943 (100-item). Average correlations within each column are calculated by first applying Fisher's *r*-to-*z* transformation to each correlation, averaging, and transforming back to *r*. All correlations are significantly greater than zero ($p < .001$).

Patterns of discriminant validity were similar across LBAs and self-report questionnaires, although self-report questionnaires discriminated slightly better between traits. The full set of correlations between self-reports and LBAs, including discriminant validity coefficients (i.e., correlations between measures of different traits), are shown in Table 2 (these analyses were repeated for subsets of users who completed the 20- and 100-item measure, see Appendices A and B). Discriminant validity coefficients for each method are shown in italics. Among all users, the average magnitude (absolute value) of discriminant validity coefficients of LBAs was significantly higher than self-report questionnaires ($r_{LBA} = .29$, $r_{self} = .19$, $z = 5.22$, $p < .001$), indicating that LBAs were relatively worse than self-report questionnaires at discriminating between traits. However, among LBAs, convergent validity coefficients were, on average, larger than discriminant validity coefficients.

We found similar patterns of convergent and discriminant validity when comparing LBAs at the more fine-grained facet-level (see Table 3). The average magnitude of convergent correlations (i.e., correlations between a domain-level and its corresponding facet-level self-report questionnaires) was significantly greater than the average discriminant correlations ($r_{convergent} = .26$, $r_{divergent} = .10$; $z = 2.18$, $p = .029$).

Patterns of facet-level convergent correlations within each domain-level trait suggested the LBAs provide broad coverage of each domain, with a few exceptions. For example, convergent correlations for the facets of cautiousness ($r =$. 08) and immoderation ($r = .10$) were noticeably smaller than for other facets in their respective domains of conscientiousness and neuroticism.

Comparison With Informant Reports

On average, LBAs were similar in agreement (or accuracy) with informant reports of personality (average self-LBA agreement, $r = .39$; average self-informant agreement, $r = .32$; $z = 1.54$, $p = .12$). Table 4 lists correlations between LBAs, self-reports, and informant reports of personality. LBAs were significantly more accurate than informants for openness ($z = 4.66$, $p < .001$). LBAs were only slightly more accurate than informants for agreeableness ($z =$

Table 2
Correlations Between Language-Based Assessments and Self-Reports of Big Five Personality Traits

	Self-reports					Language-based assessments				
	O	C	E	A	N	O	C	E	A	N
Self-reports										
Openness										
Conscientiousness	*.00*									
Extraversion	*.13*	*.19*								
Agreeableness	*.07*	*.17*	*.19*							
Neuroticism	*−.08*	*−.31*	*−.34*	*−.36*						
Language-based										
Openness	**.43**	−.12	−.08	−.05	.00					
Conscientiousness	−.13	**.37**	.16	.17	−.17	*−.25*				
Extraversion	−.07	.12	**.42**	.10	−.15	*−.17*	*.33*			
Agreeableness	−.07	.17	.13	**.35**	−.14	*−.12*	*.44*	*.27*		
Neuroticism	.05	−.17	−.18	−.13	**.35**	*.06*	*−.41*	*−.43*	*−.34*	

Note. $N = 4{,}824$. O = Openness to Experience; C = Conscientiousness; E = Extraversion; A = Agreeableness; N = Neuroticism. Convergent correlations are in bold; discriminant correlations are in italics.

1.74, $p = .081$); LBAs and informants were similar in accuracy for conscientiousness, extraversion, and neuroticism.

For comparison, the average self-LBA agreement ($r = .39$) is somewhat lower than the self-informant correlations typically found in studies using informant reports (self-informant agreement *r*s often range from .40 to .60; e.g., Vazire, 2006; Watson, Hubbard, & Wiese, 2000), suggesting that LBAs predict self-reports slightly worse than well-acquainted informants. Also, the average self-informant agreement ($r = .32$) was substantially lower than the agreement typically found with informants. The relatively low self-informant agreement in our study was likely due to the use of short, two-item informant scales.

We found substantial agreement between informant reports and LBAs (average $r = .24$), which suggests that the trait variance captured by LBAs overlaps with an outsider's perspective and is not unique to the self. To quantify the unique contribution of LBAs over informants, we calculated the partial correlations between LBAs and self-reports of each trait, controlling for informant reports. We repeated this procedure for the informant reports, controlling for LBAs. In each case, substantial partial correlations remained, suggesting that LBAs and informants have unique predictive validity.

Finally, aggregate ratings (the average of LBAs and informant reports) were consistently more accurate than informant ratings alone ($p < .001$ in each comparison) and more accurate than LBAs for all traits but openness ($p < .01$ in the remaining four comparisons).

External Correlates

To assess criterion-related validity, we compared the correlations between several relevant external criteria and (a) self-report questionnaires of personality, and (b) LBAs of personality. Figure 2 shows scatterplots of correlations with 14 external criteria. All correlations, including average correlations within measures and column-vector correlations between measures, are listed in Appendix C.

We first compared the signs within pairs of correlations (both assessments and their correlations with the same criteria). Across 70 correlation pairs, 60 shared the same sign. Among the 10 that differed in sign, the correlations tended to be close to zero. The largest discrepancies were correlations between measures of conscientiousness and self-reported recent physician visits ($r_{self} = -.05$, $r_{LBA} = .12$) and measures of openness and informant-reported extraversion ($r_{self} = .05$, $r_{LBA} = -.07$).

With few exceptions, the correlations between self-reports and external criteria were greater than those between LBAs and external criteria. This is not surprising, as the external criteria were self-reported and share method variance with the self-reported measures. What is particularly striking is that LBAs were predictive of these external criteria, without the shared variance. For example, the correlation between self-report questionnaires of extraversion and life satisfaction was $r = .24$, significantly greater than the correlation between language-based extraversion and life satisfaction, $r = .13$; $t(1{,}079) = 3.46$, $p < .001$. In 21 of 70 correlation pairs, the magnitude of the correlations between self-report questionnaires and the criterion was significantly larger than those from LBAs (at $p < .05$). This suggests that self-report questionnaires of personality shared greater variance with self-reported external criteria than LBAs.

Finally, we summarized the similarities between assessments using column-vector correlations, which ranged from $r = .83$ (openness) to $r = .96$ (neuroticism). Each column-vector correlation is listed in the lower right corner of each scatterplot in Figure 2. In general, larger correlations between a self-report questionnaires and an external criterion should be paired with relatively larger correlations between an LBA and the same criteria. This was the case across most pairs. For example, both measures of openness were moderately correlated with self-reported liberal political attitudes ($r_{self} = .32$, $r_{LBA} = .22$), and measures of extraversion were similarly correlated with number of Facebook friends ($r_{self} = .18$, $r_{LBA} = .23$).

Analysis of Distinctive Language

The most highly correlated words, phrases, and language topics with predictions of each trait were consistent with the patterns of thought, feelings, and behaviors that characterize each Big Five

Table 3
Correlations Between Language-Based Assessments and Self-Reports of Facet-Level Big Five Personality Traits

Self-reported questionnaire	Language-based assessments				
	O	C	E	A	N
Openness	***.41***	*−.12*	*.00*	*−.08*	*.01*
Liberalism	**.33**	−.23	−.02	−.14	.08
Intellect	**.34**	−.12	−.04	−.11	−.08
Adventurousness	**.12**	.01	.20	−.01	−.15
Emotionality	**.17**	.09	.05	.08	.13
Artistic interests	**.27**	.03	.12	.16	.04
Imagination	**.31**	−.24	−.03	−.15	.07
Conscientiousness	*−.03*	***.26***	*.20*	*.17*	*−.16*
Cautiousness	.02	**.08**	−.03	.11	−.03
Self-discipline	−.04	**.25**	.20	.15	−.13
Achievement-striving	.05	**.29**	.26	.16	−.14
Dutifulness	.01	**.19**	.01	.26	−.11
Orderliness	.00	**.14**	.05	.10	−.04
Self-efficacy	.03	**.18**	.26	.06	−.26
Extraversion	*−.05*	*.12*	***.36***	*.07*	*−.11*
Cheerfulness	.03	.07	**.21**	.10	−.15
Excitement seeking	−.02	−.14	**.20**	−.13	−.02
Activity level	−.04	.23	**.22**	.09	−.04
Assertiveness	−.08	.12	**.28**	−.09	−.11
Gregariousness	−.08	.02	**.30**	.07	−.08
Friendliness	−.04	.14	**.30**	.17	−.12
Agreeableness	*.02*	*.25*	*.10*	***.41***	*−.15*
Sympathy	.17	.13	−.05	**.20**	.04
Modesty	−.01	.01	−.17	**.18**	.16
Cooperation	.08	.21	.04	**.41**	−.15
Altruism	.02	.23	.17	**.28**	−.07
Morality	−.01	.17	−.06	**.31**	−.05
Trust	.00	.23	.13	**.36**	−.18
Neuroticism	*−.01*	*−.14*	*−.20*	*−.20*	***.39***
Vulnerability	.01	−.11	−.17	−.03	**.34**
Immoderation	−.05	−.05	.05	−.11	**.10**
Self-consciousness	.02	−.12	−.33	.01	**.24**
Depression	.00	−.22	−.28	−.20	**.37**
Anger	−.09	−.09	−.11	−.24	**.33**
Anxiety	.02	−.06	−.15	−.11	**.35**

Note. $N = 348$. O = Openness to Experience; C = Conscientiousness; E = Extraversion; A = Agreeableness; N = Neuroticism. Convergent correlations are bolded. Domain-level correlations are italicized.

trait. As an example, Figure 3 shows language most correlated with LBAs of extraversion, providing a simple overview of the language features that were common among those with high and low predicted extraversion. In other words, Figure 3 displays the most distinctive language of those who were predicted as high or low on extraversion. Figures for the remaining four traits are in Appendix D. A full exploration of these associations is beyond the scope of this study, but many of the patterns seen here overlap heavily with more detailed descriptive analyses of language and personality that use similar underlying methods (see Kern et al., 2014; Schwartz et al., 2013b).

Aspects of high extraversion are evident in the left panel of Figure 3, including language reflecting positive emotion (e.g., *love*, *:)*, *<3*), enthusiasm (e.g., *best*, *stoked*, *pumped*), and sociability (e.g., *party*, *hanging*, *dinner with*). On the other end, the language of low extraversion (introverts) suggested a more inward focus (e.g., *i've*, *i don't*, *i should*), relatively greater interest in things (vs. people; e.g., *computer*, *book*, *chemistry*), and tentativeness (e.g., *probably*, *suppose*, *apparently*). The absolute magnitude of the correlations between the language features and predicted extraversion in Figure 3 ranged from $r = .13$ to $r = .33$, and all correlations were significant after Bonferroni-correction ($p < .0001$). Comprehensive lists of all language features and correlations (and associated p values) used to create these figures are available as supplementary material.

Test–Retest Stability

LBAs were stable across 6-month intervals, with average test–retest correlations across consecutive 6-month intervals (Time 1–Time 2, Time 2–Time 3, and Time 3–Time 4) among openness: $r = .74$; conscientiousness: $r = .76$; extraversion: $r = .72$; agreeableness: $r = .65$; and neuroticism: $r = .62$. The average test-retest correlation of all five traits across consecutive 6-month intervals was $r = .70$. Average test–retest correlations between all pairs of intervals are shown in Table 5 (see Appendix E for corresponding tables for single traits). Test-retest correlations were attenuated as intervals were spaced farther apart. For comparison, reported test-retest correlations of Big Five self-report questionnaires typically range from .65 to .85 and increase with scale length and shorter retest intervals (e.g., Donnellan, Oswald, Baird, & Lucas, 2006; Gosling, Rentfrow, & Swann, 2003; John, Naumann, & Soto, 2008; Kosinski, Stillwell, & Graepel, 2013; Rammstedt & John, 2007).

Discussion

Social media language is rich in psychological content and can be leveraged to create a fast, valid, and stable personality assessment. Our method resulted in state-of-the-art accuracy compared with other language-based predictive models. Comparisons with informant reports and external criteria suggested that language based assessments (LBAs) are capable of capturing true personality variance. Predictions were stable over time, with test–retest correlations on par with self-report questionnaires of personality.

LBAs may complement and extend traditional measures in social media samples by providing an alternative to self-report question-

Table 4
Correlations Between Self-Reports, Informant Reports, and Language-Based Assessments of Big Five Personality Traits

	LBA and self		Informant and self		LBA and informant	LBA + Informant and self
	r	partial r[a]	r	partial r[b]	r	r
Openness	.46	.42	.25	.13	.30	.44
Conscientiousness	.34	.30	.30	.26	.20	.42
Extraversion	.43	.37	.39	.32	.24	.52
Agreeableness	.38	.34	.30	.24	.24	.44
Neuroticism	.35	.31	.34	.29	.20	.44
M	.39	.35	.32	.25	.24	.45

Note. $N = 745$. LBA = language-based assessment; LBA + Informant = aggregate ratings from informant reports and language-based assessment; M = column average correlation. [a] Partial correlation between language-based assessments and self-reports, partialling out informant reports. [b] Partial correlation between informant reports and self-reports, partialling out language-based assessments. Average correlations within each column are calculated by first applying Fisher's r-to-z transformation to each correlation, averaging, and transforming back to r. All correlations are significantly greater than zero ($p < .001$).

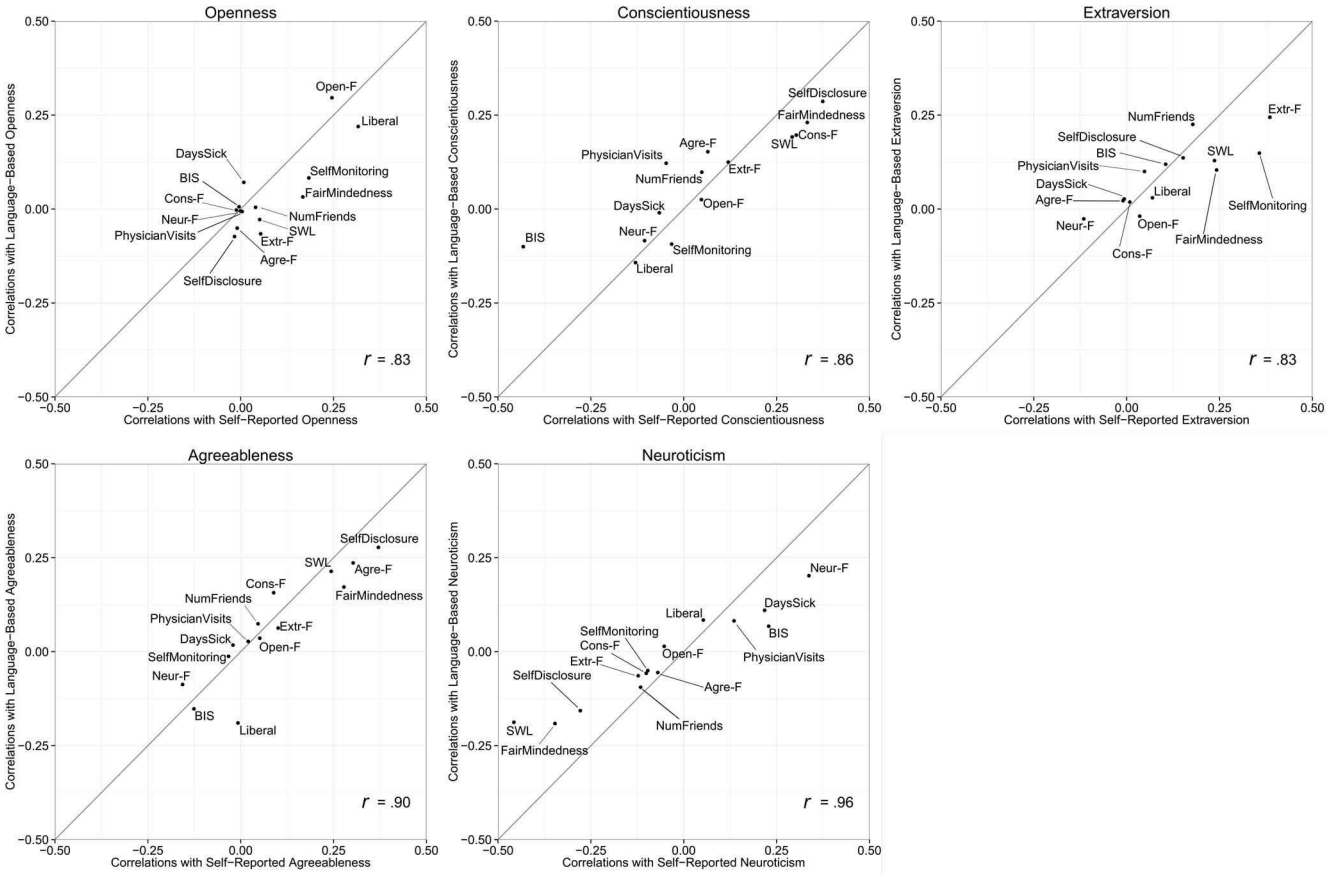


Figure 2. Scatterplots of correlations between external criteria and two assessment methods. The Pearson *r*s in the lower right of each scatterplot indicate the correlation between methods, calculated after applying Fisher's *r*-to-*z* transformation to the original measurement-external criteria correlations. Agre-F = friend-reported agreeableness; BIS = Barratt Impulsiveness Scale; Cons-F = friend-reported conscientiousness; DaysSick = self-reported number of recent sick days; Extr-F = friend-reported extraversion; FairMinded = Fair-mindedness subscale; Liberal = self-reported liberal political attitudes; Neur-F = friend-reported neuroticism; NumFriends = number of Facebook friends; Open-F = friend-reported openness; PhysicianVisits = self-reported number of recent physician visits; SelfMonitoring = Self-Monitoring Scale; SWL = Satisfaction with Life.

naires. Vazire (2006) noted the dominance of self-report questionnaires in personality research and urged researchers to consider informant reports (e.g., personality ratings from well-acquainted others) for several reasons: they are relatively fast and cheap, they avoid some biases of self-report questionnaires, and they agree with self-report questionnaires. Our results suggest that LBAs share these advantages and can improve accuracy over single informant reports.

Compared with self-report questionnaires, LBAs are extremely fast. The entire validation sample, roughly 5,000 participants, was assessed in minutes. The majority of processing time and resources was spent on the initial model building process. After training and validating a model, application of the model to a new user's language data only takes seconds.

New self-report methods are easily shared among researchers; LBAs are sharable as computer code, but application requires some specialized knowledge. Alternatively, LBAs can be distributed through a Web site interface or as weighted lexica (i.e., a list of words and phrases with corresponding regression weights). Although none of these options can match the simplicity of traditional self-report questionnaires, researchers with access to large social media datasets may be willing to trade simplicity for the speed and scale of LBAs.

Because they are derived from a target's language in a social setting, LBAs share some features of self-report (Paulhus & Vazire, 2007). To the extent that targets are aware of their own self-presentation through language, LBAs may incur biases inherent in self-reports more broadly: they are limited by a target's self-presentation and motivation to disclose information. Most self-report methods are also constrained by the target's memory, and researchers may mistrust the accuracy of retrospective self-reports (Lucas & Baird, 2006). In contrast, an advantage of LBAs is that they can be generated retroactively, giving researchers an alternative method to study past behavior without relying on participants' memories.

Statistical Language Models as a Judge

How can personality traits be accurately judged from a statistical language model? Funder's Realistic Accuracy Model (RAM; Funder, 1999, 2012) was developed to explain the accuracy of trait

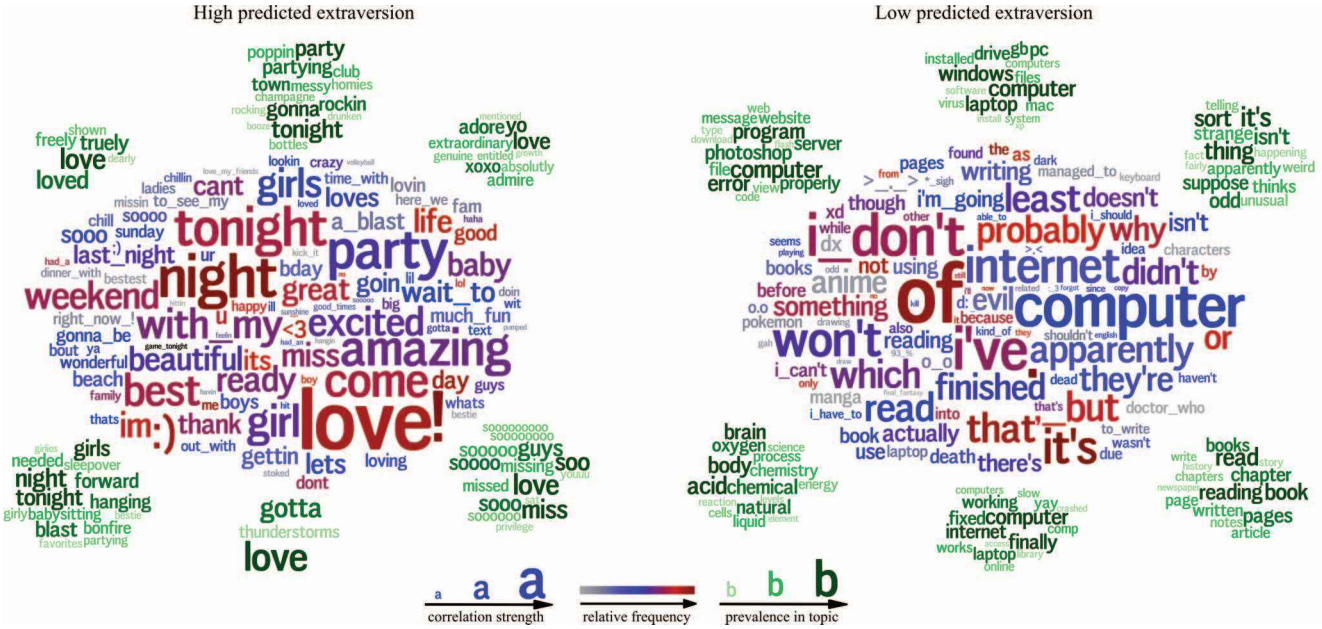


Figure 3. Words, phrases, and topics with the strongest correlations to extraversion, as predicted by language (N = 4,824). Large central word clouds (red, blue, and gray) contain the 100 words and phrases with highest correlations with high and low predicted extraversion. Word size is proportional to correlation size; color indicates word frequency. Underscores (_) are used to connect words within phrases and do not occur in the original text. The smaller surrounding word clouds (green) are the six most highly correlated topics, or clusters of semantically related words. Within topics, word size and color indicate word prevalence. All correlations are significant ($p < .001$).

predictions made by human judges, but it applies equally well to nonhuman judges. According to the RAM, accurate judgment requires that the target emits cues that are (a) relevant to the trait, (b) available to the judge, (c) detected by the judge, and (d) used correctly by the judge. Final accuracy of a judge is moderated by cue relevance, the number of available cues, the judge's capacity to detect available cues, and the judge's ability to properly use these cues. Viewing our approach in the context of the RAM is useful for understanding its relative accuracy and identifying methods for improving accuracy.

First, language is a particularly rich source of *relevant* trait cues (Tausczik & Pennebaker, 2010); it has been used to accurately predict personality by both human (Mehl et al., 2006) and automated judges (e.g., Iacobelli et al., 2011; Mairesse, Walker, Mehl, & Moore, 2007; Schwartz et al., 2013b; Sumner et al., 2012). Language from social media may be particularly relevant due to the unusually high level of self-disclosure evidenced in users (Naaman et al., 2010).

Relevant cues must be extracted from language and made *available* to the judge. Compared with closed-vocabulary methods, the open-vocabulary approach to linguistic feature extraction greatly increases the judge's amount of available cues. Still, our approach was by no means exhaustive. There are likely additional layers of relevant cues in language untouched by our approach, including syntactical, grammatical, and stylistic features.

Using a large sample size in the training phase increased the likelihood that subtle but relevant cues were *detected*. This is particularly useful when cues are rare but highly informative. For example, Schwartz et al. (2013b) found that the words *fucking* and *depression* were both highly correlated with neuroticism, but *depression* is used far less frequently. Learning the relationship between a relatively rare word like *depression* and neuroticism requires exposure to many more examples. By training a model over tens of thousands of examples and millions of words, statistical models can develop the necessary expertise to detect such rare but high-signal cues.

Finally, our statistical modeling process may be interpreted as method of optimizing cue *utilization*. The model building process detected relevant cues, removed irrelevant cues (feature selection), combined redundant cues (dimension reduction), and then optimized the weight of each cue for the judgment (regression). We

Table 5
Average Test–Retest Correlations of Language-Based Assessments of Big Five Personality

	Time 2	Time 3	Time 4
Time 1	.69 (681)	.66 (625)	.61 (331)
Time 2		.70 (1,424)	.65 (680)
Time 3			.71 (1,019)

Note. Time 1 = July 2009 to December 2009; Time 2 = January 2010 to June 2010; Time 3 = July 2010 to December 2010; Time 4 = January 2011 to June 2011. Average test–retest correlations are based on the average test–retest correlation across all five traits for each pair of intervals. Correlations were transformed to Fisher-z scores prior to averaging and then the average was transformed back to r. Sample sizes for each correlation are shown in parentheses.

used a relatively simple statistical model with the final feature set. More sophisticated modeling approaches (e.g., including interactions, ensembling multiple models) may improve sensitivity and accuracy while using the same cues.

Limitations and Future Directions

Our sample was limited in several ways. It was drawn from users with sufficient language within the larger sample of myPersonality application users on Facebook, which is a subset of the broader population of social media users. The available personality measures from this application were limited to the Big Five framework. Language models built within this specific language context may not generalize well to samples outside social media or even outside the context of Facebook status messages. Additional validation is necessary before these models can be applied in different contexts (e.g., other Facebook language outside of status messages, Twitter messages, or language sources outside of social media). Alternatively, the method described here can be adapted to alternative language contexts.

We demonstrated evidence of criterion validity by correlating LBAs with external criteria. In our comparison between LBAs and self-report personality measures, we found many similarities in their correlations with external criteria, although self-report questionnaires generally correlated more strongly with these criteria. Almost all of the external criteria available shared method variance with the self-reported personality measures. We could not determine whether the higher correlations between self-report personality measures and self-reported external criteria were due to overlapping trait variance or method variance. Future validations of LBAs should use additional external criteria that were collected outside of the computer-administered self-report context (e.g., observed behavior).

LBAs had significantly lower discriminant validity than those typically seen among self-report measures. Although discriminant correlations were smaller than convergent correlations, these differences were small (across all users, the mean discriminant correlation was .29; the mean convergent correlation was .38). Discriminant validity was poorest among LBAs of socially desirable personality traits such as conscientiousness and agreeableness. Although these traits are typically correlated regardless of method, the intercorrelation was significantly higher in LBAs. One explanation for this may be the common linguistic correlates of these traits: both traits are correlated with positive (e.g., "great," "wonderful") and negative (e.g., "damn," "bullshit") evaluations, as seen in Appendix D. Because our central goal was high predictive accuracy (convergence with self-reports), we used all informative language features when building LBAs. As a result, LBAs for any given trait often shares many language features with the LBA for a different trait, which may decrease discriminant validity. One could potentially increase discriminant validity of LBAs by filtering out these shared language features, but this would likely decrease predictive accuracy. In some applications, this may be a worthwhile tradeoff and should be considered in future work.

Convergent correlations between LBAs and self-report questionnaires of personality averaged $r = .38$, which is lower than those typically observed with novel self-report questionnaires (where *r*s typically exceed .70; e.g., Donnellan et al., 2006; Gosling et al., 2003) or informant reports (where *r*s range between .40 and .60; Vazire, 2006; Watson et al., 2000). The unreliability of the accuracy criteria (self-report questionnaires) may place a limitation on convergence. We found some evidence for this hypothesis: convergence between self-report questionnaires and LBAs was lowest when we used a 20-item measure as the criterion, and convergence was highest when using the more reliable 100-item measure. On the other hand, convergence was not higher when using the 336-item measure, so longer criterion measures do not always result in higher convergence.

Finally, we limited the accuracy criteria to self-report personality measures when building our language models. We did this for practical reasons: self-report questionnaires are the most widely used and accepted assessment method. However, alternative methods such as informant reports provide a unique "outsider" perspective, which avoids some biases and can more accurately assesses some aspects of personality than the self (Hofstee, 1994; Kolar et al., 1996; Vazire, 2010; Vazire & Mehl, 2008). Specifically, informants can be more accurate judges of highly visible and socially desirable traits (e.g., attractiveness, intelligence; Vazire & Carlson, 2011), and they may have a similar advantage in judging traits such as agreeableness and conscientiousness. For these traits, informant reports could constitute an alternative, more accurate criterion from which to build a language model.

Potential Applications

In this study, we used language to assess Big Five personality traits, but LBAs are not limited to personality. This same method can be adapted to create language models of other psychological characteristics, including psychological well-being, attitudes, traits in other personality frameworks (e.g., HEXACO; Ashton & Lee, 2007), and more temporary states such as mood, provided that the training data includes a valid measure of the target criterion. For example, Schwartz et al. (2013a) illustrated how the language from Twitter can be used to predict the average life satisfaction of U.S. counties. Refinement and further validation of these models could lead to LBAs of county-level life satisfaction and other characteristics, providing a fast and inexpensive complement to traditional survey methods.

Questionnaires can be expensive to administer and time and resource intensive. LBAs offer a practical, cost-effective alternative, allowing assessment of psychological characteristics when questionnaires are impractical. Researchers could reduce participant burden by replacing some questionnaires with a single link and sign-in procedure, allowing a research application to access participant social media language and quickly assess personality and other characteristics of interest. Alternatively, LBAs can be used to complement self- and informant reports, adding an additional measurement for multimethod study designs. The combination of reduced costs and fast assessments may offer one route to collecting samples much larger than those typically possible with traditional methods.

Combining LBAs with other features of social media data may also enable new approaches to studying geographic and temporal trends. With permission by the user, social media messages are often tagged with geographic and precise tempo-

ral metadata, providing unobtrusive measures of when and where a message was created. LBAs may provide a means to compare regional psychological differences and track psychological trends over time. Given sufficient language data from individuals, LBA may be used to study interesting within-person variation, such as patterns of psychological change over time or across locations.

Finally, a hybrid approach that combines LBAs with other rich nonverbal data sources from social media (e.g., images, preferences, social network characteristics, etc.) would likely improve predictive performance. Kosinski, Stillwell, and Graepel (2013) found that Facebook users' personality traits and other characteristics could be accurately predicted using only users' preferences or "likes." Even models built only on social network behavior, such as message frequency and message response time, have been useful in predicting users' personalities (Adali & Golbeck, 2014). Provided that each source has some unique contribution to a target trait, models combining multiple sources in addition to language may provide even better assessments.

Conclusion

In this article, we provided evidence that the language in social media can be harnessed to create a valid and reliable measure of personality. This approach is just one example of how social media can extend assessment to many more people—quickly, cheaply, and with low participant burden. Moreover, this illustrates how computational techniques can reveal new layers of psychological richness in language. Combining these techniques with psychological theory may complement existing measures, as argued here. But even more generally, using these techniques to study the words and phrases through which people express themselves, as well as their change over time, may provide us with a clearer portrait of their unfolding mental life.

References

Adali, S., & Golbeck, J. (2014). Predicting personality with social behavior: A comparative study. *Social Network Analysis and Mining, 4,* 159. http://dx.doi.org/10.1007/s13278-014-0159-7

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 11,* 150–166. http://dx.doi.org/10.1177/1088868306294907

Atkins, D. C., Rubin, T. N., Steyvers, M., Doeden, M. A., Baucom, B. R., & Christensen, A. (2012). Topic models: A novel method for modeling couple and family text data. *Journal of Family Psychology, 26,* 816–827. http://dx.doi.org/10.1037/a0029607

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,* 372–374. http://dx.doi.org/10.1177/0956797609360756

Blei, D. M., Ng., A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3,* 993–1022.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16,* 22–29.

Cohen, A. S., Minor, K. S., Baillie, L. E., & Dahir, A. M. (2008). Clarifying the linguistic signature: Measuring personality from natural speech. *Journal of Personality Assessment, 90,* 559–563. http://dx.doi.org/10.1080/00223890802388459

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (Neo-PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional manual.* Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J., & Meehl, P. E. (1955). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 52,* 281–302. http://dx.doi.org/10.1037/h0040957

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49,* 71–75. http://dx.doi.org/10.1207/s15327752jpa4901_13

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment, 18,* 192–203. http://dx.doi.org/10.1037/1040-3590.18.2.192

Fast, L. A., & Funder, D. C. (2008). Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior. *Journal of Personality and Social Psychology, 94,* 334–346. http://dx.doi.org/10.1037/0022-3514.94.2.334

Funder, D. C. (1999). *Personality judgment: A realistic approach to person perception.* San Diego, CA: Academic Press.

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science, 21,* 177–182. http://dx.doi.org/10.1177/0963721412445309

Golbeck, J., Robles, C., & Turner, K. (2011, May). *Predicting personality with social media.* In Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems - CHI '11, Vancouver, BC, 253–262.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public domain personality measures. *Journal of Research in Personality, 40,* 84–96. http://dx.doi.org/10.1016/j.jrp.2005.08.007

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37,* 504–528. http://dx.doi.org/10.1016/S0092-6566(03)00046-1

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21,* 267–297. http://dx.doi.org/10.1093/pan/mps028

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3,* 1157–1182.

Hirsh, J. B., & Peterson, J. B. (2009). Personality and language use in self-narratives. *Journal of Research in Personality, 43,* 524–527. http://dx.doi.org/10.1016/j.jrp.2009.01.006

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12,* 55–67. http://dx.doi.org/10.1080/00401706.1970.10488634

Hofstee, W. K. (1994). Who should own the definition of personality? *European Journal of Personality, 8,* 149–162. http://dx.doi.org/10.1002/per.2410080302

Holtgraves, T. (2011). Text messaging, personality, and the social context. *Journal of Research in Personality, 45,* 92–99. http://dx.doi.org/10.1016/j.jrp.2010.11.015

Iacobelli, F., Gill, A. J., Nowson, S., & Oberlander, J. (2011). Large scale personality classification of bloggers. In S. D'Mello, A. Graesser, B. Schuller, & J. Martin (Eds.), *Proceedings of the 4th international conference on affective computing and intelligent interaction* (pp. 568–577). New York, NY: Springer-Verlag. http://dx.doi.org/10.1007/978-3-642-24571-8_71

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114–158). New York, NY: Guilford Press.

Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Dziurzynski, L., Ungar, L. H., Stillwell, D. J., . . . Seligman, M. E. (2014). The online social self: An open vocabulary approach to personality. *Assessment, 21,* 158–169. http://dx.doi.org/10.1177/1073191113514104

Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of Personality, 64,* 311–337. doi:10.1111/j.1467-6494.1996.tb00513.x

Kosinski, M., & Stillwell, D. J. (2011). *myPersonality Research Wiki. myPersonality Project*. Retrieved from http://mypersonality.org/wiki

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

Lee, C. H., Kim, K., Seo, Y. S., & Chung, C. K. (2007). The relations between personality and language use. *Journal of General Psychology, 134,* 405–413. http://dx.doi.org/10.3200/GENP.134.4.405-414

Lucas, R. E., & Baird, B. M. (2006). Global self-assessment. In E. Diener & M. Eid (Eds.), *Handbook of multimethod measurement in psychology* (pp. 29–42). Washington, DC: American Psychological Association. doi:10.1037/11383-003

MacCallum, A. K. (2002). *MALLET: A machine learning for language toolkit*. Retrieved from http://mallet.cs.umass.edu

Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research, 30,* 457–500.

Martinsson, P. G., Rokhlin, V., & Tygert, M. (2011). A randomized algorithm for the decomposition of matrices. *Applied and Computational Harmonic Analysis, 30,* 47–68. http://dx.doi.org/10.1016/j.acha.2010.02.003

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,* 862–877. http://dx.doi.org/10.1037/0022-3514.90.5.862

Naaman, M., Boase, J., & Lai, C. H. (2010, February). Is it really about me?: Message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer Supported Cooperative Work* (pp. 189–192). Retrieved from http://luci.ics.uci.edu/predeployment/websiteContent/weAreLuci/biographies/faculty/djp3/LocalCopy/p189-naaman.pdf

O'Connor, B., Bamman, D., & Smith, N. A. (2011, December). Computational text analysis for social science: Model assumptions and complexity. In *Second Workshop on Computational Social Science and Wisdom of the Crowds*. Retrieved from http://www.cs.cmu.edu/~nasmith/papers/oconnor+bamman+smith.nips-ws11.pdf

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt impulsiveness scale. *Journal of Clinical Psychology, 51,* 768–774. http://dx.doi.org/10.1002/1097-4679(199511)51:6<768::AID-JCLP2270510607>3.0.CO;2-1

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). New York, NY: Guilford Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12,* 2825–2830.

Pennebaker, J. W. (1982). *The psychology of physical symptoms*. New York, NY: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4613-8196-9

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007*. Austin, TX: LIWC. net.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77,* 1296–1312. http://dx.doi.org/10.1037/0022-3514.77.6.1296

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54,* 547–577. http://dx.doi.org/10.1146/annurev.psych.54.101601.145041

Potts, C. (2011). *happyfuntokenizer* (Version 10). [Computer software]. Retrieved from http://sentiment.christopherpotts.net/code-data/happyfuntokenizing.py

Preacher, K. J. (2002, May). Calculation for the test of the difference between two independent correlation coefficients [Computer software]. Retrieved from http://quantpsy.org

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,* 203–212. http://dx.doi.org/10.1016/j.jrp.2006.02.001

Reuters. (2013). *Twitter Incorporated company profile*. Retrieved from http://www.reuters.com/finance/stocks/companyProfile?symbol=TWTR.K

Revelle, W. (2014). *psych: Procedures for Personality and Psychological Research* (Version 1.4.1) [Computer software]. Retrieved from http://CRAN.R-project.org/package=psych

Rust, J., & Golombok, S. (2009). *Modern psychometrics: The science of psychological assessment*. New York, NY: Routledge.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Agrawal, M., Park, G., . . . Lucas, R. E. (2013a, June). *Characterizing geographic variation in well-being using tweets*. In *Seventh International AAAI Conference on Weblogs and Social Media,* Boston, MA.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., . . . Ungar, L. H. (2013b). Personality, gender, and age in the language of social media: The open vocabulary approach. *PLOS ONE, 8,* e73791.http://dx.doi.org/10.1371/journal.pone.0073791

Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology, 30,* 526–537. http://dx.doi.org/10.1037/h0037039

Stanford, M. S., Mathias, C. W., Dougherty, D. M., Lake, S. L., Anderson, N. E., & Patton, J. H. (2009). Fifty years of the Barratt Impulsiveness Scale: An update and review. *Personality and Individual Differences, 47,* 385–395. http://dx.doi.org/10.1016/j.paid.2009.04.008

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245–251. http://dx.doi.org/10.1037/0033-2909.87.2.245

Sumner, C., Byers, A., Boochever, R., & Park, G. (2012, December). *Predicting dark triad personality traits from Twitter and a linguistic analysis of tweets*. Paper presented at the International Conference on Machine Learning and Applications, Boca Raton, FL.

Tausczik, Y., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29,* 24–54. http://dx.doi.org/10.1177/0261927X09351676

Vazire, S. (2006). Informant reports: A cheap, fast, and easy method for personality assessment. *Journal of Research in Personality, 40,* 472–481. http://dx.doi.org/10.1016/j.jrp.2005.03.003

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

Vazire, S., & Carlson, E. N. (2011). Others sometimes know us better than we know ourselves. *Current Directions in Psychological Science, 20,* 104–108. http://dx.doi.org/10.1177/0963721411402478

Vazire, S., & Mehl, M. R. (2008). Knowing me, knowing you: The accuracy and unique predictive validity of self-ratings and other-ratings of daily behavior. *Journal of Personality and Social Psychology, 95,* 1202–1216. http://dx.doi.org/10.1037/a0013314

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

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality, 44,* 363–373. http://dx.doi.org/10.1016/j.jrp.2010.04.001

Appendix A

Correlations Between Language-Based Assessments and 20-Item Self-Reports

	Self-reports (20-item version)					Language-based assessments				
	O	C	E	A	N	O	C	E	A	N
Self-reports										
Openness										
Conscientiousness	*−.03*									
Extraversion	*.13*	*.12*								
Agreeableness	*.05*	*.14*	*.15*							
Neuroticism	*−.04*	*−.25*	*−.31*	*−.32*						
Language-based										
Openness	**.38**	−.11	−.04	−.06	.03					
Conscientiousness	−.12	**.34**	.12	.13	−.12	*−.21*				
Extraversion	−.03	.07	**.39**	.10	−.16	*−.08*	*.28*			
Agreeableness	−.05	.14	.09	**.31**	−.09	*−.10*	*.41*	*.23*		
Neuroticism	.05	−.12	−.17	−.12	**.30**	*.01*	*−.39*	*−.41*	*−.33*	

Note. N = 2,324. O = Openness to Experience; C = Conscientiousness; E = Extraversion; A = Agreeableness; N = Neuroticism. Convergent correlations are in bold; discriminant correlations are in italics.

Appendix B

Correlations Between Language-Based Assessments and 100-Item Self-Reports

	Self-reports (100-item version)					Language-based assessments				
	O	C	E	A	N	O	C	E	A	N
Self-reports										
Openness										
Conscientiousness	*.07*									
Extraversion	*.22*	*.25*								
Agreeableness	*.13*	*.20*	*.25*							
Neuroticism	*−.14*	*−.36*	*−.41*	*−.42*						
Language-based										
Openness	**.46**	−.11	−.05	−.04	−.03					
Conscientiousness	−.09	**.38**	.17	.22	−.20	*−.25*				
Extraversion	−.04	.16	**.41**	.12	−.15	*−.16*	*.37*			
Agreeableness	−.04	.18	.16	**.40**	−.19	*−.10*	*.45*	*.30*		
Neuroticism	.03	−.20	−.18	−.14	**.39**	*.06*	*−.42*	*−.42*	*−.34*	

Note. N = 1,943. O = Openness to Experience; C = Conscientiousness; E = Extraversion; A = Agreeableness; N = Neuroticism. Convergent correlations are in bold; discriminant correlations are in italics.

(*Appendices continue*)

Appendix C

External Correlates

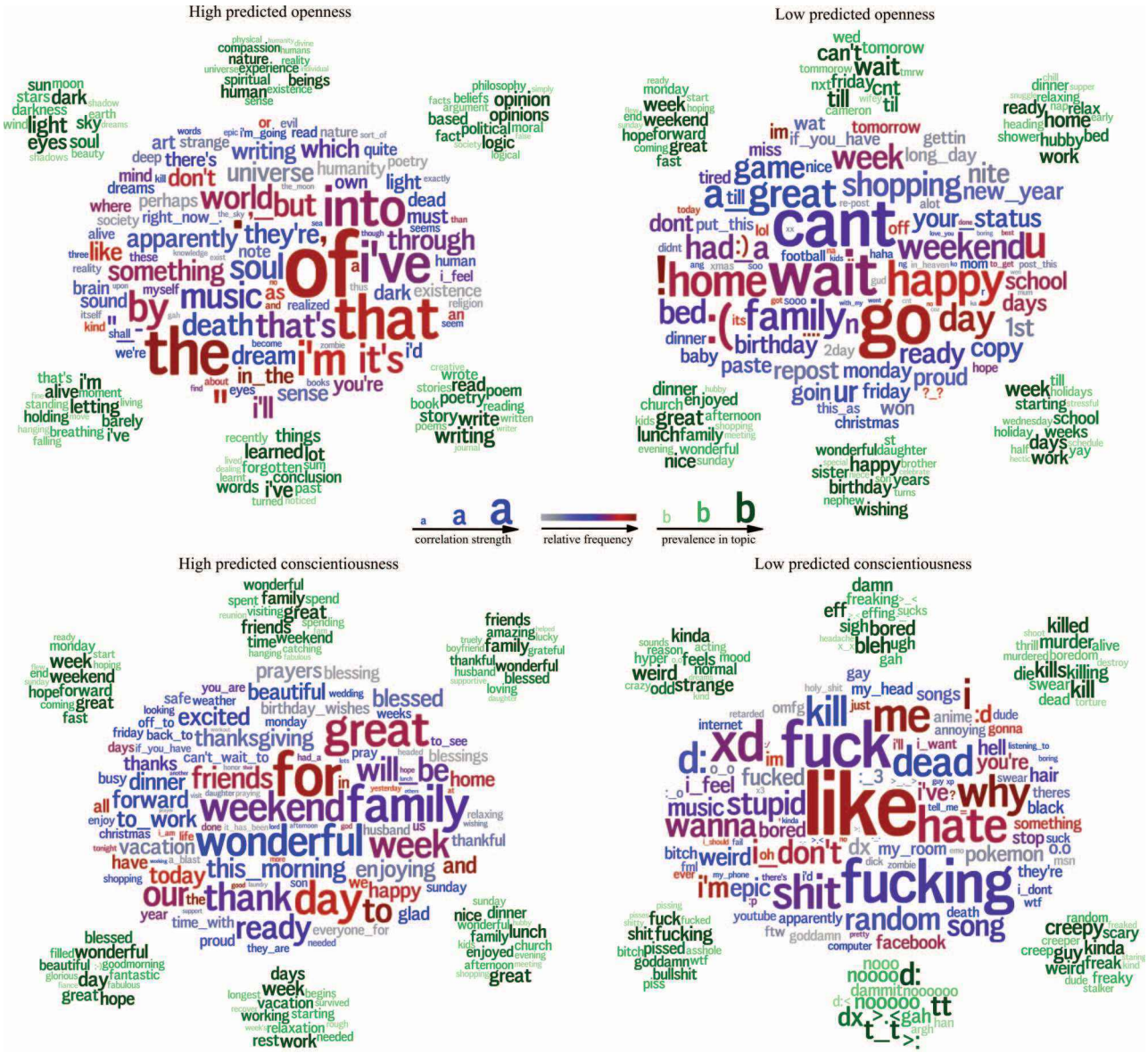
External criterion	*N*	Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism	
		SR	LBA	SR	LBA	SR	LBA	SR	LBA	SR	LBA
Satisfaction with life	1,082	.05	−.03	.29	.19	.24	.13	.24	.21	−.46	−.19
Self-monitoring	927	.18	.08	−.03	−.09	.36	.15	−.03	−.01	−.10	−.05
Fair mindedness	864	.17	.03	.33	.23	.24	.10	.28	.17	−.35	−.19
Self disclosure	864	−.02	−.07	.37	.29	.15	.14	.37	.28	−.28	−.16
Recent physician visits	736	.00	−.01	−.05	.12	.05	.10	.02	.03	.14	.08
Recent days sick	733	.01	.07	−.07	−.01	−.01	.03	−.02	.02	.22	.11
Barratt Impulsiveness Scale	549	.00	.01	−.43	−.10	.10	.12	−.13	−.15	.23	.07
Number of Facebook friends	1,842	.04	.00	.05	.10	.18	.23	.05	.07	−.12	−.09
Politically liberal	756	.32	.22	−.13	−.14	.07	.03	−.01	−.19	.05	.08
Informant reports	745										
Openness		.25	.30	.05	.03	.04	−.02	.05	.04	−.05	.01
Conscientiousness		−.01	.00	.30	.20	.01	.02	.09	.16	−.10	−.06
Extraversion		.05	−.07	.12	.12	.39	.24	.10	.06	−.12	−.06
Agreeableness		−.01	−.05	.06	.15	−.01	.02	.30	.24	−.07	.00
Neuroticism		.00	.00	−.11	−.08	−.12	−.03	−.16	−.09	.34	.20
Mean absolute correlation		.08	.07	.18	.13	.14	.10	.13	.12	.19	.10
SR-LBA column-vector correlations		.83		.86		.83		.90		.96	

Note. SR = self-report questionnaires; LBA = language-based assessment. Mean absolute and column-vector correlations are based on correlations after applying Fisher's *r*-to-*z* transformation and transforming back to *r*.

(*Appendices continue*)

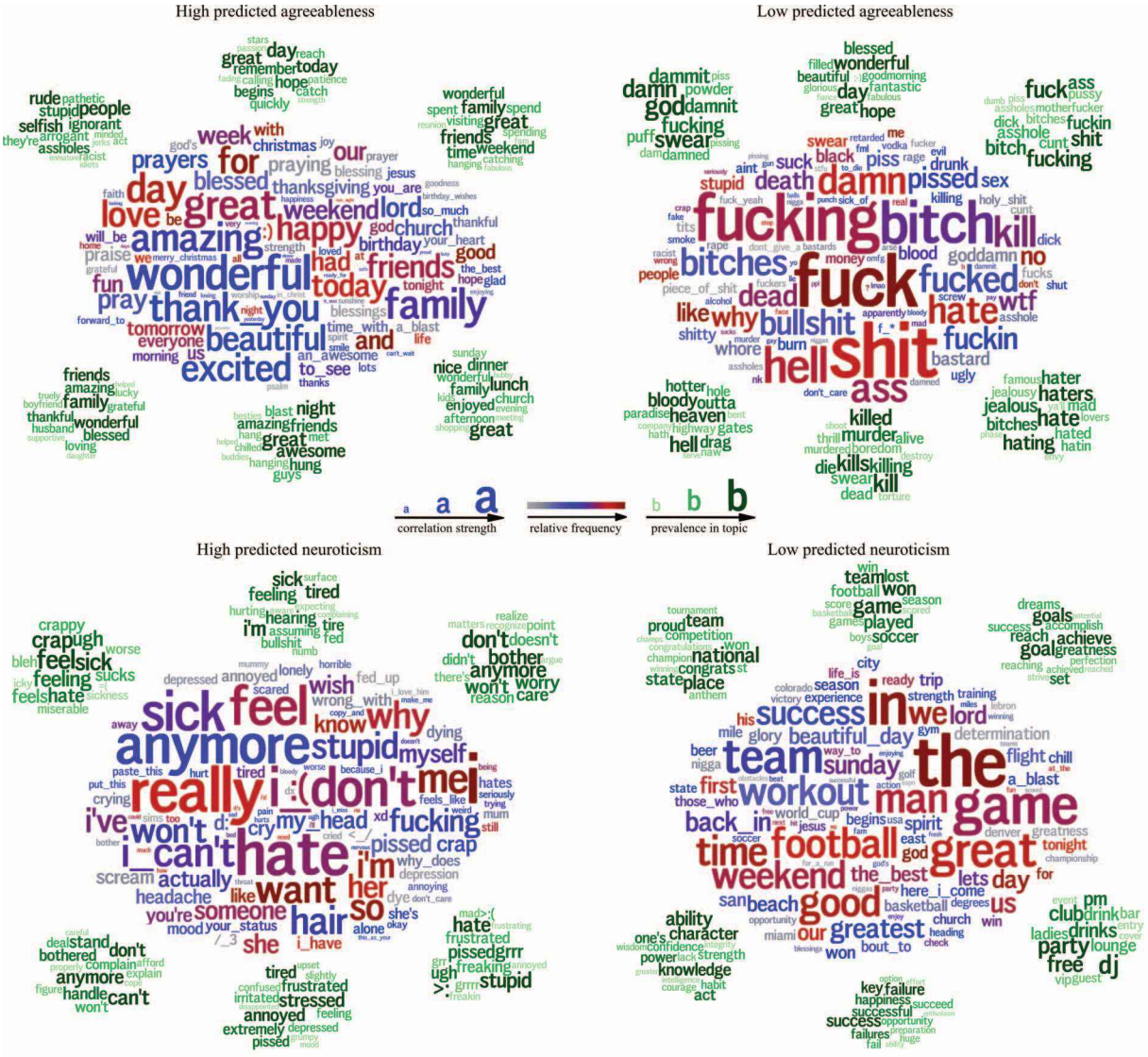
Appendix D

Words, Phrases, and Topics with Strongest Correlations to Openness, Conscientiousness, Agreeableness, and Neuroticism



Words, phrases, and topics with the strongest correlations to openness and conscientiousness as predicted by language (N = 4,824). Large central word clouds (red, blue, and gray) contain the 100 words and phrases with highest correlations with high and low levels of each trait. Word size is proportional to correlation size; color indicates word frequency. Underscores (_) are used to connect words within phrases and do not occur in the original text. The smaller surrounding word clouds (green) are the six most highly correlated topics, or clusters of semantically related words. Within topics, word size and color indicate word prevalence. All correlations are significant ($p < .001$).

(*Appendices continue*)



Words, phrases, and topics with the strongest correlations to agreeableness and neuroticism as predicted by language (N = 4,824). Large central word clouds (red, blue, and gray) contain the 100 words and phrases with highest correlations with high and low levels of each trait. Word size is proportional to correlation size; color indicates word frequency. Underscores (_) are used to connect words within phrases and do not occur in the original text. The smaller surrounding word clouds (green) are the six most highly correlated topics, or clusters of semantically related words. Within topics, word size and color indicate word prevalence. All correlations are significant ($p <$ .001).

(*Appendices continue*)

Appendix E

Test-Retest Stability

Test–Retest Correlations of Language-Based Assessments of Single Big Five Personality Traits

	Time 2	Time 3	Time 4
Openness			
Time 1	.71	.68	.64
Time 2		.74	.71
Time 3			.76
Conscientiousness			
Time 1	.75	.74	.70
Time 2		.76	.72
Time 3			.76
Extraversion			
Time 1	.72	.68	.64
Time 2		.72	.66
Time 3			.72
Agreeableness			
Time 1	.65	.61	.55
Time 2		.64	.57
Time 3			.65
Neuroticism			
Time 1	.62	.57	.51
Time 2		.62	.61
Time 3			.63

Note. Time 1 = July 2009 to December 2009; Time 2 = January 2010 to June 2010; Time 3 = July 2010 to December 2010; Time 4 = January 2011 to June 2011. Sample sizes for each correlation are the same as shown in Table 5.

Received April 9, 2014
Revision received August 8, 2014
Accepted August 28, 2014 ■