Review

Using Machine Learning to Advance Personality Assessment and Theory

Personality and Social Psychology Review
1–14

DOI: 10.1177/1088868318772990
pspr.sagepub.com
SAGE

Wiebke Bleidorn[1] and Christopher James Hopwood[1]

Abstract

Machine learning has led to important advances in society. One of the most exciting applications of machine learning in psychological science has been the development of assessment tools that can powerfully predict human behavior and personality traits. Thus far, machine learning approaches to personality assessment have focused on the associations between social media and other digital records with established personality measures. The goal of this article is to expand the potential of machine learning approaches to personality assessment by embedding it in a more comprehensive construct validation framework. We review recent applications of machine learning to personality assessment, place machine learning research in the broader context of fundamental principles of construct validation, and provide recommendations for how to use machine learning to advance our understanding of personality.

Keywords

personality assessment, machine learning, Big Five, construct validation, Big Data

Machine learning has led to remarkable advances in society including self-driving cars, speech recognition tools, and an improved understanding of the human genome. One of the most exciting applications of machine learning in psychological science has been the development of assessment tools that can predict personality traits using digital footprints such as Facebook (Youyou, Kosinski, & Stillwell, 2015) or Twitter profiles (Quercia, Kosinski, Stillwell, & Crowcroft, 2011).

Machine learning approaches to personality assessment involve automated algorithms for data extraction, cross-validation, and an emphasis on prediction, as described in detail below. These methods begin by gathering a large number of digital records with little or no relation to established theory to create scales that are associated with individual differences in enduring patterns of thoughts, feelings, and behavior (e.g., Funder, 1991; Tellegen, 1991) as assessed by more traditional personality measures. To do this, machine learning approaches focus on identifying empirical associations between digital records and established personality trait measures within specific samples. This strong empirical and mostly atheoretical focus has led to the development of potent assessment tools that can be used to reliably predict individual differences in personality traits.

However, relatively little is known about how these scales can be used to advance our understanding of personality constructs and human behavior. Machine learning approaches offer an unprecedented opportunity to advance both personality assessment and theory. The purpose of this article is to embed machine learning approaches to personality assessment in a construct validation framework that is concerned with both predicting and understanding human behavior (Cronbach & Meehl, 1955; Loevinger, 1957).

We first describe the basic approach to using machine learning algorithms for personality assessment and review recent studies that have used this approach. We next situate the findings of these studies within the broader principles of construct validation theory. We emphasize how this theory can supplement the focus on prediction characteristic of machine learning research with attention to other aspects of measurement, such as content, structural, external, and discriminant validity, and argue that doing so would appreciably enhance the potential of machine learning to generate novel tools and insights into personality traits. We conclude with nine specific recommendations for how to integrate machine learning techniques for personality assessment within a construct validation framework.

Machine Learning Approaches to Personality Assessment

People generate data whenever they go online, use their smartphones, or communicate through social media. The exponential explosion in the amount of data people are generating offers

[1]University of California, Davis, USA

Corresponding Author:
Wiebke Bleidorn, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA.
Email: wbleidorn@ucdavis.edu

Table 1. Key Concepts in Machine Learning Research on Personality Assessment.

Concept	Definition
Big data	Large and complex data sets that may be analyzed with machine learning and other advanced data analytic methods to extract patterns, detect signals, and address questions that are difficult to address with smaller data sets.
Machine learning	The study and construction of algorithms that can learn from data and make predictions based on data without being programmed to perform specific tasks.
User	A person who uses a computer, software, or network service, usually without the technical expertise required to fully understand it.
Digital footprints	Records of people's behaviors, geographical location, or physiological states in digital environments. Such footprints include web browsing logs, photos, global positioning system location logs, media playlists, voice and video call logs, and language used in Tweets or e-mails.
User–footprint matrix	Associates each individual user with a number of digital footprints (e.g., words used in Tweets).
Prediction models	Predict individual differences in users' personality traits based on the digital records in the user–footprint matrix (e.g., linear or logistic regression analyses).
K-fold validation	Multiple (k) rounds of cross-validation using different partitions of the data, for example, $k = 10$ equal-sized subsamples (referred to as folds).
Prediction accuracy	Typically quantified as Pearson product–moment correlation between a personality questionnaire score and the predicted values across users (for continuous outcomes such as personality scale scores).

researchers unprecedented opportunities for tracking, analyzing, and predicting human online behavior. In particular, recent advances in computer technology allow researchers to unobtrusively gather and automatically analyze large amounts of data from users of digital devices and services. Most often, these massive amounts of data are not collected with a specific research question in mind, but rather because it is affordable and because these data may be useful to answer future questions (Markowetz, Błaszkiewicz, Montag, Switala, & Schlaepfer, 2014). In fact, more than the actual size of the data set, a defining feature of *big data* is their use with machine learning approaches and other advanced data analytic methods to extract patterns, detect signals, and address questions that are difficult to address with smaller data sets.

In psychological science, a current question concerns the degree to which big data and digital footprints can be used to assess human personality. A guiding principle of this research is the assumption that personality characteristics influence the particular ways in which individuals use digital services and act in online environments. Consequently, data about how individuals use digital services and act in online environments should in turn be related to users' personalities (Back et al., 2010; Kosinski, Matz, Gosling, Popov, & Stillwell, 2015). To test this general hypothesis, researchers have begun to use machine learning to predict users' self- and peer-reported personality characteristics from their digital footprints. Typically, machine learning studies of personality assessment involve three steps: data collection, data extraction, and prediction of personality characteristics (cf. Kosinski, Wang, Lakkaraju, & Leskovec, 2016; Yarkoni & Westfall, 2017). Below, we describe these steps in more detail and summarize key concepts in Table 1. For the purpose of this article, we use the term machine learning personality assessment (MLPA) to contrast this methodology and type of data from more traditional and direct assessment methods such as self- or other report questionnaires.

Data Collection

People's social media accounts (e.g., Facebook, Twitter) offer a compelling source of rich and intimate digital information. For example, Facebook allows researchers to record information about users' demographic profiles (e.g., profile picture, age, gender, relationship status, place of origin, work, and education history), user-generated content (e.g., status updates, photos, videos), social network structure (e.g., list of friends and followers), and preferences and activities (e.g., group memberships, attended events). Moreover, user-generated text from messages, posts, or status updates can be further processed using text analysis tools such as the Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker, Francis, & Booth, 2001) or open vocabulary approaches (Kern, Eichstaedt, Schwartz, Dziurzynski, et al., 2014). Along with the digital data that is automatically stored, machine learning researchers can collect more traditional psychological assessment data, such as scores on personality questionnaires. Typically, online surveys or questionnaires are integrated within a social media platform and can then be easily connected with the information from users' social media profiles (Kosinski et al., 2015).

Data Extraction

Most types of digital footprints can be represented as a user–footprint matrix in which each individual user is associated with a number of possible footprints. Typically, each individual user is associated with only a small fraction of all possible footprints that leads to very large matrices with a great

majority of cells having a value of zero. For example, a user–footprint matrix might represent words and phrases used in social media posts. The rows of the user–language matrix would represent users, columns would represent words or phrases, and cells record the frequency of particular words or phrases per user (cf., Park et al., 2015). Since some users will never use a number of words, a large number of cells will be empty. Most machine learning studies use data reduction techniques and extract potentially relevant patterns to reduce the data to manageable dimensions and improve interpretability (for more details, see Kosinski et al., 2016).

Prediction of Psychological Characteristics

In a next step, machine learning researchers use the extracted set of variables to build prediction models of users' personality characteristics, often via linear and logistic regression analyses. These analyses are typically performed on a training subsample using multiple rounds of (*k*-fold) cross-validation to avoid model overfitting and to evaluate the predictive accuracy in a different subsample. That is, machine learning researchers typically split the full sample into *k* (typically 5-10) equally sized subsamples, build the regression model on a training subsample composed of all-but-one ($k - 1$) subsample, and validate this model on the excluded, testing subsample. This process is repeated *k* times for each subsample, and the prediction accuracy is averaged across all trials (Kosinski et al., 2016). For continuous outcome variables such as personality scale scores, the prediction accuracy is then typically quantified as the Pearson product–moment correlation between a questionnaire score (e.g., self-report) and the predicted values across users (cf. Kosinski, Stillwell, & Graepel, 2013).

A rapidly growing number of studies have used machine learning to predict various personality characteristics of users, with the majority focusing on the prediction of the "Big Five" personality traits of *neuroticism, extraversion, openness to experience, agreeableness*, and *conscientiousness* (John, Naumann, & Soto, 2008). Next, we review the current state of this research and evaluate the scientific evidence regarding the reliability and validity of MLPA.

Machine Learning Research on Personality Assessment

We identified three generations of studies in our literature review of machine learning research on personality assessment. The first generation of studies introduced MLPA. The second generation used large samples of social media users to finesse and optimize the predictive validity of these approaches. The third generation applied these approaches to test whether MLPA can improve upon more traditional assessment methods such as self- and peer-report.

First Generation

Early research, conducted primarily by scholars in the field of computer science, introduced MLPA (e.g., Chittaranjan, Blom, & Gatica-Perez, 2011; Iacobelli, Gill, Nowson, & Oberlander, 2011). For example, Golbeck, Robles, and Turner (2011) applied text analysis tools to the language written in the personal profiles and messages of 167 Facebook users, who also completed a Big Five self-report questionnaire. The authors analyzed more than 160 features extracted from participants' Facebook profiles and identified 74 variables (e.g., number of friends, favorite books) that were significantly ($p < .05$) correlated with at least one of the Big Five traits ($r = .16$-$.26$), which they then used to predict users' personality traits. Their regression models predicted Facebook users' self-reported personality traits better than chance, with the highest predictive accuracy occurring for openness.

Building upon this study, Golbeck, Robles, Edmondson, and Turner (2011) analyzed the public profiles and tweets of 50 Twitter users, who also completed a standard Big Five self-report questionnaire. As in their previous study on Facebook users, the authors correlated a large number of digital records obtained from Twitter and identified a set of 40 digital features (e.g., words per tweet, number of hashtags) that were significantly correlated ($p < .05$) with at least one of the Big Five personality scores ($r = .24$-$.40$).

Chittaranjan et al. (2011) and Chittaranjan, Blom, and Gatica-Perez (2013) used a similar approach to examine the relationship between digital records derived from smartphone data and self-reported Big Five personality traits in two studies of 83 and 117 smartphone users, respectively. In both studies, the authors first examined the correlations between more than 50 smartphone features (e.g., average call duration, average SMS length) and users' self-reports on the Ten-Item Personality Inventory (Gosling, Rentfrow, & Swann, 2003) for the total sample as well as separately for men and women ($r = .11$-$.26$). These analyses resulted in more than 200 statistically significant effects that the authors used to differentiate between high versus low scorers on each of the Big Five traits. Again, both studies found that users' smartphone data can be used to classify high versus low scorers on all Big Five traits better than chance.

In summary, the first generation of machine learning studies provided initial evidence that people leave digital traces in their online environments that are indicative of their personality traits. However, the relatively small sample sizes and theory-free tests of extremely large numbers of potentially relevant predictors limit the conclusion that can be drawn from these studies. In particular, while some of the observed correlations between digital footprints and personality traits seem intuitive and consistent with research in offline environments, such as the association between extraversion and the number of Facebook friends ($r = .19$), many others are not intuitive, such as a significant correlation

Table 2. Forms of Evidence for Construct Validity and Recommendations for Machine Learning Approaches to Personality Assessment.

Type	Definition	Recommendation
Substantive validity		
Content validity	Degree to which a measure samples all of the content of a construct and none of the content of other constructs	1. Examine content of MLPA models 2. Distinguishing between theory-expected and theory-unexpected content 3. Examine how different content can indicate certain traits across groups or over time
Structural validity		
Reliability	Degree to which a measure produces consistent scores across forms, time, and/or raters	4. Generalize algorithms to new samples and selected subsamples 5. Constrain indicators in principled ways based on theory and content even when the online data source is different
Generalizability	Degree to which one observation can be used to generalize to a universe of observations	
Factorial validity	Degree to which the hypothesized structure of a (multidimensional) measure is recoverable across contexts	
External validity		
Convergent validity	Degree to which measures of the same construct correlate with each other	6. Examine the nomological networks of MLPA scales 7. Use multimethod data to dissociate processes that lead to differences in MLPA scales and other personality measures 8. Enhance discriminant validity by minimizing intercorrelations of multidimensional MLPA scales 9. Examine discriminant validity using MLPA and other personality measures in MTMMs
Discriminant validity	Degree to which correlations between measures of different constructs do not exceed their associations in nature	
Criterion validity	Degree to which a measure predicts relevant outcomes	
Incremental validity	Degree to which a measure predicts a relevant outcome beyond what is known based on other measures	

Note. MLPA = Machine Learning Personality Assessment; MTMM = Multi-Trait Multi-Method Matrix.

between neuroticism and the character length of a participant's last name ($r = .18$). The questionable representation of the construct through the latter indicator (i.e., the lack of *content validity*, Haynes, Richard, & Kubany, 1995, see Table 2) suggests that this correlation may be spurious and unlikely to replicate in other samples.

Second Generation

More recent studies aimed to address the limitations of first-generation machine learning research on personality assessment by increasing statistical power and maximizing the predictive accuracy of MLPA tools. Many of these studies used data from the myPersonality Facebook application (Kosinski et al., 2015), which offered Facebook users access to 25 psychological tests including Big Five personality self- and peer-report questionnaires, well-being measures, and a computer-adaptive proxy of Raven's (1998) Standard Progressive Matrices. Between its release in 2007 and its closure in 2012, nearly 7.5 million users completed one or more measures on myPersonality. About 30% of the participants (above 2 million) volunteered to share the data on their Facebook profiles with the researchers, including information regarding their preferences, friend networks, or profile pictures. The large sample size of users who provided self- or peer-report information distinguished this project from earlier studies and allowed researchers to detect effects and patterns with high statistical power. According to the project website (http://mypersonality.org), as of March 2018, more than 50 articles and chapters have been published with these data and above 200 researchers are currently working with myPersonality data.

For example, Kosinski et al. (2013) analyzed the *likes* of over 58,000 U.S. Facebook users to predict a range of sensitive personal attributes including their personality traits. Facebook likes allow users to connect with objects that have an online presence (e.g., products, activities, places, movies, books, and music) and are shared with the public or among Facebook friends to express support or indicate individual preferences. Starting from a large number of oftentimes rare likes (~ 10 million user-like associations), the authors first trimmed and reduced the dimensionality of the data such that each user was associated with a vector of component scores. These component scores were then entered into logistic and linear regression models to predict users' psycho-demographic profiles from their likes. The results suggested that, based on users' likes, the prediction accuracy for some psychological traits approached the test–retest accuracy of a standard personality test.

The authors concluded that analyses of digital records of behavior, such as Facebook likes, may provide a convenient, accurate, and reliable approach to measuring personality traits,

which could "open new doors for research in human psychology" (Kosinski et al., 2013, p. 5805). However, while some predictive likes seemed tied in with theory and previous research, as in the case of "Cheerleading" and high extraversion, other highly predictive likes were rather elusive, as in the case of "Getting Money" and low neuroticism. This suggests that some of the content validity issues from the first generation may have persisted, despite the increase in statistical power.

Moreover, several particularly popular likes were associated with multiple attributes. For example, the brand "Hello Kitty" was predictive of younger age, high openness to experience, and low conscientiousness, agreeableness, and emotional stability. The concern with having common indicators across different dimensions is the potential to artificially increase correlations between estimates of these dimensions, compromising *discriminant validity* (see Table 2). As we will explain in more detail below, the relatively unexplored content validity and discriminant validity of MLPA measures complicate the interpretation of findings that are solely based on these scales and constrain the degree to which these measures can be used to advance personality theory (Bleidorn, Hopwood, & Wright, 2017; Wright, 2014).

Several other studies have used data from the myPersonality application to build assessment models of personality traits based on language-based information (e.g., Kern, Eichstaedt, Schwartz, Park, et al., 2015; Schwartz et al., 2013). For example, Park et al. (2015) compiled digital records of written language from 66,732 Facebook users and their Big Five personality self-reports to build and evaluate a predictive model of personality based on social media language use. The authors used an open vocabulary approach to extract more than 50,000 language features (words, phrases, topics) from users' status messages. After removing features with negligible correlations with the Big Five self-reports and reducing the dimensionality of the data, they used a final set of 5,106 language features to build prediction models for each of the five traits.

The resulting correlations between language-based and self-reported personality scores were comparable to other multimethod correlations in personality assessment (range: $r = .35$ for neuroticism to $r = .43$ for openness to experience), indicating that language features from social media texts can be used to automatically and accurately indicate personality differences among Facebook users. The results further suggested that language-based personality assessments were significantly correlated with informant reports of personality as well as external criteria, and were relatively stable above 6-month intervals. However, intercorrelations among different traits were significantly higher when measured with MLPA scales (average $r = .29$) than with self-report scales (average $r = .19$), suggesting relatively lower discriminant validity of MLPA scales using digital records of language.

Overall, the second generation of machine learning studies advanced the development of MLPA tools in at least two important ways. First, the use of substantially larger samples (typically, more than 50,000 users) enabled researchers to detect even subtle or rare but potentially informative signals in the data. Second, these studies fine-tuned the model building process to optimize the utilization of individual signals and to maximize the predictive accuracy of the assessment models. These advancements have led to the development of potent scales that can be used to make relatively accurate predictions.

Third Generation

The third and most recent generation of machine learning studies of personality assessment used the algorithms by Kosinski et al. (2013) and Park et al. (2015) to evaluate whether MLPA can outperform more traditional assessment methods for certain research questions. For example, Youyou et al. (2015) compared the accuracy of human and MLPAs using data from a large sample of Facebook users who had completed the 100-item IPIP Five-Factor Model questionnaire (Goldberg et al., 2006). MLPA models were built on participants' Facebook likes. Human personality judgments were obtained from the participants' Facebook friends, who were asked to describe a given participant using a 10-item version of the IPIP personality measure. The authors then compared the predictive accuracy of the MLPA models and human personality assessments with respect to three criteria: self-other agreement, inter-judge agreement, and external validity. With regard to self-other agreement, MLPA outperformed an average human judge ($r = .49$) when more than 100 likes were available. This advantage was largely driven by a particularly high agreement between self-reported and MLPA predicted openness to experience, whereas the self-other agreement estimates for the other Big Five traits were more comparable with those of the human judgments. The authors also found higher inter-judge agreement for MLPA models ($r = .62$) than for human judgments ($r = .38$) and higher external validity of MLPA models when predicting outcomes such as substance use, political attitudes, and physical health. MLPA models even outperformed the self-rated personality scores for some outcomes such as social network size and social network activities. Based on these findings, the authors concluded that MLPA may be superior to human personality judgments for some applications.

Two issues are important to note when interpreting these results. First, as pointed out by Youyou et al. (2015), the partial correlations between self-ratings, MLPA, and human judgments indicated that human judgments and MLPA scores may have captured distinct aspects of personality. Because content validity has not been considered in machine learning studies on personality assessment, it remains unclear what aspects of the traits MLPA scales capture. For example, it cannot be ruled out that these scales measure constructs that are related to a given personality trait (e.g., interests, motives) instead of actual trait content (i.e., relatively stable patterns of emotions, cognition, and behavior). Second, unlike the

MLPA models, the human judgments (friend reports) were not optimized to correlate with the self-ratings, which may have created an apples–oranges comparison in favor of the MLPA scales.

Youyou, Stillwell, Schwartz, and Kosinski (2017) used Facebook likes and digital records of social media language obtained through the myPersonality application to examine the similarity in personality traits between romantic partners (N = 1,101) and among friends (N = 46,483). Past research has found little evidence to suggest that romantic partners or friends are more similar in their personality than would be expected by chance. Youyou et al. proposed that the apparent lack of evidence for personality similarity effects may result from past studies' reliance on self-report and peer-report data. According to the authors, these data are at risk of obscuring the similarity among partners and friends who unconsciously treat one another as reference groups when reporting on their personality traits. MLPA, despite being optimized to predict users' self-reported personality traits, should not be prone to such reference-group effects (Heine, Buchtel, & Norenzayan, 2008), and thus might be more suited for studying personality similarity between friends and romantic partners. To examine partner and friend similarity across the different assessment methods, the authors estimated the correlations between romantic partners' and friends' personality scores using self-report, likes-based MLPA, and language-based MLPA measures, respectively. Both romantic partners (r = .20-.47) and friends (r = .12-.31) were more similar in terms of their likes-based and language-based MLPA scores as compared with their self-reported personality scores (all rs < .15 for friends and couples). These results seemed to support the authors' claim that MLPA models are more suited to detect similarities between dyads of romantic partners of friends.

The third generation of machine learning research on personality assessment foreshadows the potential of this approach for advancing personality science. However, we note two issues that should be considered when interpreting the findings of these studies. First, although an implicit message of this literature is that machine learning has the potential to generate significant advantages above traditional assessment tools, MLPA models were all initially validated on self-report questionnaires. In contemporary personality assessment research, scholars have shifted from questions about which method is optimal to questions about how different methods might assess different levels of the same construct (Bornstein, 2009). As we describe in detail below, we argue that this way of thinking would be productive for interpreting the results of comparative MLPA results as well.

Second, we know little about what aspects of personality MLPA models measure (Jensen, 2017). Their predictive accuracy notwithstanding, it remains unclear whether and to what degree these scales measure relatively stable patterns of thoughts, feelings, and behavior (i.e., personality traits) versus related psychological characteristics such as preferences, interests, attitudes, motives, or beliefs. It is generally expected that personality traits, being broad indicators of stable patterns of variation in human behavior, will be related to many psychologically relevant variables. However, from a trait realist perspective (Tellegen, 1991), it is critical to distinguish indicators of traits themselves, as opposed to indicators of other variables that are related to but not core features of traits.

The premise of this article is that the application of construct validation principles may help researchers to achieve a deeper understanding of what MLPA models are measuring (cf. Table 2). In the next section, we review key principles of construct validation and situate current evidence regarding MLPA within those principles.

Principles of Construct Validation

In the early days of quantitative personality assessment, principles of test development were not well established and the field's understanding of personality constructs was relatively impoverished. Researchers would develop tools to measure psychological concepts based on idiosyncratic theories, leading to an environment with various measures of different constructs, but no clear foundation upon which to evaluate those measures. It was entirely possible for two measures of one construct to correlate more strongly with a measure of a different construct than with one another (Loevinger, Gleser, & Dubois, 1953; Thorndike & Stein, 1937). The *Minnesota Multiphasic Personality Inventory* (MMPI; Hathaway & McKinley, 1943) attempted to solve this problem using an empirical criterion keying approach to test construction. In this atheoretical approach, items are picked from a large pool solely based on their ability to distinguish between groups of people determined to be different on some criterion. In the case of the MMPI, the criteria were psychiatric disorders as diagnosed by physicians.

The MMPI proved to be a powerful predictor of behavior that reliably differentiated people with different ostensible psychopathologies, but it also ran into problems. The content validity of some of its items was questionable, leading researchers to refer to scale numbers (i.e., "codes") rather than construct names, and to interpret these scores based on actuarial decision rules rather than an informed understanding of the constructs being measured (Hathaway & Meehl, 1951). The MMPI scales' discriminant validity was poor because the constructs it measured were not well-characterized, items overlapped between scales, and the initial criterion variables (i.e., physician diagnoses) were problematic. Successive versions of the MMPI relied less on "subtle" (i.e., content invalid) items (Gynther, Burkhart, & Hovanitz, 1979), reduced item overlap (Ben-Porath & Tellegen, 2008), and became increasingly connected to theories of personality and psychopathology (Sellbom, Ben-Porath, & Bagby, 2008) to address these issues.

The MMPI served as a sort of laboratory for the development of the principles of construct validation (e.g., Cronbach

& Meehl, 1955; Meehl, 1945). Psychological assessment and personality theory benefited from the arc of MMPI research during the 20th century and as a consequence, modern personality measures are generally designed with greater attention to construct validity than the original MMPI or its contemporaries.

The similarities between early MMPI research and recent research on MLPA are striking. Like the MMPI, machine learning algorithms are designed to maximize predictive validity. This focus is in keeping with the principles of empirical test construction (Breiman, 2001; Ratner, 2012), and there are considerable advantages to focusing on prediction (Yarkoni & Westfall, 2017). However, the largely atheoretical approach and strong focus on convergent validity proved to have limitations in the case of the MMPI. In the contemporary personality assessment literature, it is generally recognized that "an overreliance on a single parameter in item selection typically leads to a scale with one desirable psychometric property and numerous undesirable ones" (Morey, 2014, p. 186). These limitations can be addressed by considering other forms of validity within a broader construct validation framework.

The construct validation approach asserts that test development and theory development are inherently intertwined (Loevinger, 1957), such that developing and evaluating tests amounts to enhancing psychological theory. Our thesis is that refining MLPA via the principles of construct validation can enhance the utility of personality scales developed with this approach and provide novel and powerful tools with which to understand human behavior.

Table 2 classifies different forms of psychometric evidence from a construct validation perspective. Following Loevinger (1957), it distinguishes between three general classes of evidence for construct validity. *Substantive validity* refers to the degree to which the test's indicators match the theoretical contents of the construct it is designed to measure. The assumption of *content validity* is that the test indicators should be justified based on the underlying theory of the construct that the test is designed to measure (Haynes et al., 1995).

Structural validity refers to the internal characteristics of the test and involves the test's *reliability* (evidence about the consistency of test scores across time, raters, or content), generalizability (evidence regarding the degree to which the scale accurately measures the construct it is designed to measure in novel contexts), and *factorial validity* (evidence regarding the organization of scales within a multidimensional test).

External validity refers to the associations between test scores and phenomena outside of the test and involves the test's *convergent validity* (i.e., predictive validity, as the term is typically used in MLPA research; evidence regarding the association between two different measures of the same construct), *discriminant validity* (evidence regarding the associations between measures of different constructs), *criterion validity* (evidence regarding the association between a test score and other conceptually related attributes), and *incremental validity* (evidence regarding the association between a test score and some outcome variable controlling for a different test score). In what follows, we first describe the principles of construct validation in greater detail and then offer recommendations for how those principles can be brought to bear on MLPA.

Substantive Validity

Loevinger (1957) emphasized that test development is an empirical approach to theory development, and that evidence regarding construct validity is informative about both the test itself and the validity of the construct the test is designed to measure (see also Jackson, 1971). Three steps to test development, which correspond to the three types of validity presented in Table 2, are implied by this model. Loevinger asserted that "none of these steps to test construction is optional" (p. 654), and emphasized that substantive (i.e., content) validity is the foundation upon which the process of construct validation rests. She also emphasized that other aspects of construct validation interact with content validity. Negative structural and/or external validity evidence would suggest the need to revise test content, which amounts to revising the initial theory about the construct the test is designed to measure.

Like early MMPI research, investigators who have used machine learning to develop personality assessment tools have not focused on issues of content validity (Yarkoni & Westfall, 2017). Admittedly, content validity is a significant challenge for assessments developed via machine learning in big data because one of the defining characteristics of this approach is that variables are extracted after the data have been collected (Markowetz et al., 2014). This is incompatible with the traditional concept of content validity, which typically involves developing test items specifically designed to assess a certain construct prior to the generation of test content (Haynes et al., 1995). Furthermore, attempts to conceptualize the meaning of personality scales derived from big data can be challenging because the scales often comprised a very large number of diverse indicators. This is why the algorithms themselves are often treated as a "black box" (Breiman, 2001; Ratner, 2012).

Nevertheless, a consequence is that it is not clear what aspects of personality MLPA indicators identified via machine learning reference. Although there may be ways to establish or evaluate content validity in MLPA, thus far, this issue seems to have limited the degree to which machine learning approaches to personality assessment have been or could be used to advance personality theory. We see considerable value in marrying these approaches using machine learning to enhance our understanding of the content of personality traits (Bleidorn et al., 2017).

Structural Validity

At least three types of validity evidence can be used to inform the structural validity of a test (see Table 2): reliability, generalizability, and factorial validity.

Reliability. Multiple factors can influence the reliability of a test score. These factors correspond to different forms of reliability. For example, test–retest correlations estimate the impact of time on unreliability, internal consistency values estimate the influence of test content on unreliability, and measurement invariance analyses test the degree to which the covariance structure of the scales holds across time or different groups. From a classical test theory perspective, the reliability of a scale sets an upper bound for its potential validity, which is why reporting reliability statistics is standard in most psychological research.

Generalizability. With the variety of factors that can impact the value of an assessment instrument in mind, Cronbach, Gleser, Nanda, and Rajaratnam (1972) reinterpreted reliability as generalizability or the adequacy with which one can generalize from one observation to a universe of observations. They noted that the primary goal of developing personality measures involves using them in samples beyond the ones in which they were originally developed and emphasize that "a score's usefulness [. . .] largely depends on the extent to which it allows us to generalize accurately to behavior in some wider set of situations" (Shavelson, Webb, & Rowley, 1989, p. 922). Hence, it is crucial to have information about how a test will behave in new samples, at different times, or in various contexts.

As outlined above, machine learning algorithms are trained on one subsample and cross-validated on different subsamples to reduce the impacts of overfitting and demonstrate the generalizability of the algorithm within one sample (see Table 1). Thus, the procedures of generating MLPA models fit comfortably within a construct validation framework. However, this type of cross-validation does not adequately prove that a model will generalize to new samples. It is likely that there are sources of user sample homogeneity (e.g., due to historic events, technological developments), even in very large samples, that are not population-general. Examples that are standard to psychological assessment include demographic and cohort effects.

In addition, there are even more challenging factors specific to big data, namely that the digital footprints used to build the scales usually differ from one sample to the other. Unlike traditional personality tests which can be administered across samples and, if indicated, tested for measurement equivalence (Drasgow, 1984), machine learning researchers cannot directly compare content from different social media platforms (e.g., Facebook and Twitter) or indicators (e.g., likes and word counts). This poses a serious challenge for establishing and evaluating generalizability.

Factorial validity. Factorial validity refers to the inter-relationships of different indicators and scales of a test. It is relevant whenever more than one psychological dimension is hypothesized by theory and measured by an instrument. For example, personality trait theory and available evidence supports modest correlations between Big Five factors. As with reliability, it has become standard to include factor analyses or other evidence regarding the factorial structure of multidimensional tests in the personality and assessment literatures.

External Validity

Cronbach and Meehl (1955) emphasized the importance of the pattern of associations between a test score and external attributes or criteria that are not "operationally defined." The more we know about the external correlates of a particular scale, the more we know about the construct it is designed to measure, and the more confidence we can have in inferences based on estimates of that construct via the scale. Thus far, MLPA studies have focused nearly exclusively on external validity by maximizing the convergence of MLPA models with questionnaire measures of the Big Five.

From a construct validation perspective, negative external validity evidence can be particularly informative and imply (a) a problem with the test, (b) a problem with the theorist's understanding of the construct the test was designed to measure, or (c) some combination of both. The external validation of a scale can thus be seen as a progression of iterative bootstrapping studies of test score associations, each of which provides new and incremental information about the nomological network for both the latent construct and the scale designed to measure that construct. Four types of external validity evidence can be distinguished (see Table 2): convergent, discriminant, criterion, and incremental.

Convergent validity. Associations between MLPA scales and established Big Five questionnaires have been the predominant focus of MLPA studies up to this point. These studies have implicitly accepted the notion that Big Five questionnaires represent the "truth" about personality traits, insofar as they have optimized algorithms to predict people's scores on those questionnaires rather than endeavoring to identify personality factors independent of this theoretical framework. This approach has allowed researchers to maximize the convergent validity of MLPA with a well-known personality trait model. It is possible that MLPA could also be used to develop alternative or elaborated models of this framework, if embedded in a construct validation framework.

Discriminant validity. Campbell and Fiske (1959) observed that most validity research has a bias to focus on test score convergences, an issue that persists in the present (Bornstein, 2009). They asserted that discriminant validity, or evidence that a test does not measure attributes that it is

not intended to measure, is equally important for demonstrating construct validity. They presented the multi-trait multimethod matrix (MTMM) as a model for evaluating both convergent and discriminant validity. In an MTMM, there are multiple measures of at least two or more constructs. The idea is that correlations between different measures of the same construct should be stronger than correlations between measures of different constructs. In the history of psychological assessment, demonstrating discriminant validity has been more challenging than one might have expected because multimethod measures of the same trait often diverge (e.g., self-report and behavioral task; cf. Sharma, Markon, & Clark, 2014) and measures of different traits sometimes do not (e.g., personality disorders; Bornstein, 1998).

Critical in this framework is that different tests are treated as complementary rather than competitive. That is, the question is not so much "which test works best," but rather, "what do different tests tell us about an underlying construct." There are many reasons that test scores might converge or diverge. Convergence between measures of different constructs could occur because of a natural co-occurrence between those constructs, overlapping items, or poor content validity. The second two reasons would suggest the need to modify the tests, whereas the first would imply that the test is capturing covariance that is valid in nature. Divergence between measures of the same constructs can occur because different methods capture different aspects of the construct or because of weak content validity. Again, the second reason would indicate the need to modify the tests, whereas the first reason points to an opportunity to learn something interesting about the construct.

As of yet, machine learning studies focused less on evaluating the discriminant validity of MLPA scales with respect to other personality instruments. An exception is the study by Park et al. (2015) which showed that the intercorrelations among different traits are significantly higher when measured with MLPA than self-report scales, suggesting relatively lower discriminant validity for MLPA coefficients. For conscientiousness and agreeableness, the discriminant correlations even exceeded the scales' convergent correlations. Indicator overlap across MLPA scales may explain, in part, their poor discriminant validity. Frequent or popular digital footprints are often associated with multiple traits. To maximize convergent validity, machine learning researchers tend to use all informative digital indicators when building their assessment models. This results in scales that share many indicators, potentially compromising the discriminant validity of these scales.

Criterion validity. Criterion validity involves the degree to which a test is related to other, extra-test indicators of theoretical relevance. Park and colleagues (2015) also compared the correlations between some external criteria (e.g., self-monitoring, impulsiveness, and physical symptoms) and self-report personality questionnaires on one hand, and MLPA assessments, on the other hand. Overall, they found similar patterns of correlation between MLPA and self-report personality assessments with external criteria. This study offers an important initial contribution toward understanding the criterion validity of MLPA models, which could be expanded with studies using multimethod assessments of various external criteria.

Research including other person characteristics, such as interests, motives, attitudes, or life outcomes would be particularly useful to advance our understanding of the nature and composition of MLPA scales. As it stands, it is mostly unknown whether Big Five scores from machine learning algorithms have the same pattern of correlates as scores from questionnaires, even though the algorithms were optimized to predict those questionnaires.

Incremental validity. Incremental validity is the degree to which a test provides information over and above measures of other, typically related constructs for understanding criterion variables. The potential incremental validity of MLPA relative to more traditional forms of assessment has been highlighted recently in the personality assessment literature (Youyou et al., 2015; Youyou et al., 2017). There may indeed be cases in which MLPA may be optimal for certain kinds of personality assessments; however, from a construct validation perspective, this type of inference would be most comfortably embedded in a fuller understanding of what MLPA algorithms are measuring.

Summary

Predicting behavior is important for personality assessment (Wiggins, 1973) and doing so efficiently and in a way that maximizes the potential of rich data sources is a powerful application of MLPA. However, from a construct validation perspective, personality assessment should serve the interconnected purposes of predicting and understanding personality (Cronbach & Meehl, 1955; Loevinger, 1957). Thus, machine learning research on personality assessment has been based on a philosophy that emphasizes prediction (Yarkoni & Westfall, 2017) and has thus focused almost exclusively on the convergence of MLPA models with established personality measures. This focus reflects one aspect of the third stage of the construct validation framework. In contrast, little research has been done to use machine learning and digital records of behavior to further our understanding of personality, as could be accomplished via focus on other aspects of construct validation. We see appreciable potential to use machine learning technology to develop improved tools as well as new insights into personality through an enhanced focus on construct validity.

Recommendations for Machine Learning Approaches to Personality Assessment

The juxtaposition of contemporary MLPA and general principles of construct validation leads to specific recommendations for how machine learning research can advance personality science. Below, we offer recommendations relevant to each of the three major steps of construct validation (Loevinger, 1957) outlined above (see Table 2).

Substantive Validity

Machine learning models have been referred to as a "black box" because scales are typically composed of so many indicators that their systematic examination is challenging. Moreover, as commonly applied, any indicator that "works" in the sense that it converges with an established measure is retained and any indicator that does not is discarded. This leads to algorithms with some indicators that have a clear conceptual connection to the construct they intend to measure and others that do not, a result that stands in marked contrast to the emphasis on content validity in construct validation approaches to test development (Loevinger, 1957).

Our *first recommendation* is to move past the "black box" perspective by considering more carefully the contents extracted from machine learning for the assessment of personality traits (Yarkoni & Westfall, 2017). Previous examinations of MLPA models have revealed two broad classes of indicators: those that are intuitive and tie in with theory and previous research and those that are surprising and not intuitive. Explicitly distinguishing these two classes offers a powerful means of learning about the validity of MLPA scales and the theoretical constructs they are designed to measure. Indicators that are not surprising and connect well with past research suggest that the algorithm "worked." A lesson from early MMPI research was that such indicators are more likely to be effective in new samples than indicators with an unclear relation to the target construct. Surprising, nonintuitive indicators are less likely to generalize to other samples and should be treated with some skepticism in terms of generalizability. However, sometimes surprising indicators may work consistently (Waljee, Higgins, & Singal, 2014), and in such cases, they may provide novel insights into the constructs to which they relate. From this perspective, machine learning approaches have the potential to broaden and refine our understanding of the structure and content of the Big Five if they turn out to replicate (Bleidorn et al., 2017).

One approach to examining the content validity of MLPA scales would begin with a precise definition of the targeted trait constructs, including examples of relevant digital indicators of behaviors, thoughts, and feelings. The large body of literature on the Big Five provides circumscribed definitions of these five personality traits including rich descriptions of theoretically relevant and irrelevant content (e.g., John et al., 2008). Using these theoretical specifications as a guide, the content validity of the MLPA scales could then be evaluated using an expert-rating approach. That is, a group of subject matter experts rates the Big Five MLPA scales regarding relevance, representativeness, specificity, and clarity of their content. The resulting expert-consensus ratings could then be used to guide evaluations of content validity and help identify irrelevant, overrepresented, or missing content (Haynes et al., 1995). Admittedly, the large numbers of diverse digital indicators that go into MLPA scales complicate the expert-rating procedure. To simplify this approach, we would suggest to rate the most predictive digital indicators that are endorsed by a nontrivial number of users (e.g., 100 users; cf. Kosinski et al., 2013).

Our *second recommendation* is to examine the content of MLPA models across time and groups. For example, identifying trait indicators that are similarly effective for different age groups has posed a significant challenge for personality researchers because development can influence the way a trait is expressed (Caspi, Roberts, & Shiner, 2005). Given that personality traits reflect underlying dispositions that can be expressed differently at different ages, some indicators of personality will be age-general whereas others will tend to be age-specific. Machine learning research has the potential to both improve developmentally sensitive personality assessment to permit enhanced prediction of behavior at different life stages and contribute to a better theoretical understanding of personality development by identifying and distinguishing age-general and age-specific indicators. To the degree that the same indicators can be used to assess personality differences across age groups, machine learning research can help develop measures that are more effective across the lifespan. Likewise, the detection of indicators that are valid for certain age groups (but not others) could inform the development of age-specific personality tests.

Similarly, MLPA could greatly advance research on cross-cultural personality differences. To the degree that there are digital footprints that are systematically related to personality differences across cultures, the use of machine learning to identify digital indicators can help develop culture-free personality measures whereas indicators that are valid in certain cultures but not in other could inform the development of culture-specific tests.

Structural Validity

Our *third recommendation* is to evaluate and report reliability and factorial validity statistics whenever possible. This is standard for other approaches to personality assessment and should be more routinely adapted by MLPA. Park et al. (2015) illustrated how the traditional test–retest approach can be approximated using timestamps to split the digital records per person into 6-month subsets. However, to the best of our knowledge, no study has examined the factorial

validity of MLPA tools in a multidimensional (e.g., five-factor) context.

Our *fourth recommendation* is to generalize machine learning algorithms across different samples from different populations of users whenever possible. In contrast to other approaches, machine learning studies have the distinct advantage of pairing oftentimes large samples of users with cross-validation techniques to reduce the risk of overfitting the data (Kosinski et al., 2016). In cases where the samples are so large that they approach the population of interest, such studies may generate models relevant for all people whose behavior a researcher might want to predict or understand. However, in most cases, cross-validation within a sample does not guarantee generalizability to new samples (e.g., from different cultures or collected at different times). Furthermore, certain aspects of MLPA studies make generalization challenging, such as the fact that different studies often train algorithms on different kinds of digital records (e.g., Facebook likes vs. language-based indicators). Yet, generalizability is a critical feature of a well-functioning personality assessment tool, and both successes and failures in generalization are informative about personality. As such, more efforts are needed to evaluate the generalizability of MLPA measures across different samples of both users and digital records.

To our knowledge, this has never been done in MLPA studies probably because this would only be possible in rare situations in which the same kind of data (e.g., written text from social media platforms) was available across samples of users of different online services (Twitter, Facebook, Instagram), from different countries, or different times. A less optimal, but nevertheless, informative alternative, particularly in very large samples, would be to split up validation subsamples in ways that are not random. For example, having trained an algorithm on a randomly selected subsample, cross-validation samples selected based on demographic or other characteristics could be used to evaluate how well the algorithms function in different groups. Effective cross-validation would support the validity of the indicators whereas problems with cross-validating would be informative regarding differences in the way personality is expressed across different groups.

Our *fifth recommendation* rests upon a more careful consideration of content validity to design MLPA measures that are generalizable across samples of users of different social media or other online services. Knowledge about the content of the most predictive indicators would provide a means for constraining indicators based on their conceptual relations to the construct across studies and samples even when those indicators are derived from different online platforms or even reflect different types of digital footprints. For example, Facebook likes of "dancing" and "beerpong" (cf. Kosinski et al., 2013) may be conceptually similar to party-related words such "party" or "a blast" in Facebook posts (cf. Park et al., 2015), and both might assess extraversion. Identifying and constraining such indicators based on their conceptual meaning (i.e., liking parties) would help provide the means to replicate and generalize MLPA models across samples and contexts.

External Validity

Our *sixth recommendation* is to carefully examine the associations between MLPA scores and external criteria (e.g., Park et al., 2015). Particular attention should be paid to the similarity of external correlates between machine learned and traditional trait scores. For example, profiles of correlations with other variables can be compared (e.g., Westen & Rosenthal, 2003) to provide a quantitative index of the degree to which two sets of variables are measuring the same construct in terms of their network of relationship with other variables. Often, other variables are readily available in the data sets in which MLPA algorithms are developed (e.g., myPersonality). Evaluating the similarity of the pattern of criterion correlates is an important and relatively straightforward means of evaluating construct validity.

Instances of mismatch can be particularly informative about the nature of personality. As an example, narcissism researchers used this approach to show that different measures of that construct had rather different patterns of external validity (Maxwell, Donnellan, Hopwood, & Ackerman, 2011; Miller & Campbell, 2008), which corresponded to different underlying theoretical models of narcissism (Pincus & Lukowitsky, 2010). This finding helped move this literature from a confusing array of seemingly discrepant results to systematic programs research on the core features of (Wright et al., 2017; Wurst et al., 2017) and origins of different perspectives (Ackerman, Hands, Donnellan, Hopwood, & Witt, 2017; Miller et al., 2014) on narcissism. That is, comparing the external correlates of different measures both improved the measurement of narcissism and contributed to a deeper understanding of what narcissism actually is.

Our *seventh recommendation* is to use multimethod data to dissociate the processes that contribute to differences in scores on MLPA scales and other types of assessments. For example, in the Youyou et al. (2015) study, informants were probably not trying to respond to the questionnaire in the way they thought the target would respond—they were offering their own take using a short 10-item measure. When MLPA scales correspond more closely to self-reports, it is not necessarily because they are "better" than informants, it may be because they are trying to do different things with different means and potentially measure different aspects of the broader constructs (as indicated by the partial correlations between self-ratings, MLPA scores, and friend ratings). This has informative implications for personality because it implies that both the measure and the motivations of the rater will impact the ratings.

Because many personality traits relate to one another (e.g., Digman, 1997), establishing the discriminant validity

of scales designed to assess different traits is a significant challenge. Our *eighth recommendation* is to try to enhance discriminant validity by minimizing the intercorrelations of multidimensional MLPA scales that are intended to measure constructs with low correlations in theory or on other instruments (e.g., the Big Five). To maximize convergent validity, machine learning researchers have typically used all informative digital indicators when building their assessment models. This resulted in MLPA scales that share many indicators, the presence of which ensures discriminant validity problems. A corollary of our eighth recommendation is thus that the same indicators should generally not be used to measure different traits. This will lead to more useful MLPA tools and a more precise picture of the kinds of digital indicators that are specifically related to certain traits.

A *ninth and related recommendation* is to evaluate discriminant validity using other types of personality measures in MTMMs. From a construct validation perspective, a well-functioning scale should correlate more strongly with a scale designed to measure the same construct from a different instrument than with any other scale. However, method variance makes this challenging because different scales from a common instrument will tend to correlate highly with one another and cross-method convergent correlations may tend to be low. An explicit and detailed examination of the MTMM of MLPA scales and other measures of personality (e.g., self-report, peer-report) can thus contribute to the development of more effective assessment tools and be informative about how different trait indicators capture variance associated with different traits.

Conclusion

We have reviewed recent machine learning research on personality assessment, which has focused on the prediction of individual differences and comparisons to established personality measures. We have discussed this research with regard to a broader construct validation perspective on personality assessment, which sees prediction and explanation as two aspects of an iterative developmental process of theory and test building. We specifically reviewed three broad steps to test development from a construct validation perspective: substantive validity, structural validity, and external validity. From this perspective, the existing MLPA literature has focused nearly exclusively on one aspect of external validity and could be fruitfully enhanced by a consideration of other aspects of construct validation. Using principles related to each of these steps as a guide, we offered nine specific recommendations for how machine learning research on personality assessment can contribute to personality science in the form of both more robust assessment tools and new insights into personality structure, processes, and development. These included the following:

Substantive Validity: (1) examine the content of MLPA models in terms of what can be learned about the underlying construct, with a specific focus on distinguishing between theory-expected and theory-unexpected content, (2) examine how different content can validly indicate certain traits across groups or over time.

Structural Validity: (3) report basic statistics on test reliability and structure, (4) generalize algorithms to new samples and selected subsamples when possible, and (5) constrain indicators in principled ways based on content even when the data source is different.

External Validity: (6) compare the pattern of validity correlates (nomological network) of MLPA to those of other measures, (7) when scores differ across methods, try to understand the sources of those differences in terms of the underlying process of test score generation, (8) elevate the importance of discriminant validity as a criterion for multidimensional MLPA, and (9) examine discriminant validity with respect to different methods designed to measure the same multidimensional traits.

The use of machine learning and big data to develop personality assessment tools may eventually replace other forms of test development technologies for a wide variety of applications. This technology will be more potent to the degree that the development of these tools occurs within a construct validation framework that links the empirical processes of test development with personality theory. With greater attention to all aspects of the construct validation process, machine learning research can become a powerful tool for both predicting behavior and developing novel insights into the nature of personality.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Ackerman, R. A., Hands, A. J., Donnellan, M. B., Hopwood, C. J., & Witt, E. A. (2017). Experts' views regarding the conceptualization of narcissism. *Journal of Personality Disorders*, *31*, 346-361.

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.

Ben-Porath, Y. S., & Tellegen, A. (2008). *MMPI-2-RF manual for administration, scoring, and interpretation*. Minneapolis, MN: NCS Pearson.

Bleidorn, W., Hopwood, C. J., & Wright, A. G. (2017). Using big data to advance personality theory. *Current Opinion in Behavioral Sciences*, *18*, 79-82.

Bornstein, R. F. (1998). Reconceptualizing personality disorder diagnosis in the DSM-V: The discriminant validity challenge. *Clinical Psychology: Science and Practice*, *5*, 333-343.

Bornstein, R. F. (2009). Heisenberg, Kandinsky, and the hetero-method convergence problem: Lessons from within and beyond psychology. *Journal of Personality Assessment*, *91*, 1-10.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and change. *Annual Review of Psychology*, *56*, 453-484.

Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2011). Who's who with big-five: Analyzing and classifying personality traits with smartphones. In *2011 15th Annual International Symposium on Wearable Computers (ISWC)* (pp. 29-36). New York, NY: IEEE.

Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2013). Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, *17*(3), 433-450.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements*. New York: Wiley, 1972.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology*, *73*, 1246-1256.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, *95*, 134-135.

Funder, D. C. (1991). Global traits: A neo-Allportian approach to personality. *Psychological Science*, *2*, 31-39.

Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011). *Predicting personality from twitter*. Paper presented at the 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (SocialCom), Boston, MA, USA.

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, pp. 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.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*, 504-528.

Gynther, M. D., Burkhart, B. R., & Hovanitz, C. (1979). Do face-valid items have more predictive validity than subtle items? The case of the MMPI Pd scale. *Journal of Consulting and Clinical Psychology*, *47*, 295-300.

Hathaway, S. R., & McKinley, J. C. (1943). *Manual for the Minnesota Multiphasic Personality Inventory*. New York, NY: Psychological Corporation.

Hathaway, S. R., & Meehl, P. E. (1951). *An atlas for the clinical use of the MMPI*. Oxford, UK: University of Minnesota Press.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*, 238-247.

Heine, S. J., Buchtel, E. E., & Norenzayan, A. (2008). What do cross-national comparisons of personality traits tell us? The case of conscientiousness. *Psychological Science*, *19*, 309-313.

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.

Jackson, D. N. (1971). The dynamics of structured personality tests: 1971. *Psychological Review*, *78*, 229-248.

Jensen, E. A. (2017). Putting the methodological brakes on claims to measure national happiness through Twitter: Methodological limitations in social media analytics. *PLoS ONE*, *12*, e0180080. doi:10.1371/journal.pone.0180080

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy. 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. doi:10.1177/1073191113514104

Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Park, G., Ungar, L. H., Stillwell, D. J., . . . Seligman, M. E. (2014). From "sooo excited!!!" to "so proud": Using language to study development. *Developmental Psychology*, *50*, 178-188.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, *70*, 543-556.

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*, 5802-5805.

Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods*, *21*, 493-506.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635-694.

Loevinger, J., Gleser, G. C., & Dubois, P. H. (1953). Maximizing the discriminating power of a multiple-score test. *Psychometrika*, *4*, 309-317.

Markowetz, A., Błaszkiewicz, K., Montag, C., Switala, C., & Schlaepfer, T. E. (2014). Psycho-informatics: Big Data shaping modern psychometrics. *Medical Hypotheses*, *82*, 405-411. doi:10.1016/j.mehy.2013.11.030

Maxwell, K., Donnellan, M. B., Hopwood, C. J., & Ackerman, R. A. (2011). The two faces of Narcissus? An empirical comparison of the Narcissistic Personality Inventory and the Pathological Narcissism Inventory. *Personality and Individual Differences*, *50*, 577-582.

Meehl, P. E. (1945). The dynamics of "structured" personality tests. *Journal of Clinical Psychology*, *1*, 296-303.

Miller, J. D., & Campbell, W. K. (2008). Comparing clinical and social-personality conceptualizations of narcissism. *Journal of Personality*, *76*, 449-476.

Miller, J. D., McCain, J., Lynam, D. R., Few, L. R., Gentile, B., MacKillop, J., & Campbell, W. K. (2014). A comparison of the criterion validity of popular measures of narcissism and narcissistic personality disorder via the use of expert ratings. *Psychological Assessment*, *26*, 958-969.

Morey, L. C. (2014). The Personality Assessment Inventory. In R. P. Archer & S. M. Smith (Eds.), *Personality assessment* (2nd ed., pp. 181-228). New York, NY: Routledge.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology*, *108*, 934-952

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: LIWC*. Mahwah, NJ: Lawrence Erlbaum.

Pincus, A. L., & Lukowitsky, M. R. (2010). Pathological narcissism and narcissistic personality disorder. *Annual Review of Clinical Psychology*, *6*, 421-446.

Quercia, D., Kosinski, M., Stillwell, D., & Crowcroft, J. (2011, October). Our twitter profiles, our selves: Predicting personality with twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (pp. 180-185). IEEE.

Ratner, B. (2012). *Statistical and Machine Learning Data Mining: Techniques for better predictive modeling and analysis of big data*. Boca Raton, FL: CRC Press.

Raven, J. C. (1998). *Raven's progressive matrices*. Oxford, UK: Oxford Psychologists Press.

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*, e73791. doi:10.1371/journal.pone.0073791

Sellbom, M., Ben-Porath, Y. S., & Bagby, R. M. (2008). Personality and psychopathology: Mapping the MMPI-2 Restructured Clinical (RC) scales onto the five factor model of personality. *Journal of Personality Disorders*, *22*, 291-312.

Sharma, L., Markon, K. E., & Clark, L. A. (2014). Toward a theory of distinct types of "impulsive" behaviors: A meta-analysis of self-report and behavioral measures. *Psychological Bulletin*, *140*, 374-408.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*, 922-932.

Tellegen, A. (1991). Personality traits: Issues of definition, evidence, and assessment. In D. Cichetti & W. Grove (Eds.), *Thinking clearly about psychology: Essays in honor of Paul Everett Meehl* (pp. 10-35). Minneapolis: University of Minnesota Press.

Thorndike, R. L., & Stein, S. (1937). An evaluation of the attempts to measure social intelligence. *Psychological Bulletin*, *34*, 275-285.

Waljee, A. K., Higgins, P. D., & Singal, A. G. (2014). A primer on predictive models. *Clinical and Translational Gastroenterology*, *5*(1), e44.

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, *84*, 608-618.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Boston, MA: Addison-Wesley.

Wright, A. G. C. (2014). Current directions in personality science and the potential for advances through computing. *IEEE Transactions on Affective Computing*, *5*, 292-296.

Wright, A. G. C., Stepp, S. D., Scott, L., Hallquist, M., Beeney, J. E., Lazarus, S. A., & Pilkonis, P. A. (2017). The effect of pathological narcissism on interpersonal and affective processes in social interactions. *Journal of Abnormal Psychology*, *126*, 899-910.

Wurst, S. N., Gerlach, T. M., Dufner, M., Rauthmann, J. F., Grosz, M. P., Küfner, A. C., . . . Back, M. D. (2017). Narcissism and romantic relationships: The differential impact of narcissistic admiration and rivalry. *Journal of Personality and Social Psychology*, *112*, 280-290.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons learned from machine learning. *Perspectives on Psychological Science*, *12*, 1100-1122. doi: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*, 1036-1040.

Youyou, W., Stillwell, D., Schwartz, H. A., & Kosinski, M. (2017). Birds of a feather do flock together: Behavior-based personality-assessment method reveals personality similarity among couples and friends. *Psychological Science*, *28*, 276-284.