Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

Language-based personality: a new approach to personality in a digital world☆

Ryan L Boyd and James W Pennebaker

Personality is typically defined as the consistent set of traits, attitudes, emotions, and behaviors that people have. For several decades, a majority of researchers have tacitly agreed that the gold standard for measuring personality was with self-report questionnaires. Surveys are fast, inexpensive, and display beautiful psychometric properties. A considerable problem with this method, however, is that self-reports reflect only one aspect of personality — people's explicit theories of what they think they are like. We propose a complementary model that draws on a big data solution: the analysis of the words people use. Language use is relatively reliable over time, internally consistent, and differs considerably between people. Language-based measures of personality can be useful for capturing/modeling lower-level personality processes that are more closely associated with important objective behavioral outcomes than traditional personality measures. Additionally, the increasing availability of language data and advances in both statistical methods and technological power are rapidly creating new opportunities for the study of personality at 'big data' scale. Such opportunities allow researchers to not only better understand the fundamental nature of personality, but at a scale never before imagined in psychological research.

Address
The University of Texas at Austin, Department of Psychology, United States

Corresponding authors: Boyd, Ryan L (ryanboyd@utexas.edu), Pennebaker, James W (pennebaker@utexas.edu)

Current Opinion in Behavioral Sciences 2017, **18**:63–68

This review comes from a themed issue on **Big data in the behavioral sciences**

Edited by **Michal Kosinski** and **Tara Behrend**

http://dx.doi.org/10.1016/j.cobeha.2017.07.017

People differ dramatically in the ways they think, feel, and behave in general, forming the basis for what we refer to as personality. Going back to the ancient Greeks, formal thinking about personality has relied on different methods to measure and explain personality. Classically, Galen posited four general temperaments — sanguine, phlegmatic, melancholic, and choleric — based on his observations of biology and the theories of Hippocrates [1]. Freud [2] revolutionized the broader discussion about personality by arguing that inborn temperament and early experiences shaped what people were like later in life. Temperament researchers focused on the activity levels and emotionality of infants to posit the likely genetic and biological bases of individual differences [3]. Others, such as Gordon Allport [4] pointed to the enduring and stable behavioral styles that people possessed — including the ways they walked, gestured, or chewed gum. Even the most nuanced behaviors revealed people's basic characteristics.

Not until the advent of modern social science did psychologists begin to focus on the careful measurement of personality [5–7]. In the last quarter of the 20th century, the trait approach emerged that effectively defined modern personality theory, ushering in detailed factor models of the construct [8,9]. The new trait approach energized the field of personality research, in part because it leaned heavily on self-reports of participants' self-concepts for understanding their general personality characteristics. This was a profound development in personality research: widespread adoption of self-reports meant that it was now possible to have very large groups of people complete extensive personality scales rather than relying on more time-intensive and resource-intensive approaches. Paired with advances in statistical and other computational methods, the adoption of self-report scales resulted in new ways of studying the domains and correlates of traits.

Self-report questionnaires can provide rich information about peoples' conscious, explicit self-concepts. However, most personality experts have harbored occasional doubts about the degree to which people's self-reported traits reflect who they really are [10]. For example, to what degree do self-theories map onto their actual behaviors? Across thousands of studies, we know that self-reports correlate nicely with other self-reports from the same people, yet often show lackluster overlap with more objective measures that presumably capture the same underlying traits. Researchers consistently find that widely-used and well-validated self-report measures are insufficient when it comes to forming an accurate understanding of even basic human patterns such as workplace behaviors [11], physical activity [12], and expressions of happiness [13••] or other emotional states [14].

☆ Preparation of this manuscript was aided by grants from the National Institutes of Health [5R01GM112697-02], John Templeton Foundation [#48503], and the National Science Foundation [IIS-1344257].

Are we thinking about personality in the right way? Are people's self-theories the appropriate gold standard for assessing personality? If not self-reports, does a gold standard exist? As we outline below, we must move beyond the gold standard way of thinking. Self-reports reflect one dimension of personality, while nervous system activity may serve as another, genetic factors may be the basis of a third, and so on.

Beyond self-reports and biological markers, recent research has demonstrated that a powerful reflection of personality can be gleaned from the words people use in everyday life. As an increasing number of studies demonstrate, the ways in which people use words is reliable over time, internally consistent, predictive of a wide range of behaviors and even biological activity, and varies considerably from person to person. Language, then, is yet another fundamental dimension of personality. Of great benefit to researchers, and unlike other standard personality markers, people do not need to complete questionnaires or submit to invasive blood or genetic tests in order to provide useful personality data in the form of language.

Language and personality in the land of big data

Over half of the planet's population uses the internet, and over 80% of people in developed countries are internet users [15]. Every minute, more than 350 000 tweets are posted to Twitter, approximately 3 million Facebook posts are shared, 4 million Google queries are submitted, and over 170 million e-mails are sent [16,17]. In more human terms, the average office worker sees over 120 e-mails per day [18], the typical teen in the United States sends over 60 text messages per day from their mobile phones [19] and the average Facebook user writes 25 comments daily [20]. In short, the amount of language data generated by humans on a minute-by-minute basis around the world is nothing short of staggering.

As with the unprecedented availability of human-generated data, the field of psychology has witnessed a recent cascade in psychometric techniques that are well-suited to a big data research culture. Of the more recent psychological assessment methods, perhaps the most accessible and refined to date is that of automated language analysis, which is currently experiencing rapid adoption and growth across a wide range of academic fields. Historically, psychologists have long believed that a person's words can be revealing of deeper, meaningful psychological constructs [21–23]. For example, classical research on motivation found that the individual's personal strivings, such as the needs for affiliation and achievement, were manifest in their everyday words [24], and it has long been believed that linguistic cues can be used to identify different states of consciousness [25]. However, the modern rejuvenation of language research in the field of personality psychology has been primarily driven by the adoption of modern statistical methods and technological innovations, such as the boom of personal computing power and data accessibility [26].

Unlike most classical research on language and psychology, which typically treated linguistic measures as indicators of a person's transient mental state [14,27], several key studies were conducted early on in the current language analysis renaissance which demonstrated that the properties of language-based psychological measures behave in much the same way as traditional measures of personality. For example, Pennebaker and King [28] explored the psychometric properties of language as a psychological measure, finding that the majority of measures provided by the Linguistic Inquiry and Word Count method [29] exhibited all of the hallmarks of a standard individual differences measure: test–retest reliability, external validity, and internal consistency. A considerable amount research within the LIWC domain has expanded these initial findings, establishing the word-counting paradigm as a robust tool for measuring stable individual differences [30•,31,32].

In the modern research world, where psychologically-relevant data is available in great abundance, psychometric techniques like language analysis allows researchers to indirectly probe and better understand how lower level psychological processes function and interact to manifest in the form of personality in the real world. In other words, techniques such as language analysis are particularly well-suited to the proximal measurement the lower level processes that cohere to form personality, especially in relation to traditional self-report measures. Countless patterns of attention, behaviors, and emotions are deeply embedded in a person's language [31], and psychologists now have access to an ever-growing number of methods to extract these patterns for deeper study.

Given the modern surge of language data, as well as methods for extracting psychological information from such data, a logical next step for social scientists is to begin benefiting from the trait-like qualities of language-based measures in psychological research. In the current climate of the 'big data' revolution, many of the logistical properties for which self-report measures are often lauded ring even truer for language-based measures of personality. While self-reports are relatively easy to collect compared to other measures such as physiological data, language analysis often relies on data that *already exists*. Moreover, pre-existing digital data from the web, smart phones, and social media are inherently ecologically valid, having originated from thoughts and behaviors that occur in the absence of researcher intervention.

It is vital to note that the analysis of language for personality research can be performed *at scale* in nearly any