Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

Using big data to advance personality theory

Wiebke Bleidorn[1], Christopher J Hopwood[1] and Aidan GC Wright[2]

Big data has led to remarkable advances in society. One of the most exciting applications in psychological science has been the development of computer-based assessment tools to assess human behavior and personality traits. Thus far, machine learning approaches to personality assessment have been focused on maximizing predictive validity, but have been underused to advance our understanding of personality. In this paper, we review recent machine learning studies of personality and discuss recommendations for how big data and machine learning research can be used to advance personality theory.

Addresses
[1] University of California, Davis, United States
[2] University of Pittsburgh, United States

Corresponding author: Bleidorn, Wiebke (wiebkebleidorn@gmail.com)

Current Opinion in Behavioral Sciences 2017, **18**:79–82

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

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

http://dx.doi.org/10.1016/j.cobeha.2017.08.004

2352-1546/© 2017 Published by Elsevier Ltd.

In the digital age people generate behavioral footprints nearly constantly. These footprints agglomerate to 'big data' that offer psychological researchers unprecedented opportunities for tracking, analyzing, and predicting human behavior. A guiding assumption of this kind of research is that psychological characteristics (e.g., traits) 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 predictive of users' psychological characteristics [1,2••].

To test this hypothesis, researchers have begun to use machine learning approaches to predict users' personality characteristics from their digital footprints such as Facebook likes [3•] or Twitter profiles [4]. Most of this work has focused on developing reliable estimates of the 'Big Five' personality traits *neuroticism*, *extraversion*, *openness to experience*, *agreeableness*, and *conscientiousness* [5]. Identifying markers of these traits in big data has significant potential for furthering research on the structure and development of personality across languages and cultures. In this paper, we outline how this potential can be more fully achieved by situating machine learning research within a construct validation framework of test and theory development, with a particular focus on the content validity of computer-based assessments [6–9].

Construct validation and big data

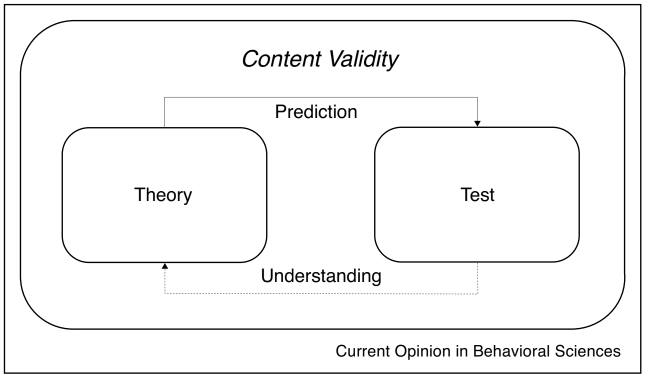
Construct validation emphasizes the bidirectional relationship between test development and theory development. Any scientific theory must be operationalized so that its variables can be measured and used in experiments. Any particular measure that operationalizes the variables in a theory will be imperfect. Thus, evidence that a measure is not performing as expected could mean that there is something inaccurate about the theory or that there is something inadequate about the measure (cf., [10,11]). From this perspective, establishing *content validity* amounts to connecting a test with the theoretical variable that test is meant to measure.

Figure 1 depicts the bidirectional relationship between a latent or theoretical concept and a manifest or measured variable. The solid arrow from theory to test development represents the half of the theory-test relationship focused on the degree to which tests adequately operationalize theories. This arrow is labeled with the term 'prediction' because tests that adequately operationalize theories should be more effective for predicting behavior. For instance, an adequate personality measure should be able to predict individual differences in personality traits in a manner that corresponds to previous research on how individuals differ as assessed by other instruments.

The dashed arrow in Figure 1 from test to theory development represents the other half of the relationship. This arrow is labeled 'understanding' because it focuses on how the development and refinement of a personality measure's content can provide new insights about personality theory.

So far, machine learning research has predominantly focused on the 'prediction' of personality differences [12••]. These studies generally maximize the convergence between computer-based and other (typically questionnaire) measures of personality traits. This process assumes the content validity of the other measure in

Figure 1



Construct validation.

the absence of any theoretical concerns. This approach stands in contrast to deductive/iterative approaches to establishing content validity in construct validation. Normally, researchers first define a universe of indicators based on their understanding of the latent constructs they are trying to measure, sample systematically within this universe to develop the test, and then refine the content of the measure based on additional validity data [13].

Thus far, little research has been done to establish or evaluate content validity in big data personality measures. For example, Kosinski, Stillwell, and Graepel (2013) found that Facebook users' personality traits can be predicted to a high degree of accuracy based on their *likes*. Facebook likes allow users to connect with objects that have an online presence (e.g., products, movies, etc.) and are shared with the public or among Facebook friends to express support or indicate individual preferences [14]. Some of the most highly predictive likes seemed face valid and tied in with previous research, as in the case of 'Cheerleading' and high extraversion. Yet, many other highly predictive likes were rather elusive and raise questions concerning the measure's content validity; as in the case of 'Getting Money' and low neuroticism. The relatively unexplored content validity of computer-based personality measures complicates the interpretation of findings that are solely based on these scales and constrains the degree to which these measures can be used to advance personality theory. As we will outline below, considerations of content validity can be particularly fruitful for theory, precisely in those cases where the content might bear limited apparent relations to the trait [8].

In summary, previous machine learning studies have predominantly focused on the 'prediction' of personality differences. Machine learning algorithms have rarely been used to further our understanding of personality structure, processes, and development, as indicated by the dashed 'understanding' arrow in Figure 1. As such, machine learning research on personality has only focused on one half of the construct validation process. We see this as an appreciable gap in this line of work, and believe that there is significant potential, because of some of the unique features of big data, to use machine learning to develop new insights about personality through an enhanced focus on content validity.

What can big data tell us about personality?

We focus on three broad areas in which big data could inform personality theory, all of which would require a more thorough investigation of the content validity of computer-based personality measures: (1) the *delimitation of trait content*, (2) the articulation of how *developmental processes* impact personality measures, and (3) the *identification of culture-specific personality markers*. Fundamental to each of these domains is the distinction between manifest indicators and latent variables. In Figure 1, the test represents the manifest indicators of a latent variable or set of variables as indicated by the theory. As described above, from a construct validation perspective, establishing content validity essentially amounts to closing the gap between manifest (measurement) and latent (theoretical) variables.

Trait content

The content of the Big Five were generated lexically [15]. Early personality researchers assumed that most of the important trait-relevant information would be contained within language, because an important function of language is to communicate about what people are like and how they differ from each other. This information was combined empirically primarily through factor analyses of the trait descriptors derived from the lexicon. Decades of research led to the general consensus that the Big Five represent broad traits that capture how more specific behaviors, thoughts, and feelings tend to covary in the population (e.g., [16•]). Yet, the emerging structure and content of traits inevitably depend on the universe of items that were considered.

Big data and machine learning approaches have the potential to broaden and refine our understanding of the structure and content of the Big Five. A unique feature of big data is that they are wide ranging and inductive. The relatively unconstrained access to digital traces of personality allows researchers to detect and include personality indicators that might not have been conceived of by lexical or deductive approaches. In other words, big data offer researchers access to a new lexicon which contains a wealth of data that are often personal and otherwise difficult to assess [2••,17]. To the degree that these data contain new and hitherto unexplored trait-relevant content, they could greatly advance our understanding of the specific ways in which personality traits manifest in online environments and beyond.