Contextualized Construct Representation: Leveraging Psychometric Scales to Advance Theory-Driven Text Analysis

Mohammad Atari[1], Ali Omrani[2], Morteza Dehghani[2,3]

[1]Department of Human Evolutionary Biology, Harvard University

[2]Department of Computer Science, University of Southern California

[3]Department of Psychology, University of Southern California

Author Note

Mohammad Atari https://orcid.org/0000-0002-4358-7783

Ali Omrani https://orcid.org/0000-0002-0507-0759

Morteza Dehghani https://orcid.org/0000-0002-9478-4365

All materials are available at https://osf.io/bu6wg/. Mohammad Atari was supported by the John Templeton Foundation (Grant ID 62161). Ali Omrani and Morteza Dehghani were supported by DARPA HR001121C0165. We thank members of the Culture, Cognition, and Coevolution Lab at Harvard University for their valuable feedback on earlier stages of this project. We thank Tom Zhang for his help with developing and debugging the accompanying R package. We are also grateful to Brendan Kennedy and Aida Davani for their feedback on an earlier draft of the manuscript. Our R package, Python library, and online demo are available at: https://github.com/Ali-Omrani/CCR

Correspondence concerning this article should be addressed to Mohammad Atari, Department of Human Evolutionary Biology, Harvard University, 11 Divinity Ave, Cambridge, MA 02138. E-mail: matari@fas.harvard.edu

Abstract

Over the past decades, text-analysis methods have been slowly integrated into the repertoire of methods used to reliably measure psychological constructs. Yet, many of the existing computational methods in psychological text analysis remain atheoretical and lack the interpretability that social sciences are accustomed to and desire. Here, we introduce a novel method for theory-driven text analysis by bridging the power of contextual language models and common psychometric scales. The new technique, which we call Contextualized Construct Representation (CCR), retains high levels of interpretability and top-down flexibility but makes use of state-of-the-art language models developed in natural language processing (NLP). CCR is a flexible technique that will be able to adapt to the continuously progressing set of tools for language modeling. We discuss how our proposed technique quantifies psychological information in textual data, and demonstrate in two studies (N = 2,996) that CCR outperforms other top-down methods (i.e., word-counting and word-embedding representations) in predicting an array of psychological outcomes common in social and personality psychology, including moral values, the need for cognition, political ideology, strength of norms, and cultural orientation. We provide an accompanying R package, a Python library, and an online interface for researchers to conveniently use CCR in their research.

Keywords: theory-driven text analysis, natural language processing, large language models, computational psychology, psychometric scales.

Contextualized Construct Representation: Leveraging Psychometric Scales to Advance Theory-Driven Text Analysis

Throughout the history of psychological science, researchers have mapped and measured the human mind primarily through questionnaires. Validated questionnaires still remain one of the most commonly used tools across the social sciences. In just over a century since the first questionnaires were developed by psychologists, a tremendous literature on multi-item psychometric scales has amassed, measuring constructs as diverse as explicit bias, personality traits, psychopathology, and cultural orientation. Through precise, validated, and consistent wording, these tools (often self-reports) provide a direct way to measure complex psychological phenomena (Clark & Watson, 1995; Schwarz, 1999).

More recently, contemporaneous with the emergence of "big data" methods, psychologists have been using indirect assessments of psychological constructs, for example quantifying people's personality traits and values based on their social-media posts (H. A. Schwartz et al., 2013). An array of text-analytic methods — often developed in Natural Language Processing (NLP) — have been adapted for use in psychological studies (Park et al., 2015). Previous methods fall under two broad categories: top-down and bottom-up text analysis. Top-down methods take expert knowledge and apply them to language data using computational techniques, whereas bottom-up methods use NLP algorithms to represent language (Kennedy et al., 2022). The priority of the former is extracting psychologically relevant information from text, primarily by developing *dictionaries*, categories of words each with a meaningful link to a construct of interest. The latter class of methods is primarily focused on modeling language wherein a successful language model looks at language as a whole rather than purely in terms of a particular construct. Research has shown that top-down methods are highly desirable because of their interpretability and ease of use, but they tend to fall short of capturing sufficient variance in intended constructs. Bottom-up methods, on the other hand, substantially outperform the first class of methods in terms of explained variance, but are considered a "black box" in the sense that they are uninterpretable and atheoretical (Kennedy et al., 2021).

Here, aiming to bridge between these classes of methods, we advance theory-driven psychological text analysis by using the power of large language models. We introduce a novel technique called Contextualized Construct Representation (CCR) which quantifies psychological constructs using contextual language models developed in NLP, by measuring the distance between the vectorized representation of a given text and that of questionnaire items. Critically, representing text using language models has revolutionized the field of bottom-up text analysis, with the recent wave of contextualized language models able to capture the compositional meaning of sentences (and items). We move from "dictionaries" to questionnaire items, as the latter gives us the context we need to better capture complex psychological phenomena in language data. In what follows, we briefly review how multi-item scales have become the bedrock of psychological science — especially in social and personality psychology — and how modern computational methods can be sensibly applied to advancing the representation of constructs by embedding entire questionnaire items. Finally, we demonstrate the predictive power of CCR in two studies, showing that it outperforms other top-down methods that rely on dictionaries.

A Brief History of Psychometric Scales

At the beginning of scientific psychology in early 20th century, psychologists faced the challenge of developing valid measures without having access to a knowledge base or theoretical framework to rely on. The lack of knowledge about psychological concepts was a hurdle for constructing tests. As researchers developed somewhat primitive measures and tested simple hypotheses about human behavior, a basic body of knowledge began to shape (Strauss & Smith, 2009). While there are some questionnaires developed in the last decades of the 19th century (see White, 1992), one of the earliest measures in the history of test-construction efforts is the Woodworth Personal Data Sheet, a measure created during World War I to help the U.S. Army screen out individuals who might be at risk for "shell shock."

In the early 1950s there was an emerging concern with theory development that led to Meehl and Challman's introduction of the concept of "construct validity" — collecting evidence that measurement instruments actually measure the constructs researchers claim they measure — as a part of the work of the American Psychological Association's Committee on Psychological Tests (American Psychological Association, 1954; Cronbach & Meehl, 1955). In the last several decades, psychologists have designed numerous programs of research and thousands of studies to identify, define, and measure some postulated attributes of people, which are typically unobservable (e.g., attitudes). This has led to the development and validation of thousands of multi-item psychometric scales, that is, multiple items measuring a focal construct in a reliable and valid manner and yielding numeric data (Clark & Watson, 1995; Robinson, 2018).

While debates and disagreements about the co-evolution of theory and method continue (e.g., Fried, 2015), multi-item psychometric scales remain the most powerful and omnipresent mode of measurement in social and personality psychology (see Flake et al., 2017; Flake & Fried, 2020). Researchers from neighboring fields such as cultural studies, political science, marketing, and public policy often use psychometric methods to develop new multi-item scales. Even when researchers do not follow a rigorous construct-validation process, they often use multiple face-valid items to measure the construct of interest (Flake et al., 2017). There are typically multiple alternative scales for psychological constructs, and in the case of more common constructs (e.g., depression, religiosity), there are sometimes dozens of available psychometric scales, translated into many languages and adapted for different populations (DeVellis & Thorpe, 2021).

Psychometric scales — while relatively "cheap" tools that only take minutes to complete — have substantially expanded our understanding of human sociality, personality, psychopathology, culture, and cognition. While psychology has been considered the "science of behavior", introspective self-reports and questionnaire ratings have become increasingly popular in *direct* measurement of human behavior and cognition (Baumeister et al., 2007). Psychometric scales typically comprise a number of items (i.e., a question or a declarative statement about the variable of interest) that respondents are required to rate and are, therefore, usually considered a self-report research method (Robinson, 2018; Stone et al., 1999); although the same scales can be used to rate others, as is the case in peer-, parent-, clinician-, and supervisor-ratings (e.g., Kolar et al., 1996). A common element, and perhaps limitation, of psychometric scales, is that the individual should be known and accessible to be assessed (either via self- or other-reports). There are at least two limitations to this approach: First, psychometric scales have limited utility in some cases as they can suffer from socially-desirable responses (Paulhus, 1984). For example, in some cases people actually believe their positive self-reports (i.e., self-deception), and sometimes the respondent consciously dissembles (i.e., other-deception) (Sackeim & Gur, 1979). Second, researchers sometimes want to quantify psychological attributes in "unreachable" individuals, that is, individuals who are not present and/or willing to complete a scale (e.g., politicians, social-media users) (Kosinski et al., 2013) or people who have lived in the past (Muthukrishna et al., 2021). But some psychological processes can be *indirectly* inferred by looking at the language that people have produced.

Indirect Psychological Assessment Through Language

Considering the above-mentioned limitations with self-report scales, researchers have turned to using people's textual data to unobtrusively infer psychological attributes such as attitudes, moral values, and political preferences. For example, Körner et al. (2022) conducted text analysis on over 15,000 tweets made during the U.S. presidential election in 2020 and found that, in comparison to Trump, Biden used a greater number of words associated with virtue, honesty, and accomplishment. In the case of "dead minds", we can similarly text-analyze historical written records to infer important psychological constructs (Atari & Henrich, 2023). For example, Boyd and Pennebaker (2015) analyzed the works of three authors — William Shakespeare, John Fletcher, and Lewis Theobald — and found that all these authors displayed highly unique patterns of language use across their solo works. Importantly, psychologists often resort to text analysis because they are not able to use their gold-standard measures to directly observe and record real-world behavior.

A number of unobtrusive text-analytic approaches have been developed, often adapted from NLP, to measure cultural and psychological attributes based on people's (or societies') textual records (Boyd & Schwartz, 2021; Jackson et al., 2021). Broadly, psychological language analysis can fall under two categories: top-down and bottom-up methods.

Psychological Language Analysis Using Top-Down Methods

In top-down text analysis, expert knowledge is "offloaded" to computers. Typically, psychologists have developed word lists, or *dictionaries*, to measure a construct (see Kennedy et al., 2022; Pennebaker & King, 1999). Dictionary-based text analysis is highly intuitive and interpretable. The key assumption is that the prevalence of a set of terms meaningfully corresponds to the construct of interest. For example, when someone is in a church, they use a lot of words related to religious beliefs (e.g., "pray," "God"). Someone who comes from a collectivist culture might refer more frequently to group-related units such as "community" and "family" compared with their counterpart in an individualistic culture who might, in contrast, use "unique" and "independent" more often (e.g., Greenfield, 2013). If words from a dictionary are being used more often in a given text, it is more likely that the speaker is thinking in a certain way, experiencing a particular emotional state, values a particular set of principles, or belongs to a certain religious affiliation.

Although a priori dictionaries are a useful method for quantifying psychological constructs in text, researchers must subjectively specify a set of words (without considering the diverse contexts in which they might appear) that capture a construct, which is time-dependent, labor-intensive, influenced by researchers' biases, and sometimes practically impossible. If a dictionary's conceptual coverage is too narrow, important parts of the construct will be omitted, inflating false-negatives — compromising the validity of such text-based measures. However, if the coverage of a dictionary is too broad, the text would match when it should not, producing false-positives and jeopardizing a dictionary's construct validity. Finally, another problem with applying pre-defined dictionaries to textual data is *polysemy* which means that some words — in fact, many words — have more than one distinct meaning. For instance, the word "bank" can refer to a river bank or a financial institution (see Kennedy et al., 2022).

Psychological Language Analysis Using Bottom-Up Methods

While expert knowledge can be effectively applied to the study of textual data using top-down approaches, much of modern NLP research takes a less theory-informed starting point (Kennedy et al., 2022). Such bottom-up approaches, also known as data-driven approaches, include latent semantic analysis, topic models, and text embedding.

Latent semantic analysis uses word co-occurrence patterns (i.e., collocations within the same document) to create a semantic space that can then be used to quantify the semantic similarity of terms (Deerwester et al., 1990). This method has been applied in the study of morally loaded topics such as terrorism and abortion on social media (Sagi & Dehghani, 2014). Although latent semantic analysis remains a useful method in small corpora, its principal historical contribution has been motivating probabilistic methods to achieve the same outcome. One of the most popular and effective probabilistic models is Latent Dirichlet Allocation (LDA; Blei et al., 2003) which assumes that words in each document are sampled from a set of "topics" where each topic is a cluster of semantically coherent words. LDA can serve as an exploratory and useful method to discover new categories of words by correlating them with psychological variables. In the "Open Vocabulary" approach to text analysis (H. A. Schwartz et al., 2013), patterns of language use are discovered in a bottom-up way that are correlated with dimensions of personality

(Park et al., 2015), symptoms of depression (Guntuku et al., 2017), and geographic distribution of psychological variables such as well-being (Jaidka et al., 2020).

Deep-learning methods have recently become the standard for modeling language data. Deep learning, defined as multi-layer networks of artificial “neurons” (LeCun et al., 2015), refers to a family of data-driven techniques that can approximate any function (Hornik et al., 1989) yet are generally viewed as “black boxes,” which produce predictions from inputs without interpretability. These models are an extension of the “bottom–up” methods typified by models like LDA. These models offer a similar approach to LDA, only with more modeling capacity. In NLP, this capacity includes the ability to optimize the prediction of linguistic modeling objectives, such as predicting the “next word” in a sequence or predicting the distribution of a word’s “context window” (for a review, see Kennedy et al., 2022).

The deep-learning revolution in NLP led to two main families of techniques: word embeddings and language models. Word embeddings are an efficient numerical representation of words that allow words with similar meanings to have a similar representation (Mikolov et al., 2013; Pennington et al., 2014). Word embedding models map words into a continuous vector space, similar to LSA. Word embeddings refer to a broad class of methodologies that have become a staple of NLP research in the past 10 years. Interestingly, the initial word embeddings were based on (and continue to be motivated by) the idea of distributed word semantics, pursued via methods such as LSA. The application of the distributional hypothesis (“You shall know a word by the company it keeps”; Firth, 1957), in particular, continues to be the foundation of representation learning for language, extracting semantic information about words from their “neighbors,” or the other words with which they frequently co-occur.

While the pre-trained word embeddings such as GloVe have been useful in many domains, they have a limitation – word embedding models presume that a word’s meaning is relatively stable across contexts (e.g., sentences). Word embeddings are useful in capturing the meanings of words, but context should be incorporated into holistic language modeling. In recent years, large pre-trained transformer-based language models have taken NLP by storm. These families of language models include the Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2018) and the Generative Pre-trained Transformer (GPT; Radford et al., 2018) models. These innovations were made possible after the development of a new neural architecture in 2018, the Transformer (Vaswani et al., 2017).

Combining Top-Down and Bottom-Up Methods

With the aim of combining top-down theories in psychology with the power of *word embeddings*, Garten et al. (2018) introduced *Distributed Dictionary Representation* (DDR), which functions by (a) specifying a small researcher-specified list of words that represent a construct; (b) representing these single words using a word-embedding model; (c) finding the centroid of the representations of these words as the representation of that dictionary; (d) finding the centroid of the embeddings of the words in the text (e.g., document); (e) calculating the cosine similarity between the low-dimensional representation of the dictionary and that of the text. Importantly, in using DDR, the text does not need to contain any of the words in the pre-specified dictionary in order for similarity to be computed. For instance, a text containing "dinner" might be considered highly similar to a dictionary consisting of "night," "supper," and "food." This entails that word lists can be short themselves, yet cover a construct somewhat adequately. This eliminates one of the major limitations of word-counting methods. In validation studies, DDR has been shown to outperform traditional word-counting methods using only four words for each construct in some domains (Garten et al., 2018). DDR and similar methodologies based on word embeddings (e.g., Charlesworth et al., 2021; Garg et al., 2018) provide evidence that a short set of quintessential words for each psychological construct can be used to effectively measure the construct of interest in textual data.

Although DDR has been shown to be a more reliable text-based measure than word-counting, at least in some domains such as moral values (Hoover et al., 2020; Kennedy et al., 2021), this approach is not without limitations. First, some psychological constructs are too complex to be represented using a few words. Second, if the representations of chosen words are not close enough, their centroid might completely fail to represent the desired construct of interest; in other words, the representation of the construct is too sensitive to what seed words are selected to represent the construct. Third, polysemous words are almost always bad choices for DDR and similar methods because static representations for individual dictionary words are decontextualized.

A different combinatory method that does not rely on dictionaries is called the "human-annotation-based" approach (Atari & Dehghani, 2022; Kennedy et al., 2022). In this approach, sentences or documents are expert-coded based on a theory-driven typology (i.e., the top-down component). Then, supervised machine-learning analyses are conducted to link the representations of texts (e.g., text embeddings) to the resulting labels. A presupposition of the human-annotation-based approach is that textual data include complex information; therefore, human annotators can best code nuances and compositionality of written text (rather than a dictionary). After a supervised classifier is trained, it can automatically label unseen texts. This approach is particularly useful in the study of framing and rhetoric wherein fixed dictionaries likely fail (e.g., moral rhetoric; Hoover et al., 2020).

Contextualized Construct Representation (CCR)

Since 2018, NLP research has been characterized by the emergence of many *large language models* (Bender et al., 2021; Devlin et al., 2018; Vaswani et al., 2017), that is, the use of statistical and probabilistic techniques to determine the probability of a given sequence of words occurring in a sentence. These models, sometimes referred to as "contextualized language models," allow us to reliably represent whole pieces of text, such as sentences or questions or vignettes (for an application in psychology, see Atari et al., 2020), rather than looking at individual words as is done in dictionary-based psychological text analysis.

In the last few years, after the introduction of transformers (Vaswani et al., 2017) as the dominant architecture for NLP, thousands of contextual language models have been developed that generate vectorized representations for text. Although these contextual language models were developed primarily to do language-completion tasks, they have been found to be highly accurate in other complex psychologically relevant tasks such as ethical reasoning (Jiang et al., 2021), inferring human personality structures from large corpora (Cutler & Condon, 2022), and analogical reasoning (Webb et al., 2022). Some authors have even made the case that theory-of-mind abilities may have spontaneously emerged as a byproduct of large language models' improving language skills (Kosinski, 2023).

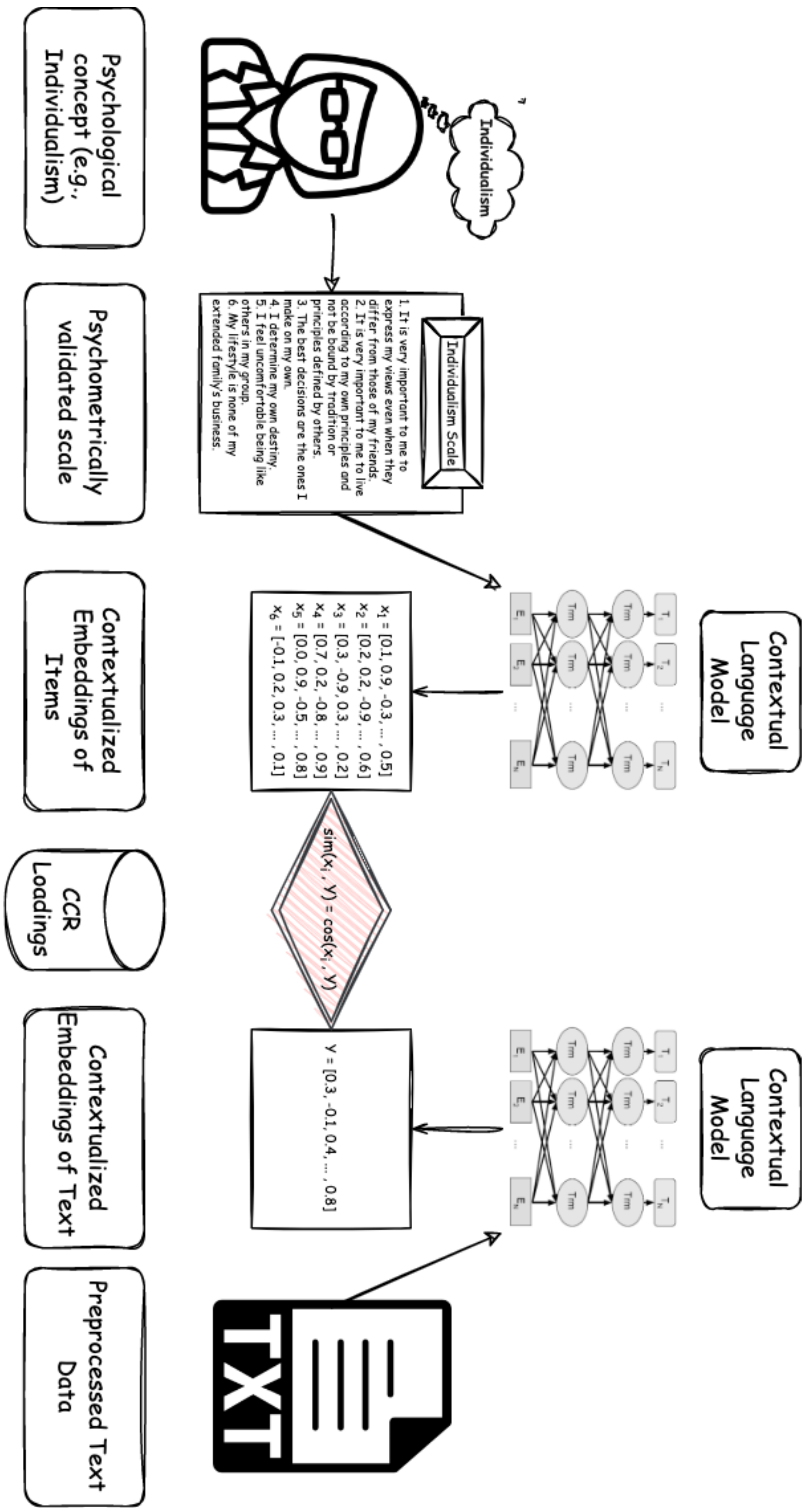
Here, we propose a novel method for psychological text analysis which (a) takes advantage of powerful language models developed in NLP, (b) does not need selecting "seed words" to represent a psychological construct; (c) takes advantage of rigorously validated measures using well-known psychometric methods in social and personality psychology. This method, which we refer to as *Contextualized Construct Representations* (CCR), can be summarized in five steps. The pipeline is schematically shown in Figure 1.

1. Select a psychometrically validated self-report measure for the construct of interest (e.g., the 6-item Individualism Scale, Oyserman, 1993).

2. Represent self-report items as embeddings using contextualized language models (e.g., BERT; Devlin et al., 2018).

3. Generate the embedding of the text to be analyzed using contextualized language models (e.g., BERT; Devlin et al., 2018).

4. Compute the similarity between the representation of text and those of questionnaire items (e.g., using cosine similarity) to arrive at a "loading" score. The higher this


Psychological concept (e.g., Individualism)
Individualism
Psychometrically validated scale
Individualism Scale
Contextualized Embeddings of Items
Contextual Language Model
E1
E2
...
EN
Trm
T1
T2
TN
CCR Loadings
sim(xi , Y) = cos(xi , Y)
Contextualized Embeddings of Text
Contextual Language Model
Preprocessed Text Data
TXT

1. It is very important to me to express my views even when they differ from those of my friends.
2. It is very important to me to live according to my own principles and not be bound by tradition or principles defined by others.
3. The best decisions are the ones I make on my own.
4. I determine my own destiny.
5. I feel uncomfortable being like others in my group.
6. My lifestyle is none of my extended family's business.

x_1 = [0.1, 0.9, -0.3, ... , 0.5]
x_2 = [0.2, 0.2, -0.9, ... , 0.6]
x_3 = [0.3, -0.9, 0.3, ... , 0.2]
x_4 = [0.7, 0.2, -0.8, ... , 0.9]
x_5 = [0.0, 0.9, -0.5, ... , 0.8]
x_6 = [-0.1, 0.2, 0.3, ... , 0.1]

Y = [0.3, -0.1, 0.4, ... , 0.8]

Figure 1

The Schematic Pipeline of the Contextualized Construct Representation (CCR) Method similarity, the higher the loading on the construct of interest.

CCR proceeds by representing both textual data of interest and theory-representative descriptions (here, questionnaire items) and calculates a “loading” score which is the semantic similarity of these vectorized representations. CCR has multiple advantages over prior methods. First, CCR relies on items in psychometrically validated scales. Social and personality psychology have a century-long history of validating self-report scales using a variety of methodologies. In other words, CCR obviates the use of dictionaries in contexts in which valid psychometric scales exist. Second, CCR makes use of contextual language models, but it is not dependent on a particular model. Therefore, as better language models are trained in NLP, their off-the-shelf models can be easily applied to CCR’s pipeline. Third, while translating psychological word lists to other languages is labor-intensive, expensive, and sometimes practically impossible (e.g., Matsuo et al., 2019), CCR can be easily applied to many languages as contextual language models are widely available for many languages, including low-resource languages. Of course, this also requires a valid questionnaire in the target language. For example, Multilingual BERT (mBERT) trained on 104 languages has shown good cross-lingual performance on a number of tasks (Wu & Dredze, 2020). This feature helps psychologists go beyond English-centric text analysis (Blasi et al., 2022).[1]

The Present Studies

Here, we report two studies showcasing that CCR outperforms existing top-down methods. In Study 1, we show how CCR outperforms both word-counting and DDR in predicting people’s self-reported (i.e., ground truth) moral values. Specifically, we apply the Moral Foundations Dictionary (Graham et al., 2009), DDR (Garten et al., 2018), and CCR to people’s Facebook updates to predict their scores on the Moral Foundations

[1] We caution against the blind application of CCR in other languages without adequate validation, as such translation-based analysis has yet to be tested.

Questionnaire (MFQ; Graham et al., 2009), a psychometrically validated measure of moral values. In Study 2, to make sure that the predictive power and high interpretability of CCR are not unique to the domain of morality and social-media data, we conduct a study in which people write about different topics and completed a broad array of psychometric scales commonly used in social and personality psychology. We demonstrate that CCR outperforms both word-counting and DDR in predicting an array of psychological constructs.

Study 1

In this study, we apply word-counting, DDR, and CCR to people's social-media posts in order to predict their self-reported moral values based on a psychometrically validated scale. In other words, we examine how well alternative methods can extract moral values from people's texts. In addition, we explore the kinds of moral values where CCR is most predictive compared with values that are less capturable using CCR.

Methods

Participants

As part of a collective effort on YourMorals.org, some users voluntarily completed self-report measures and consented to have their Facebook posts analyzed for research purposes. Initially, 4,414 respondents completed the study, volunteered their Facebook data for research, and had at least one post on Facebook (see Kennedy et al., 2021, for more information about the dataset and data-cleaning procedures). We removed participants who were younger than 18 or older than 65 years old (592 participants). All hyperlinks, picture links, and "mentions" were removed using regular expressions and the corpus was tokenized using the Natural Language Toolkit (NLTK; Loper & Bird, 2002). Overall, 53,901 of the 165,787 posts were removed that were either too short (less than five tokens) or were not classified as English during our initial text preprocessing. Finally, participants with fewer than 10 Facebook posts (1,131 participants) were excluded from the study, similarly to prior work (Kennedy et al., 2021; Park et al., 2015). This procedure resulted in 2,691 participants. In the full sample of 2,691, participants self-reported age (M_{age} = 32.8 years, SD_{age} = 11.9 years) and sex (57.0% male).

Measures

All participants completed the 30-item Moral Foundations Questionnaire (MFQ; Graham et al., 2011), which measures the five moral foundations of care, fairness, loyalty, authority, and purity. In the first section of the measure, items are rated along a 6-point Likert-type scale ranging from 0 (*Not at all relevant*) to 5 (*Extremely relevant*). In the second section, the items are rated along a 6-point Likert-type scale ranging from 0 (*Strongly disagree*) to 5 (*Strongly agree*). An example item is "I think it's morally wrong that rich children inherit a lot of money while poor children inherit nothing". The internal consistency coefficients were .70, .64, .72, .78, and .87 for care, fairness, loyalty, authority, and purity, respectively.

Word Counting

To get the word count of each dictionary for each participant's language data, we first construct a regular expression by expanding each word to its other morphological forms. Then we use the Regular Expression Operations library in Python to find the number of occurrences of terms in each dictionary in each participant's language data.

Distributed Dictionary Representation (DDR)

There are two components to calculate a DDR *loading*: (1) the input text's embedding and (2) the centroid of embeddings of seed words. To get the input text's embedding we first use the CountVectorizer from scikit-learn library (Pedregosa et al., 2011) to tokenize each response into words. Then we take the average embedding of all the words in the response as the response embedding. To construct the dictionary embedding, we calculate the average embedding of a sample of 25 words from each dictionary. Finally, the DDR loading is simply computed as the cosine similarity between the response embedding and the dictionary embedding. Our current analysis uses the GloVe pre-trained word embeddings (Pennington et al., 2014).

Contextualized Construct Representation (CCR)

Since CCR relies on semantic sentence similarity, we use Sentence-BERT (SBERT; Reimers & Gurevych, 2019) which achieves state-of-the-art performance on semantic-textual-similarity tasks. SBERT derives sentence embeddings that can be compared using cosine similarity by further finetuning pooled BERT embeddings using a "siamese" and triplet network (Schroff et al., 2015) on two Natural Language Inference (NLI) datasets. Applying the "siamese" and triplet architecture results in a mapping from text to vector space that allows for the comparison of sentence embeddings using cosine similarity.

For our current studies, we first get the embedding of all items in a psychometric scale using the pre-trained weights from "all-MiniLM-L6-v2" SBERT's implementation in Python (Reimers & Gurevych, 2019). These models take the whole context of each item into account when calculating the embeddings. We follow the same approach to get the embedding of the input text. The CCR score for each question is then defined as the cosine similarity between the contextualized embeddings of the input text and the scale item(s).

Analytic Strategy

We use indices based on word-counting, DDR, and CCR to train a linear model with 10-fold cross-validation. We trained a linear regression model using a randomly selected subsample including 90% of the data and then calculated the explained variance (R^2) in the remaining 10%. In order to produce variance around these out-of-sample effect size estimates, we bootstrapped 100 times for each construct (care, fairness, loyalty, authority, and purity).

Results

Comparisons of different methods in all domains can be seen in Figure 2. As can be seen, CCR substantially outperformed both word counting and DDR in predicting people's actual self-reported values in all five domains. In care, the out-of-sample explained variance of CCR was larger than that of word counting (t = 91.04, $p < .001$) and DDR (t = 25.20, $p < .001$). In fairness, CCR outperformed word counting (t = 60.55, $p < .001$) and DDR (t = 82.70, $p < .001$). In loyalty, CCR outperformed word counting (t = 162.17, $p < .001$) and DDR (t = 135.54, $p < .001$). In authority, CCR outperformed word counting (t = 149.05, $p < .001$) and DDR (t = 136.01, $p < .001$). Finally, in purity, CCR outperformed word counting (t = 198.24, $p < .001$) and DDR (t = 205.07, $p < .001$).

Discussion

The results of this study showed that applying CCR to textual data from social media returns substantially more accurate text-based measures of moral psychological constructs. These results are particularly important since much of prior work in moral text analysis has used dictionary-based methods (e.g., Brady et al., 2017; Buttrick et al., 2020) and DDR (e.g., Candia et al., 2022; Wang & Inbar, 2021). The power of contextual language models in two domains can explain the power of CCR in outperforming its predecessors: (a) quantifying the nuances of moral constructs that are captured in psychometric scales, but absent in dictionaries; and (b) quantifying the contextual information present in social-media language such as Facebook updates. Notably, CCR is most capable of extracting information about purity and authority, and least powerful in extracting signal about people's fairness values, replicating the findings of (Kennedy et al., 2021). Importantly, CCR explains more variance compared with its counterparts within the top-down class of text analysis (i.e., word-counting and DDR), but comparing our results with those of Kennedy et al. (2021) suggests that using bottom-up methods (e.g., using all BERT features in a regression) explains more variance. As such, it is important

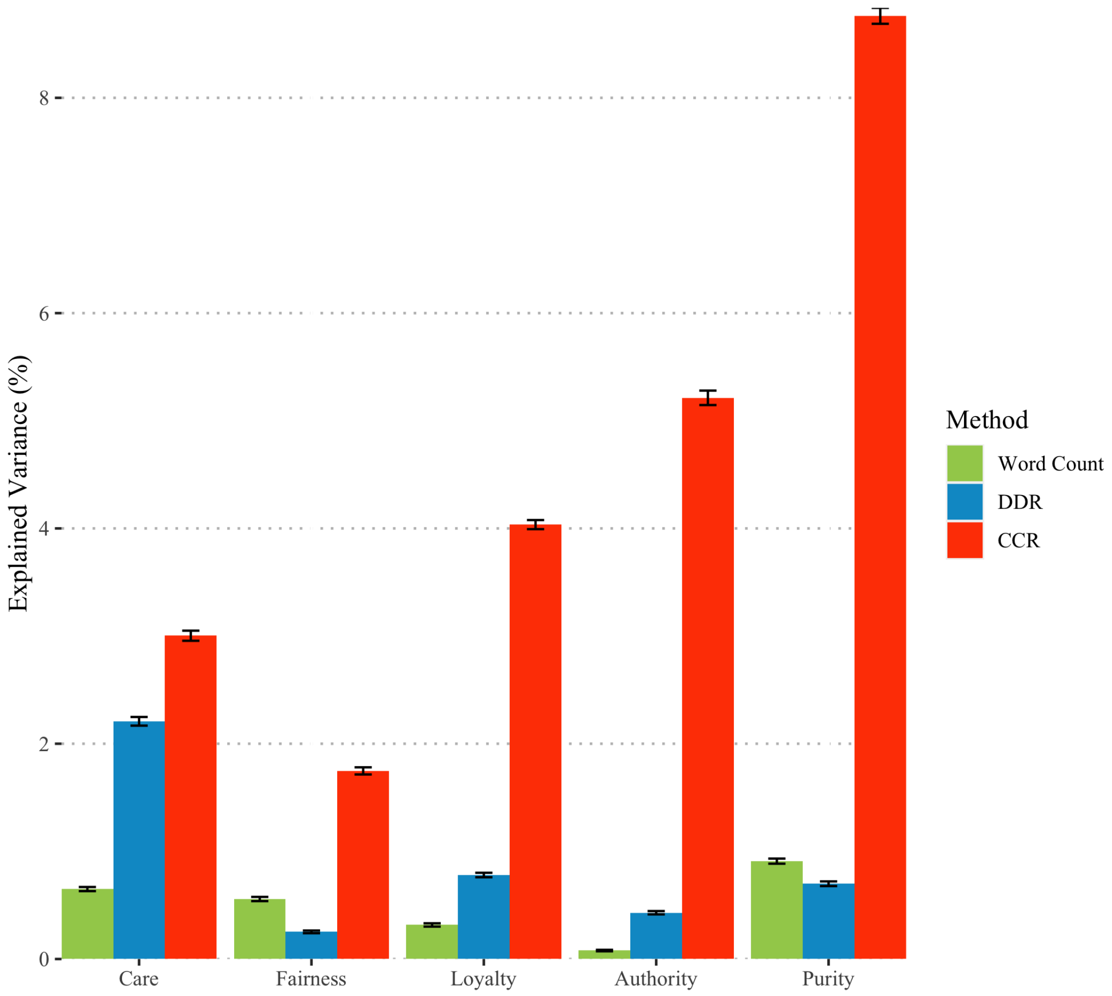


Figure 2

Comparing out-of-sample explained variance in predicting self-reported moral values (Study 2). Whiskers represent 95% Confidence Interval.

that each method is compared within its own class of methods.

Study 2

We designed this study to go beyond the domain of morality and to replicate our findings with more "structured" open-ended data rather than social-media posts. Here, we compare the performance of CCR and other methods to predict people's scores on an array of self-report measures commonly used in social and personality psychology. Specifically, we gave participants two open-ended prompts – one regarding their values and one regarding their everyday life – and asked them to write short essays in response (Boyd et al., 2015). Then all participants completed psychometric measures of moral foundations, personal values, religiosity, cultural tightness-looseness, individualism and collectivism, political conservatism, and the need for cognition.

Methods

Participants

We recruited 305 participants (63.6% women, $M_{age} = 40.0$ years, $SD_{age} = 13.9$ years) from the U.S. using TurkPrime. In terms of race-ethnicity, the majority of participants were White (68.9%), followed by Black (11.5%), and Asian (5.9%). All participants' first language was English.

Self-Report Measures

Participants first completed the two writing tasks (see below) in a pre-randomized order. Then, they completed all self-report measures in a pre-randomized order. In the end, participants provided their demographic details.

Values Essay. In order to assess participants' values in their own words, based on Boyd et al. (2015), we asked them to respond to the following prompt:

> *For the next 5 minutes (or more), write about your central and most important values that guide your life. Really stand back and explore your deepest thoughts and feelings about your basic values. You might think about the types of guiding principles that you use to make difficult decisions, interact with other people, and determine the things that are important in your life and the lives of those around you. Try to describe each of these values and their relationship to who you are. Once you begin writing, try to write continuously as much as you want.*

Everyday Essay. The following prompt was given with the aim of collecting natural language related to everyday behaviors (Boyd et al., 2015). This prompt was not intended to acquire a list of all behaviors in which all participants engaged. Rather, our goal was to acquire a snippet of participants' behavioral inventory that reflected common, psychologically meaningful behaviors. The writing prompt read as follows:

> *For the next 5 minutes (or more), write about everything that you have done in the past 7 days. For example, your activities might be simple, day-to-day types of behaviors (such as eating dinner with your family, making your bed, writing an e-mail, and going to work). Your activities in the past week might also include things that you do regularly, but not necessarily every day (such as going to church, playing a sport, writing a paper, having a romantic evening) or even rare activities (such as skydiving, taking a trip to a new place). Try to recall each activity that you have engaged in, starting a week ago and moving to the present moment. Be specific. Once you begin writing, try to write continuously as much as you want.*

Moral Foundations. After Study 1 and before Study 2, a new version of the MFQ was developed. All participants completed the new measure of moral foundations, that is, the 36-item Moral Foundations Questionnaire-2 (MFQ-2; Atari, Haidt, et al., 2022), which measures six moral foundations of care, equality, proportionality, loyalty, authority, and purity. One novel aspect of MFQ-2 compared with MFQ is that "fairness" has been split to "equality" and "proportionality." All items were rated along a 5-point Likert-type scale ranging from 1 (*Does not describe me at all*) to 5 (*Describes me extremely well*). An example item is "I admire people who keep their virginity until marriage."

Personal Values. All participants completed the 21-item Portrait Values Questionnaire (PVQ-21; S. H. Schwartz, 2003), which measures the importance that a person ascribes to each of the ten basic values. Each basic value is measured using two or three items. Each of the 21 items consists of a vignette that describes goals, aspirations, and wishes, which point to the importance of a value. Participants indicate on a six-point scale, ranging from 1(*Not like me at all*) to 6 (*Very much like me*). The items were adapted to reflect personal values of oneself. An example item is “It is important for me to behave properly at all times and not do anything that people consider wrong.”

Cultural Tightness-Looseness. This 6-item measure was designed to assess the degree to which social norms are pervasive, clearly outlined and imposed within groups (Gelfand et al., 2011). Participants rated their responses on a 6-point Likert-type scale ranging from 1 (*Strongly disagree*) to 5 (*Strongly agree*). Low scores indicate looseness while high scores indicate tightness of culture. An example item is “People in this country almost always comply with social norms.”

Individualism. All participants completed the 6-item Individualism Scale developed by Oyserman (1993) and recently adapted by Lin et al. (2021). All items were rated along a 5-point scale ranging from 1 (*Strongly disagree*) to 5 (*Strongly agree*). An example item is “I determine my own destiny.”

Collectivism. All participants completed the 6-item Collectivism Scale developed by Oyserman (1993) and recently adapted by Lin et al. (2021). All items were rated along a 5-point scale ranging from 1 (*Strongly disagree*) to 5 (*Strongly agree*). An example item is “In general, I accept the decisions made by my group.”

Need for Cognition. The need for cognition refers to people’s tendency to engage in and enjoy thinking (Cacioppo & Petty, 1982). All participants completed the 6-item Need for Cognition Scale (NCS-6; Lins de Holanda Coelho et al., 2020). Responses were given on a 5-point scale (1 = *Extremely uncharacteristic of me*; 5 = *Extremely characteristic of me*). An example item is “I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.”

Self-Rating of Religiosity. Participants completed a single-item measure of religiosity rated along an 11-point scale (0-10) (Abdel-Khalek, 2007; Afhami et al., 2017).

Political Ideology. All participants responded to an item asking about their ideology along 7-point scale ranging from 1 (*Very liberal*) to 7 (*Very conservative*). They also were asked to indicate their identification with the Democratic vs. Republican party on a 7-point scale, ranging from 1 (*Very Democrat*) to 7 (*Very Republican*). These two items were averaged to arrive at a single index of political conservatism (Jost & Thompson, 2000).

Text-Based Measures

To arrive at dictionary-based, DDR-based, and CCR-based measures of text, we used different dictionaries, short sets of seed words, and questionnaire items, described in the Supplementary Materials.

Analytic Strategy

Pre-trained Models. For DDR we used the GloVe pre-trained word embeddings (Pennington et al., 2014) to map words to a 300-dimensional vector space. For CCR, we used the "all-MiniLM-L6-v2" sentence encoder (Reimers & Gurevych, 2019) implementation and pre-trained weights from huggingface[2] to map both questions and input texts to a 384-dimensional vector space.

Text Preprocessing and Regressions. For our DDR analysis, we removed all hyperlinks, hashtags, and non-alphabetic characters using regular expressions and tokenized each input text using the Natural Language Toolkit (Bird et al., 2009). For our CCR analysis, we used the accompanying tokenizer for the "all-MiniLM-L6-v2" model using the huggingface library.

Statistical Analysis. We conducted linear regression models to predict scores on self-report measures. We used R^2 as a measure of performance. Here, since we had a large number of constructs, we examined the performance of CCR with that of word-counting and DDR across constructs (rather than within constructs using bootstraps).

[2] available at https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Results

We ran word-counting, DDR, and CCR on the text to predict people's scores on self-report measures. Figure 3 shows the power of these methods in predicting psychometric ground truth using the values essay. Figure 4 shows the power of these methods in predicting self-report scores using the everyday essay. In the values essay (Figure 3) and across 22 self-report measures, CCR outperformed word counting (t = 2.08, Welch-corrected df = 39.41, p = .044) and DDR (t = 2.88, Welch-corrected df = 33.00, p = .007). In the everyday essay (Figure 4) and across 22 self-report measures, CCR outperformed word counting (t = 3.65, Welch-corrected df = 21.89, p = .001) and DDR (t = 2.96, Welch-corrected df = 25.93, p = .007).

Next, we examined how strongly the performance of CCR is correlated with the performance of word counting and DDR. Across 22 psychological constructs, we correlated the explained variance of CCR with that of word counting and DDR. The performance of CCR was highly correlated with the performance of word counting (r = .84, df = 20, p < .001), suggested that for constructs in which word counting fares well, CCR is also highly powerful. In other words, some constructs are simply more capturable for both dictionary-based methods and CCR. Next, we correlated CCR's and DDR's performances for all 22 constructs, finding a strong relationship (r = .74, df = 20, p < .001). Again, for constructs that DDR performs well, CCR also performs well.

Finally, we explored how close CCR can get to bottom-up methods in predicting the ground truth. To do so, we represented people's open-ended responses using BERT (Devlin et al., 2018) and entered all 768 BERT features to predict the outcome. Extended results are present in Supplementary Materials. As expected for bottom-up methods, BERT features collectively explained more variance than did CCR (or other top-down methods). This pattern is consistent with prior work in which less interpretable, bottom-up methods were able to explain more variance in predicting self-report scores (Kennedy et al., 2021). Therefore, if researchers are primarily focused on prediction (rather than explanation), using bottom-up methods, or a combination of top-down and bottom-up methods such as human annotations, should be prioritized over highly interpretable methods such as DDR and CCR (see Yarkoni & Westfall, 2017).

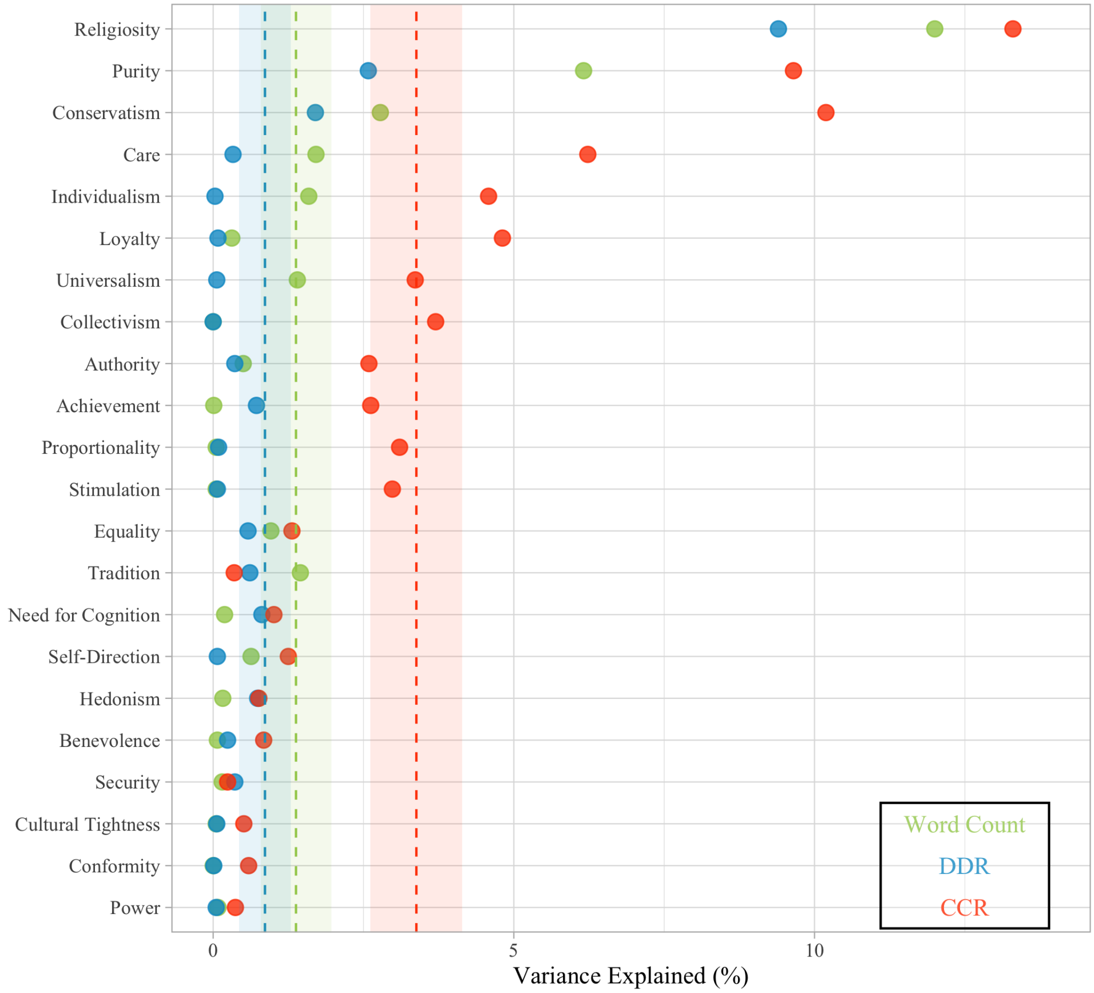


Figure 3

Comparing word-counting, DDR, and CCR on the values essay to predict self-report scores. Shaded areas represent Standard Errors.

Discussion

The findings of Study 2 further demonstrate the power of CCR in predicting the ground truth (i.e., scores on self-report measures) across a wide range of constructs common in social and personality psychology. These results show that CCR's performance

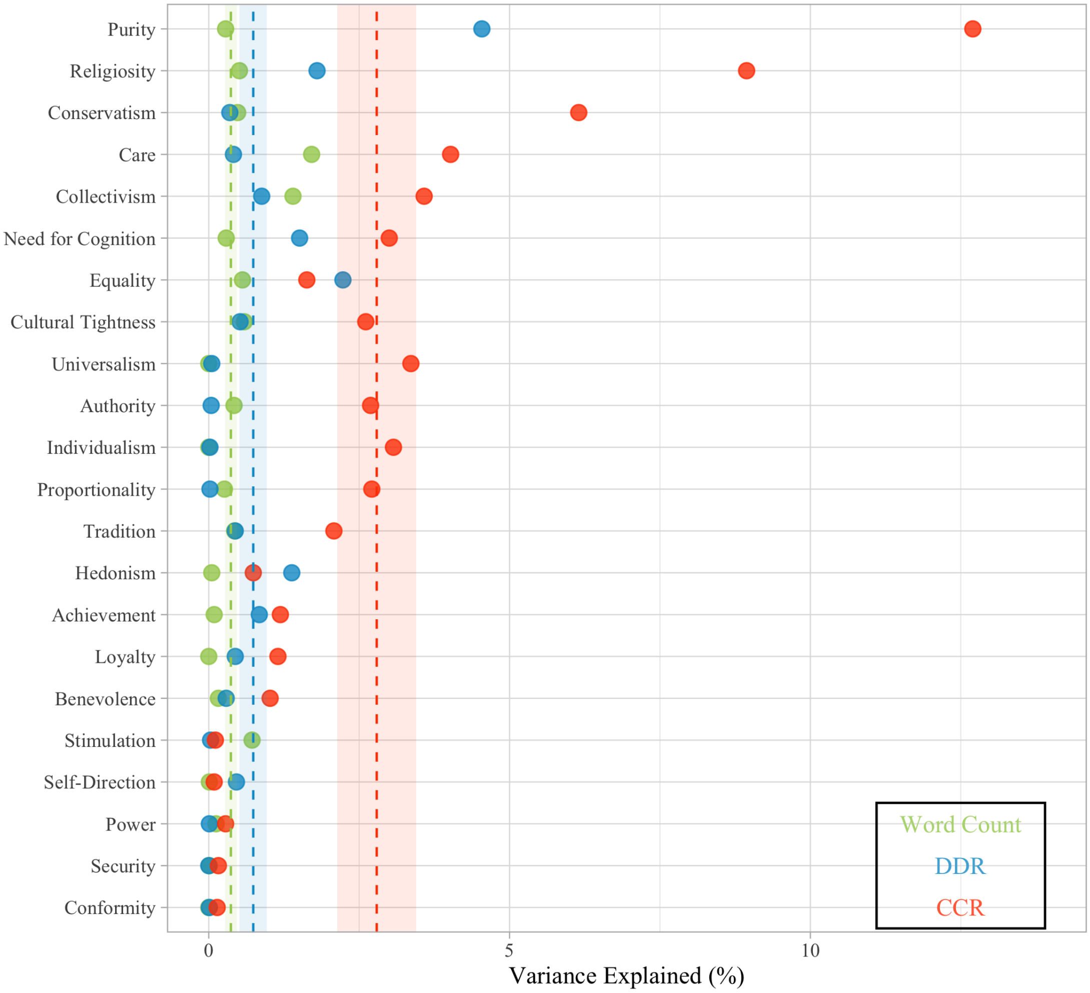


Figure 4

Comparing word-counting, DDR, and CCR on the everyday essay to predict self-report scores. Shaded areas represent Standard Errors.

is not limited to a particular domain of psychology, nor is it dependent on the nature of textual data (social media vs. open-ended questions).

General Discussion

With the advent of psychological text-analytic methods, it has become increasingly common for psychologists to use textual data to infer psychological states and dispositions. Measuring psychological construct in text generates new opportunities to expand the remit of psychological science by allowing us to ask new questions about new sources of data such as social media, avoid social desirability biases inherent in self-reports, or extend previous theories to a wider variety of contexts such as temporal variations in psychology and cultural change. Here, we propose a new technique for capturing psychological constructs in textual data. This new method (CCR) makes use of items in psychometrically validated scales to quantify the similarity between the semantic representation of a given text and that of scale items. Unlike prior studies that have largely relied on word lists (or context-free *dictionaries*), here we combine holistic approaches to language modeling (i.e., contextual language models) and psychometric scale items which include carefully worded and nuanced statements about psychological phenomena. Collectively, our results in two studies demonstrate that CCR substantially outperforms its dictionary-based predecessors across common topics in psychology (e.g., morality, need for cognition, cultural orientation), demonstrating its potential to integrate insights from psychometric scales and state-of-the-art language models developed in NLP.

Importantly, our proposed approach relies on an extensive body of work that has developed and psychometrically validated scales. CCR preserves the theory-driven nature of psychological text analysis while removing the barrier of constructing word lists. We, therefore, bridge decades of work in psychometric measurement with cutting-edge NLP models to capture psychological information. This method can be of utility across many subdisciplines beyond psychology, ranging from political history to data science since it is easy to use and flexible to adapt to future developments in NLP.

Many social psychological phenomena (e.g., stereotypes, ambivalent sexism, power) are hard to capture in textual data using context-free dictionaries. For example, stereotypes are dependent on context and social groups, and can drastically change in a matter of years (Charlesworth & Banaji, 2022; Charlesworth et al., 2022). While there have been important studies in developing word lists to study dynamic constructs, such as stereotypes, in textual data (e.g., Nicolas et al., 2021), CCR can be particularly helpful in capturing such constructs because it relies on a holistic approach to language modeling rather than fixed word lists. Future research is recommended to incorporate psychometric scales (as well as validated vignettes, see for example, Clifford et al. 2015) into theory-driven text analysis to quantify social psychological phenomena.

In terms of applications to personality psychology, CCR can complement bottom-up approaches in developing models of human personality as well as measuring personality dimensions in textual data. Natural language has been central in the study of human personality since the field's inception. Lexical approaches to personality factor-analyze the words (or descriptors) that people use to describe one another to arrive at a small and manageable number of personality dimensions. This approach generated the Big Five model that remains a major framework in contemporary personality science (Ireland & Mehl, 2014). The Lexical Hypothesis (or postulate) states that people encode in their everyday languages all those differences *between* individuals that they perceive to be salient and that they consider to be socially relevant in their everyday lives. In other words, language allows people to describe the specific ways in which we differ (e.g., "smart", "talkative", or "brave"). While word-counting methods have been applied to the study of personality for at least two decades (Holtzman et al., 2019; Pennebaker & King, 1999), researchers have recently begun to incorporate large language models into personality research. In their "Deep Lexical Hypothesis", Cutler and Condon (2022) introduced a method to extract adjective similarities from language models as done with survey-based ratings in traditional psycholexical studies but using millions of times more text in a natural setting.

CCR can address several of the ways that traditional methods of personality research are limited. CCR eliminates the need to focus on a small number of traits or descriptors. As evidenced in our Study 2, CCR-based measures can reliably capture people's attributes, values, and cultural orientations, hence assuming adequate text from a person, we can "ask" that person thousands of contextualized questions without participant fatigue. Descriptors or adjectives can be transformed into contextualized full sentences as done by Cutler and Condon (2022) (e.g., “talkative” can be structured into “she has a talkative personality” or “everyone thinks that she is quite a talkative person”), which can be fed into the CCR algorithm. Text-analysis methods have already increased the scope of personality psychology’s long-standing questions and generated novel methods of addressing old debates (e.g., Bleidorn & Hopwood, 2019; Park et al., 2015). CCR expands and complements this emerging methodological toolbox. Finally, CCR can be applied to different (alternative or rival) personality theories (Srivastava, 2020) as long as that theoretical framework is accompanied by a multi-item scale — which is almost always the case for established models such as the Big Five and HEXACO (Ashton & Lee, 2007).

In addition to the above applications, CCR can benefit historical text analysis in the emerging field of historical psychology (Atari & Henrich, 2023). What would it be like if we could hand out questionnaires to people who lived hundreds of years ago, long before psychological theories were born? While reading “dead minds” was forlorn only a few decades ago, text-analytic methods get us closer to making this previously far-fetched goal attainable. The “dead minds” represent an extraordinarily diverse subject pool (Muthukrishna et al., 2021), hence studying them via text-analytic tools can uncover new insights about human sociality and culture that were previously invisible to psychologists (Atari & Henrich, 2023). CCR holds the promise of extracting reliable psychological signal from historical texts. Future research is encouraged to replicate prior historical psychological findings (e.g., Atari, Reimer, et al., 2022; Charlesworth et al., 2022; Greenfield, 2013; Jackson et al., 2019; Martins & Baumard, 2020) using CCR. Importantly, however, we recommend adopting historical language models in such analyses (see Manjavacas & Fonteyn, 2021).

Finally, another application of CCR is to examine psychological processes in social-media data. In the last decade, a large number of social scientists have turned to social-media data (e.g., Twitter, Facebook) to examine psychological theories in an ecologically valid environment. Since many of these efforts are theory-driven, they have primarily relied on top-down methods such as dictionaries (e.g., Brady et al., 2017; Burton et al., 2021; Simchon et al., 2020), DDR (e.g., Candia et al., 2022; Wang & Inbar, 2021), and human annotations (e.g., Hoover et al., 2020; Mooijman et al., 2018). Since we show that CCR can outperform word-counting and DDR methods, future theory-driven work can rely on CCR as a more robust technique that captures the complexities of social-media language. We note that expert annotations of textual data remain the modus operandi of extracting psychological information from social-media textual data, especially when investigating constructs that require the complex social and contextual knowledge that people use on social media (e.g., Atari et al., 2023; Kennedy et al., 2022). However, such efforts are often time-consuming and labor-intensive. CCR provides a less demanding replacement for prior text-analytic methods such as DDR and word counting. Furthermore, human annotations can be used to validate CCR scores when applied to new domains.

Concluding Remarks

Psychometric scales are the backbone of measurement in psychology, and most of these scales are comprised of sentences, questions, or vignettes. Advances in NLP provide accessible approaches to represent such contextualized pieces of text. Here, we combine the construct validity of top-down theory (and associated psychometric scales) and the computational power of contextual language models in NLP. The result is advancing theory-driven text analysis, which can give us the desired interpretability of top-down methods such as word lists, but also the predictive power of modern computational linguistic tools (see Yarkoni & Westfall, 2017). In two studies, we demonstrate that our proposed method, CCR, substantially outperforms both word-counting methods and prior methods that rely on word embeddings (Garten et al., 2018). Importantly, CCR is not dependent upon a particular language model; rather, as large language models become more sophisticated and powerful, CCR will be able to flexibly use them to outperform predecessor language models. Our methodology has broad applications in social and personality psychology including, but not limited to, social-media “big data” analytics which is gaining momentum in the modern science of mind and behavior.

References

Abdel-Khalek, A. M. (2007). Assessment of intrinsic religiosity with a single-item measure in a sample of Arab Muslims. *Journal of Muslim Mental Health*, *2*(2), 211–215.

Afhami, R., Mohammadi-Zarghan, S., & Atari, M. (2017). Self-rating of religiosity (SRR) in Iran: Validity, reliability, and associations with the big five. *Mental Health, Religion & Culture*, *20*(9), 879–887.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, *51*(2), 1–38.

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and social psychology review*, *11*(2), 150–166.

Atari, M., & Dehghani, M. (2022). Language analysis in moral psychology. In M. Dehghani & R. L. Boyd (Eds.), *Handbook of language analysis in psychology* (pp. 207–228). Guilford.

Atari, M., Haidt, J., Graham, J., Koleva, S., Stevens, S. T., & Dehghani, M. (2022). Morality beyond the weird: How the nomological network of morality varies across cultures. psyarxiv.com/q6c9r

Atari, M., & Henrich, J. (2023). Historical psychology. *Current Directions in Psychological Science*, *32*, 1–8. https://doi.org/10.1177/09637214221149737

Atari, M., Mehl, M. R., Graham, J., Doris, J. M., Schwarz, N., Davani, A. M., Omrani, A., Kennedy, B., Gonzalez, E., Jafarzadeh, N., Hussain, A., Mirinjian, A., Madden, A., Bhatia, R., Burch, A., Harlan, A., Sbarra, D. A., Raison, C. L., Moseley, S. A., . . . Dehghani, M. (2023). The paucity of morality in everyday talk. *Scientific Reports*, *13*, 5967.

Atari, M., Mostafazadeh Davani, A., & Dehghani, M. (2020). Body maps of moral concerns. *Psychological science*, *31*(2), 160–169.

Atari, M., Reimer, N. K., Graham, J., Hoover, J., Kennedy, B., Davani, A. M., Karimi-Malekabadi, F., Birjandi, S., & Dehghani, M. (2022). Pathogens are linked to human moral systems across time and space. *Current Research in Ecological and Social Psychology*, 100060.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on psychological science*, *2*(4), 396–403.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit.* " O'Reilly Media, Inc."

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in cognitive sciences.*

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

Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, *23*(2), 190–203.

Boyd, R. L., & Pennebaker, J. W. (2015). Did Shakespeare write double falsehood? identifying individuals by creating psychological signatures with text analysis. *Psychological science*, *26*(5), 570–582.

Boyd, R. L., & Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, *40*(1), 21–41.

Boyd, R. L., Wilson, S., Pennebaker, J., Kosinski, M., Stillwell, D., & Mihalcea, R. (2015). Values in words: Using language to evaluate and understand personal values.

Proceedings of the International AAAI Conference on Web and Social Media, *9*(1), 31–40.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318.

Burton, J. W., Cruz, N., & Hahn, U. (2021). Reconsidering evidence of moral contagion in online social networks. *Nature Human Behaviour*, *5*(12), 1629–1635.

Buttrick, N., Moulder, R., & Oishi, S. (2020). Historical change in the moral foundations of political persuasion. *Personality and Social Psychology Bulletin*, *46*(11), 1523–1537.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of personality and social psychology*, *42*(1), 116.

Candia, C., Atari, M., Kteily, N., & Uzzi, B. (2022). Overuse of moral language dampens content engagement on social media. *Under review*, *10*, xx.

Charlesworth, T. E., & Banaji, M. R. (2022). Patterns of implicit and explicit attitudes: Iv. change and stability from 2007 to 2020. *Psychological Science*, 09567976221084257.

Charlesworth, T. E., Caliskan, A., & Banaji, M. R. (2022). Historical representations of social groups across 200 years of word embeddings from google books. *Proceedings of the National Academy of Sciences*, *119*(28), e2121798119.

Charlesworth, T. E., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, *32*(2), 218–240.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319.

Clifford, S., Iyengar, V., Cabeza, R., & Sinnott-Armstrong, W. (2015). Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior research methods*, *47*(4), 1178–1198.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281.

Cutler, A., & Condon, D. M. (2022). Deep lexical hypothesis: Identifying personality structure in natural language. *Journal of Personality and Social Psychology*.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications*. Sage publications.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 10–32.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465.

Fried, E. I. (2015). Problematic assumptions have slowed down depression research: Why symptoms, not syndromes are the way forward. *Frontiers in psychology*, *6*, 309.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior research methods*, *50*(1), 344–361.

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., Duan, L., Almaliach, A., Ang, S., Arnadottir, J., et al. (2011). Differences between tight and loose cultures: A 33-nation study. *science*, *332*(6033), 1100–1104.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, *96*(5), 1029.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of personality and social psychology*, *101*(2), 366.

Greenfield, P. M. (2013). The changing psychology of culture from 1800 through 2000. *Psychological science*, *24*(9), 1722–1731.

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences*, *18*, 43–49.

Holtzman, N. S., Tackman, A. M., Carey, A. L., Brucks, M. S., Küfner, A. C., Deters, F. G., Back, M. D., Donnellan, M. B., Pennebaker, J. W., Sherman, R. A., et al. (2019). Linguistic markers of grandiose narcissism: A LIWC analysis of 15 samples. *Journal of Language and Social Psychology*, *38*(5-6), 773–786.

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., et al. (2020). Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, *11*(8), 1057–1071.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359–366.

Ireland, M. E., & Mehl, M. R. (2014). Natural language use as a marker. *The Oxford handbook of language and social psychology*, 201–237.

Jackson, J. C., Gelfand, M., De, S., & Fox, A. (2019). The loosening of American culture over 200 years is associated with a creativity–order trade-off. *Nature Human Behaviour*, *3*(3), 244–250.

Jackson, J. C., Watts, J., List, J.-M., Puryear, C., Drabble, R., & Lindquist, K. A. (2021). From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science*, 17456916211004899.

Jaidka, K., Giorgi, S., Schwartz, H. A., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2020). Estimating geographic subjective well-being from twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*, *117*(19), 10165–10171.

Jiang, L., Hwang, J. D., Bhagavatula, C., Le Bras, R., Liang, J., Dodge, J., Sakaguchi, K., Forbes, M., Borchardt, J., Gabriel, S., et al. (2021). Can machines learn morality? the delphi experiment. *arXiv e-prints*, arXiv–2110.

Jost, J. T., & Thompson, E. P. (2000). Group-based dominance and opposition to equality as independent predictors of self-esteem, ethnocentrism, and social policy attitudes among African Americans and European Americans. *Journal of Experimental Social Psychology*, *36*(3), 209–232.

Kennedy, B., Ashokkumar, A., Boyd, R. L., & Dehghani, M. (2022). Text analysis for psychology: Methods, principles, and practices. In M. Dehghani & R. L. Boyd (Eds.), *Handbook of language analysis in psychology* (pp. 3–64). Guilford.

Kennedy, B., Atari, M., Davani, A. M., Hoover, J., Omrani, A., Graham, J., & Dehghani, M. (2021). Moral concerns are differentially observable in language. *Cognition*, *212*, 104696.

Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of personality*, *64*(2), 311–337.

Körner, R., Overbeck, J. R., Körner, E., & Schütz, A. (2022). How the linguistic styles of Donald Trump and Joe Biden reflect different forms of power. *Journal of Language and Social Psychology*, 0261927X221085309.

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, *110*(15), 5802–5805.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Lin, Y., Zhang, Y. C., & Oyserman, D. (2021). Seeing meaning even when none may exist: Collectivism increases belief in empty claims. *Journal of Personality and Social Psychology*.

Lins de Holanda Coelho, G., HP Hanel, P., & J. Wolf, L. (2020). The very efficient assessment of need for cognition: Developing a six-item version. *Assessment*, *27*(8), 1870–1885.

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Manjavacas, E., & Fonteyn, L. (2021). Macberth: Development and evaluation of a historically pre-trained language model for English (1450-1950). *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, 23–36.

Martins, M. d. J. D., & Baumard, N. (2020). The rise of prosociality in fiction preceded democratic revolutions in Early Modern Europe. *Proceedings of the National Academy of Sciences*, *117*(46), 28684–28691.

Matsuo, A., Sasahara, K., Taguchi, Y., & Karasawa, M. (2019). Development and validation of the Japanese moral foundations dictionary. *PloS one*, *14*(3), e0213343.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018). Moralization in social networks and the emergence of violence during protests. *Nature human behaviour*, *2*(6), 389–396.

Muthukrishna, M., Henrich, J., & Slingerland, E. (2021). Psychology as a historical science. *Annual Review of Psychology*, *72*, 717–749.

Nicolas, G., Bai, X., & Fiske, S. T. (2021). Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, *51*(1), 178–196.

Oyserman, D. (1993). The lens of personhood: Viewing the self and others in a multicultural society. *Journal of personality and social psychology*, *65*(5), 993.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of personality and social psychology*, *108*(6), 934.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of personality and social psychology*, *46*(3), 598.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, *77*(6), 1296.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Robinson, M. A. (2018). Using multi-item psychometric scales for research and practice in human resource management. *Human resource management*, *57*(3), 739–750.

Sackeim, H. A., & Gur, R. C. (1979). Self-deception, other-deception, and self-reported psychopathology. *Journal of consulting and clinical psychology*, *47*(1), 213.

Sagi, E., & Dehghani, M. (2014). Measuring moral rhetoric in text. *Social science computer review*, *32*(2), 132–144.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, *8*(9), e73791.

Schwartz, S. H. (2003). A proposal for measuring value orientations across nations. *Questionnaire package of the European social survey*, *259*(290), 261.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*(2), 93–105.

Simchon, A., Brady, W. J., & Van Bavel, J. J. (2020). Troll and divide: The language of online polarization. *PNAS Nexus*.

Srivastava, S. (2020). Personality structure: Who cares? *European Journal of Personality*, (4), 550–551.

Stone, A. A., Bachrach, C. A., Jobe, J. B., Kurtzman, H. S., & Cain, V. S. (Eds.). (1999). *The science of self-report*. Psychology Press. https://doi.org/10.4324/9781410601261

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual review of clinical psychology*, *5*, 1.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, S.-Y. N., & Inbar, Y. (2021). Moral-language use by us political elites. *Psychological Science*, *32*(1), 14–26.

Webb, T., Holyoak, K. J., & Lu, H. (2022). Emergent analogical reasoning in large language models. *arXiv preprint arXiv:2212.09196*.

White, S. H. (1992). G. stanley hall: From philosophy to developmental psychology. *Developmental Psychology*, *28*(1), 25–34.

Wu, S., & Dredze, M. (2020). Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.