Received: 12 March 2020 | Accepted: 29 September 2020

DOI: 10.1002/ejsp.2724

RESEARCH ARTICLE

EASP WILEY

Comprehensive stereotype content dictionaries using a semi-automated method

Gandalf Nicolas | Xuechunzi Bai | Susan T. Fiske

Department of Psychology, Princeton University, Princeton, NJ, USA

Correspondence
Gandalf Nicolas, Department of Psychology, Rutgers University – New Brunswick, Tillet Hall 607, Piscataway, NJ 08854, USA.
Email: gandalf.nicolas@rutgers.edu

Abstract

Advances in natural language processing provide accessible approaches to analyze psychological open-ended data. However, comprehensive instruments for text analysis of stereotype content are missing. We developed stereotype content dictionaries using a semi-automated method based on WordNet and word embeddings. These stereotype content dictionaries covered over 80% of open-ended stereotypes about salient American social groups, compared to 20% coverage from words extracted directly from the stereotype content literature. The dictionaries showed high levels of internal consistency and validity, predicting stereotype scale ratings and human judgments of online text. We developed the R package *Semi-Automated Dictionary Creation for Analyzing Text* (SADCAT; https://github.com/gandalfnicolas/SADCAT) for access to the stereotype content dictionaries and the creation of novel dictionaries for constructs of interest. Potential applications of the dictionaries range from advancing person perception theories through laboratory studies and analysis of online data to identifying social biases in artificial intelligence, social media, and other ubiquitous text sources.

KEYWORDS

dictionaries, stereotype content, text analysis, word embeddings, WordNet

Text data are everywhere. Researchers may obtain text data from sources such as the internet, literary collections, archival entries, and experimental psychology's open-ended responses. Compared to traditional response scales in psychological research, embracing text data allows more unobtrusive and unconstrained approaches to measurement. For example, social media data provide information about participants' cognitions, free from demand characteristics associated with some laboratory studies (see Meshi et al., 2015).

Using open-ended (vs. forced-choice) responses in controlled settings also enables more ecologically valid and data-driven study of psychological processes and content. These benefits appear in studying emotion (Gendron et al., 2015) and racial categorization (Nicolas et al., 2018), challenging previously held findings by employing free-response measures that circumvent researcher constraints on participants' responses. For example, despite several studies showing that Americans categorize Black-White mixed-race faces as Black when only allowed to make Black versus White categorizations (see Nicolas & Skinner, 2017), in a free response task participants most frequently indicated perceiving these targets to be Hispanic or Middle-Eastern (Nicolas et al., 2018). These kinds of online and open-ended text data, however, often need some form of dimensionality reduction and numerical representation for interpretation (e.g., due to the large number of words that may refer to the same overarching construct of interest), making text analysis methods necessary.

Language and text analysis have a long history in psychology (e.g., Dewey, 1910; Miller, 1951; see Boyd, 2017 for a review) and affiliated fields ranging from Sociology and Political Science (see Lucas et al., 2015) to Computer Science (see Nerbonne, 2003). In social and personality psychology in particular, numerous studies have made use of text analysis to obtain novel insights into human traits and behaviors. For example, studies into status differences in language use have shown that higher (vs. lower) status individuals tend to use *we* more often than *I* as pronouns (e.g., Kacewicz et al., 2014).

Other studies have applied text analysis to interpersonal relations, finding for example that members of longer-lasting relationships tend to match their linguistic style and use more positive emotional words (Slatcher & Pennebaker, 2006). Tapping into the vast amounts of online data, text analysis in the field even shows promise of predicting health-related behaviors (see Chung & Pennebaker, 2019) and bringing in more explanation into descriptive frameworks of personality (see Boyd & Pennebaker, 2017).

Despite the advantages, creating and validating text analysis instruments such as dictionaries differs considerably from developing traditional scales, and currently not many appropriately reviewed guidelines exist. As a result, many areas have yet to fully incorporate text analysis methods into their repertoire. An example is stereotyping, which despite being one of the largest research areas within social psychology, suffers from a dearth of specialized text analysis methods and literature that may support new avenues of research (reviewed below).

1 | CURRENT APPROACHES TO TEXT ANALYSIS IN PSYCHOLOGY

Recently, advances in natural language processing in machine learning allow easier extraction of information about psychological processes and content. The most common method to analyze text data in psychology has traditionally been human coding. In this approach, each text is evaluated by a group of human judges in terms of how much it reflects a construct of interest. Measures of agreement between human judges often document reliability. Evidently, however, this approach is time-consuming and resource-demanding, and these limitations rapidly worsen the more data that need to be coded (Iliev et al., 2015). Furthermore, this approach for text analysis lacks standardization—that is, judges coding may vary across studies or laboratories.

An increasingly popular alternative to per-study human coding of text is offered by dictionaries (see Iliev et al., 2015). Dictionaries list words that are indicators of the construct of interest. Once created, dictionaries are a standardized approach for coding text data, across studies, without additional human judge intervention. For this reason, they are also less resource-intensive and time-consuming for users. Dictionaries are also easy to use in analysis (vs. some more advanced natural language processing methods). The analysis process most often consists of counting the number of words in a text that are included in the dictionary. The larger the number of words from the dictionary that are present in the text, the higher the score for the construct of interest measured by the instrument. To illustrate, if evaluating the positivity of a particular text (e.g., a self-description, or a diary entry), a researcher would count the number of words that fall into a positive valence dictionary (e.g., "good," "nice," "amazing") as a measure of the constructs.

The most widely used set of dictionaries in psychology and akin areas is the Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2015). LIWC has been the benchmark for studying text related to content as varied as emotion, social relationships, thinking styles, among others (see Tausczik & Pennebaker, 2010). The original creation and wide usage of the LIWC dictionaries highlights both that available dictionaries are cheaper and easy to implement, and that the cost of creating new dictionaries in the first place may be prohibitively expensive and time-consuming for many researchers. For example, many of the LIWC dictionaries used up to 8 judges across several stages in order to manually expand the word lists (Pennebaker et al., 2015).

Given that the constructs measured by existing dictionaries are inevitably limited compared to the diversity of constructs studied by psychologists, more accessible methods to facilitate dictionary creation are useful. For example, even topic areas as central to social psychology as stereotype content (see Fiske et al., 2010) lack comprehensive specialized instruments for their measurement in text, an issue we sought to address in the current paper.

2 | STEREOTYPE CONTENT

Stereotypes are beliefs about social groups and are encoded and shaped through language (Maass, 1999). In fact, a myriad of linguistic factors matter in the perception of social groups: from the natural language itself (e.g., the presence of grammatical gender in a language affecting gender representations; see Sato et al., 2013), to the concreteness of the language used (e.g., ingroups being described more abstractly than outgroups when performing positive actions; Maass et al., 1989; Semin & Fiedler, 1992), to the part of speech used (e.g., group cues presented through nouns rather than adjectives lead to stronger stereotype-congruent inferences; Carnaghi et al., 2008). The study of stereotypes through explicit person descriptions, in particular, has one of the longest traditions within social psychology (Bergsieker et al., 2012, Study 4; Katz & Braly, 1933). However, to date, no comprehensive instruments for the analysis of text data have been developed in the area. This article provides such an instrument for measuring several relevant dimensions of content.

The stereotype content model (SCM; Fiske et al., 2002), a well-known current framework, proposes that people primarily use two dimensions to think about individuals and groups: warmth (i.e., is this target a friend or a foe?) and competence (i.e., can this target act on their intentions?). A large body of research has corroborated that evaluations along these dimensions occur cross-culturally (Fiske, 2018). The combination of the two core dimensions also predicts intergroup emotions and behavioral tendencies (Cuddy et al., 2007).

More recent models of stereotype content have either defined different facets of warmth and competence, or proposed novel, distinct dimensions of stereotype content. For example, Abele and colleagues (2016) suggest subdividing Warmth (also called Communion) into friendliness/sociability and morality facets and Competence (also called agency) into ability and assertiveness (see also Ellemers, 2017; Goodwin, 2015). The recent Agency-Beliefs-Communion model (ABC; Koch et al., 2016) introduces beliefs (i.e.,

religious-secular beliefs and political orientation) and status. Thus, stereotype content dimensions are still contested, and open-ended data could shed some light on this issue.

3 | TEXT ANALYSIS IN STEREOTYPE CONTENT

The stereotype content literature has so far largely relied on traditional metrics of measurement, in particular Likert-type scales measuring how much a social group allegedly possesses a particular dimension of content. A couple of studies (Decter-Frain & Frimer, 2016; Dupree & Fiske, 2019) have used some LIWC dictionaries to measure warmth (e.g., the *family* and *friend* dictionaries) and competence (e.g., the *work* and *achievement* dictionaries). However, because these dictionaries were not designed to measure those constructs, they may cover both a small subset of appropriate words and correlated constructs rather than the target concepts. For example, the LIWC affiliation dictionary includes words such as *friend* or *friendly*, but not low-directional antonyms (e.g., *enemy* or *unfriendly*). In the absence of a specialized indicator or separate dictionary for the antonyms of these dimensional constructs, text data that include responses along the whole dimension (such as stereotypes) will suffer from lack of coverage or loss of information, depending on the application. Finally, a recent study (Pietraszkiewicz et al., 2018) developed dictionaries of communion (similar to warmth) and agency (similar to competence) using the LIWC development approach, but these included only a subset of possible words (e.g., only high directional), did not provide explicit indicators for the different facets of these dimensions, and did not cover other stereotype dimensions.

For an area such as stereotype content, where responses go beyond a small set of categories, to a large number of possible nouns and adjectives, developing a more comprehensive instrument becomes even more vital to faithfully characterizing text content. Potentially, this instrument could expand current theoretically derived models of social cognition by exploring open-ended responses in controlled experiments, in addition to examining stereotype content in multiple untapped sources of text data online. For example, Fiske et al. (in press) argue that stereotypes obtained through open-ended and text measures may differ from traditional scale-based stereotypes, providing information into which stereotypes are more central to social groups' representation and improving predictive models of discrimination arising from stereotyping. For example, while traditional scale ratings of Warmth and Competence would place Doctors and Nurses as similarly high on both dimensions, spontaneous text responses (e.g., coded through dictionaries) suggest that Warmth is more representative of the stereotype content of Nurses while Competence is more representative of the stereotypes of Doctors. Furthermore, this type of information derived from text responses significantly improves predictions of attitudes toward social groups (see Nicolas et al., 2020). Others argue that stereotypical explicit person descriptions extracted from large online corpora (e.g., social media) may sometimes function more like implicit than explicit stereotypes measured in traditional laboratory scales (Kurdi et al., 2019). These kinds of theoretical advances are greatly facilitated, and sometimes only possible, through the use of automated methods dependent on the existence of valid stereotype content dictionaries. In fact, the simple exploration of the structure of dictionaries in this article may provide some insights into the structure of stereotypes in natural language, as we briefly explore.

In this article, we introduce novel stereotype content dictionaries that fill a void in the study of stereotyping in text. We develop these dictionaries using an approach (incorporating some natural language processing methods in novel ways) that is described in the text and made available through an R package for other researchers to use when developing dictionaries for other constructs of their interest. Finally, we use traditional and emergent techniques to evaluate the coverage, reliability, and validity of the dictionaries. This new approach provides a complementary way to automatize many processes, in order to facilitate new dictionaries that are also less coder-reliant, may handle more words, and address distinctive topics, among other benefits. We make available helper functions used to create dictionaries using this approach in the R package *Semi-Automated Dictionary Creation for Analyzing Text (SADCAT)*, available at https://github.com/gandalfnicolas/SADCAT. The package also contains functions to code text into the stereotype content dictionaries developed here. All data and code for the analyses presented here are also available at https://osf.io/yx45f/?.

4 | COVERAGE, RELIABILITY, AND VALIDITY

Dictionary creation aimed to achieve three indicators of quality: coverage, internal reliability, and convergent validity.

4.1 | Coverage

A traditional psychological scale can measure a construct with a few items sampled from a larger pool of intercorrelated items without wasting any data. However, with text measures, where participants choose the items (i.e., words) they wish to convey about the construct, a larger pool of items is needed to code the participants' responses. Coverage refers to the proportion of possible participant responses that is covered by the dictionary (i.e., the pool of items). Coverage will be domain-dependent. Thus, our dictionaries aim to explain a majority of participants' responses when prompted to provide stereotype content of social groups (i.e., stereotypes). Here, we use WordNet (Miller, 1995), a lexical database with semantic relations between words, in order to automatically expand an initial set of words into their synonyms, antonyms, etc., to increase the coverage of open-ended stereotypes provided by a sample of American respondents.

4.2 | Internal reliability

We refer to internal reliability as the consistency and intercorrelations of the pool of items that make up the dictionaries. In other words, reliability measures whether words within a dictionary bear higher semantic similarity than words in different dictionaries. To measure semantic similarity, we adapt recent methods in natural language processing (Mikolov et al., 2013; Pennington et al., 2014) to generate numeric vector representations of text data. Obtaining pairwise similarities from these vectors enables calculations of traditional metrics of internal reliability, such as the average inter-item "correlation" (in this case average cosine similarity) or Cronbach's alpha.

4.3 | Validity

Validity is relatively straightforward as it is most similar to scales. Using the dictionaries allows us to code the construct of interest from participants' responses, which may then correlate with other constructs expected to be theoretically related (i.e., convergent validity) or unrelated (i.e., divergent validity). Here we test validity against multiple data sources and compare our comprehensive dictionaries with existing dictionaries used to measure stereotype content in text.

We note that the current dictionaries are validated for the domain of explicit person descriptions. This is one of the most relevant and widely studied topics in social psychology, ranging from studies on social group stereotypes (see Fiske et al., 2010) to face impressions (see Todorov, 2017). Explicit and blatant stereotyping is very much alive and widespread (e.g., Kteily & Bruneau, 2017; Roberts & Rizzo, in press), and these instruments aim to measure their use in experimental and online settings. However, based on factors such as social desirability, text data may instead portray stereotypes implicitly or indirectly, or may provide information on the author rather than targets described in text. Our dictionaries may likely be useful for these applications as well (e.g., based on correlations with dictionaries more explicitly designed to measure such indicators, Pietraszkiewicz et al., 2018), but future validation of these applications is necessary.

5 | DICTIONARY CREATION OVERVIEW

Dictionaries evolved through an iterative process that subsequent sections will explain, and that is summarized in a flowchart in Figure 1. To anticipate: In Study 1 we identified from the literature words covering relevant stereotype and person perception dimensions, forming an initial set of seed words dictionaries. In Study 2 we collected stereotype content text data to test how much the initial seed words accounted for participants' responses (i.e., coverage). In Study 3, we used WordNet to expand the seed words to a larger dictionary. We iterated the process of testing coverage and adding words until we reached a good proportion of dictionary coverage. After completing the dictionaries, in Study 4 we tested dictionary reliability using the similarity metrics discussed above. Finally, we tested the validity of the dictionaries in four ways. First, we explored the convergent and discriminant validity of our dictionaries in comparison to existing dictionaries that measure related constructs (Study 5). Second, we tested validity in relation to scale ratings, that is, whether experimentally requested responses coded with our dictionaries correlated with stereotypes measured by

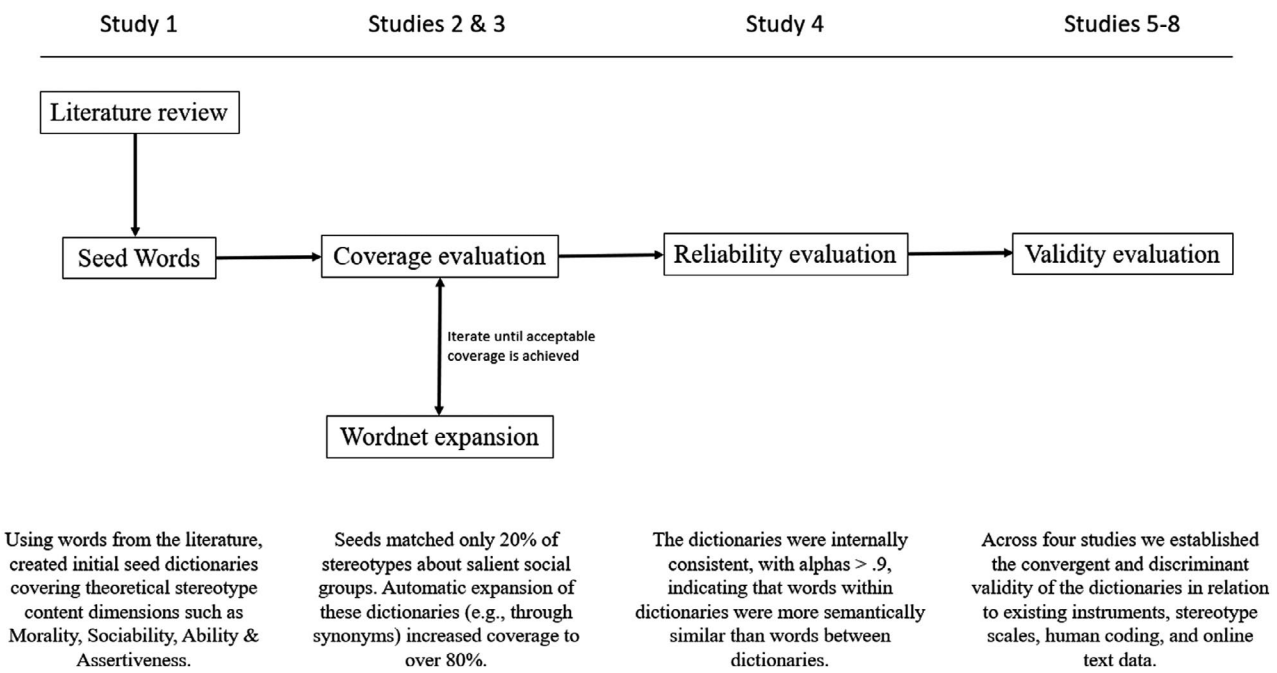

FIGURE 1 Flowchart and summary of procedures and results

TABLE 1 Example words for each seed dictionary

Dimension	Term	Direction
Sociability	Sociable	High
Sociability	Unsociable	Low
Sociability	Friendly	High
Sociability	Unfriendly	Low
Sociability	Warm	High
Sociability	Cold	Low
Sociability	Liked	High
Sociability	Disliked	Low
Sociability	Outgoing	High
Sociability	Shy	Low
Morality	Moral	High
Morality	Immoral	Low
Morality	Trustworthy	High
Morality	Untrustworthy	Low
Morality	Sincere	High
Morality	Insincere	Low
Morality	Fair	High
Morality	Unfair	Low
Morality	Tolerant	High
Morality	Intolerant	Low
Ability	Competent	High
Ability	Incompetent	Low
Ability	Competitive	High
Ability	Uncompetitive	Low
Ability	Intelligent	High
Ability	Unintelligent	Low
Ability	Able	High
Ability	Unable	Low
Ability	Educated	High
Ability	Uneducated	Low
Assertiveness	Confident	High
Assertiveness	Diffident	Low
Assertiveness	Assertive	High
Assertiveness	Unassertive	Low
Assertiveness	Independent	High
Assertiveness	Dependent	Low
Assertiveness	Active	High
Assertiveness	Inactive	Low
Assertiveness	Determined	High
Assertiveness	Doubtful	Low
Status	Wealthy	High
Status	Poor	Low
Status	Powerful	High
Status	Powerless	Low
Status	Superior	High

(Continues)

TABLE 1 (Continued)

Dimension	Term	Direction
Status	Inferior	Low
Status	Influential	High
Status	Uninfluential	Low
Status	Successful	High
Status	Unsuccessful	Low
Politics	Traditional	High
Politics	Modern	Low
Politics	Conventional	High
Politics	Unconventional	Low
Politics	Conservative	High
Politics	Liberal	Low
Politics	Republican	High
Politics	Democrat	Low
Politics	Narrow-minded	High
Politics	Open-minded	Low
Religion	Religious	High
Religion	Irreligious	Low
Religion	Christian	High
Religion	Muslim	High
Religion	Jewish	High
Religion	Atheist	Low
Religion	Secular	Low
Religion	Believer	High
Religion	Nonbeliever	Low
Religion	Skeptic	Low

scales (Study 6). Then, we tested validity in relation to human ratings, that is, whether human coders identified the semantic meaning of each dictionary from a small subset of its items (Study 7). Finally, we used the dictionaries to code for real-world data and correlate these with human coding along the dimensions (Study 8).

In these studies, we report all measures, manipulations and exclusions. Depending on the within-subject variance, power analyses for all studies reveal over 80% power to detect small effects of r or f between 0.1 and 0.2 in our main tests. Sample size was determined before any data analysis. All validity studies with human subjects were approved by the University ethics committee, and adhered to the ethical guidelines specified in the APA Code of Conduct and the US Federal Policy for the Protection of Human Subjects (including informed consent, right to withdraw, and debriefing).

6 | STUDY 1: CREATING SEED DICTIONARIES

In Study 1, we identify stereotype content dimensions that have been previously formalized in stereotype content models and create initial, theory-driven seed dictionaries.

6.1 | Methods

We reviewed the literature (Abele et al., 2008, 2016; Fiske et al., 2002; Koch et al., 2016; Oosterhof & Todorov, 2008; Wojciszke et al., 2011) for lists of words used to measure friendliness/sociability, morality/trustworthiness, ability, assertiveness/dominance, status, political beliefs, and religious beliefs in relation to social groups. For every word, if not already included, we also obtained its antonym.

6.2 | Results

The final seed dictionaries consisted of 341 distinct words, with their corresponding theoretical direction (i.e., high or low on their corresponding dimension, based on how they were labeled in the reviewed literature). See Table 1 for example words; for a full list see online repository. Because words can have multiple senses (e.g., *warm* can refer to psychological or physical warmth) the researchers independently went through the list of seed words and decided on the most appropriate sense(s), based on their part of speech, definition, and example sentences, which resulted in a list

of 455 senses. The final senses were those which two of the researchers agreed on, 90% of the total senses selected by either of the two researchers.

Dictionaries were mostly balanced in terms of high and low senses, but there were some slight imbalances such as more low (vs. high) Morality words and more high (vs. low) Ability words. Note that by high and low we do not mean valence: It is simply an indicator of which end of the antonymy dimension the word refers to; whether one or the other antonym is coded as high versus low is arbitrary. For example, we coded beliefs as ranging from progressive to traditional, and thus high direction in this dictionary means that the word is more about traditional beliefs than progressive beliefs. For more information about the seed dictionaries please refer to the Supplement.

6.3 | Summary

In an initial theory-driven and human-dependent step, we collected from the literature small dictionaries containing seed words for the constructs of stereotype content. These seed word dictionaries would be expanded in subsequent steps to obtain the final instruments.

7 | STUDY 2: SEED DICTIONARIES COVERAGE

In Study 2 we perform an initial test of coverage on development data. That is, we explore how many of participants' open-ended stereotypes about salient U.S. social groups are accounted for by our seed dictionaries.

7.1 | Methods

Development data allowed for initial tests of coverage and validity of the dictionaries. The development data consisted of a survey (N = 201, M_{age} = 37.8, 55% female; 85% White, 6% Black, 3% Hispanic, 3% Asian) asking for participants' spontaneous thoughts about characteristics that different social groups would have. We used a total of 20 social groups (e.g., "Asian", "Elderly", "Wealthy"), sampled from the literature, and showed five to each participant, in random order. Participants provided 10 open-ended single-word responses for each target. Next, participants saw the same social groups again and rated them on warmth (items: friendly, sincere) and competence (items: efficient, competent) using a scale ranging from 1 (*not at all*) to 5 (*extremely*), as well as a measure of familiarity with the social group. Finally, participants completed some demographic questions.

The open-ended responses were preprocessed (e.g., lower cased, deleted grammatical signs; see Supplement).

7.2 | Results

As expected, the words used in the existing literature to describe content dimensions were not a good measure of the diversity of open-ended responses, accounting for only 20.2% of our development data (6.2% of distinct responses). Mapping the content of spontaneous stereotypes requires accounting for most of the responses. However, open-ended responses allow for any number of synonymous terms that have not been exhaustively listed in previous studies. For instance, even though we were able to find in the literature words such as *thief* referring to morality, other synonyms such as *robber* were absent. For this reason, in the next study we expand the dictionaries using WordNet to improve coverage.

7.3 | Summary

In this study we tested how many spontaneous stereotypes provided by a sample of American participants in response to a salient sample of social groups were covered by our seed dictionaries. This coverage was very low, meaning that deploying the seed dictionaries to analyze laboratory or online text data would result in large amounts of missing data and undercounting of construct-relevant words. In the next study we address this limitation.

8 | STUDY 3: EXPANSION AND FINAL DICTIONARIES

Given the low coverage of the seed dictionaries, in Study 3, we use WordNet (Miller, 1995) to automatically expand the seed dictionaries and improve coverage.

8.1 | Methods

Although one could manually gather many words using suggestions from field experts, that labor- and time-consuming method would be limiting. WordNet offers one automated way to obtain a large pool of items by adding words that are semantically associated with a smaller pool of seed words obtained from the literature. WordNet (Miller, 1995) is a large lexical database for the English language. The database contains metadata about English words, including part-of-speech (i.e., noun, adjective, verb, and adjective), glosses (i.e., short definitions), and usage examples in sentences. Most importantly, WordNet distinguishes words' different senses (e.g., warmth may refer to both psychological warmth and temperature), and these senses then associate with other words/senses through several relations such as synonyms and antonyms. Previous research in other fields has used WordNet to expand dictionaries (e.g., Maks et al., 2014). Here, we apply the procedure to the creation of Stereotype Content

TABLE 2 Dictionary characteristics

Dictionary	Words	High	Low	Pos	Neg	Preprocessed	High	Low	Pos	Neg
Sociability	1,210	505	430	0.21	0.24	1,148	479	421	0.21	0.24
Morality	2,523	477	1,865	0.14	0.19	2,404	458	1,791	0.14	0.19
Ability	999	611	303	0.2	0.14	950	590	298	0.2	0.14
Assertiveness	774	453	269	0.16	0.16	731	423	255	0.17	0.17
Health	1,477	39	1,432	0.07	0.22	1,427	35	1,384	0.07	0.22
Status	595	291	193	0.17	0.14	560	279	183	0.17	0.14
Work	2,051	NA	NA	0.03	0.02	1,957	NA	NA	0.02	0.02
Politics	400	87	109	0.08	0.09	391	86	107	0.08	0.09
Religion	818	784	30	0.06	0.05	804	771	30	0.06	0.05
Beliefs - other	119	NA	NA	0.09	0.07	117	NA	NA	0.09	0.07
Inhabitant	664	NA	NA	0	0.01	657	NA	NA	0	0.01
Country	312	NA	NA	0	0.01	306	NA	NA	0	0.01
Feeling	1,164	NA	NA	0.2	0.3	1,088	NA	NA	0.2	0.31
Relative	215	NA	NA	0.04	0.02	214	NA	NA	0.04	0.02
Clothing	602	NA	NA	0.01	0.02	567	NA	NA	0.01	0.02
Ordinariness	147	52	88	0.17	0.23	146	52	88	0.17	0.23
Body part	390	NA	NA	0.03	0.03	353	NA	NA	0.02	0.03
Body properties	349	NA	NA	0.12	0.13	329	NA	NA	0.12	0.13
Skin	59	NA	NA	0.1	0.16	55	NA	NA	0.09	0.17
Body covering	216	NA	NA	0.01	0.04	208	NA	NA	0.01	0.04
Beauty	223	168	47	0.3	0.13	208	155	46	0.31	0.13
Insults	40	NA	NA	0.04	0.34	39	NA	NA	0.04	0.35
STEM	781	NA	NA	0.02	0.01	726	NA	NA	0.02	0.01
Humanities	83	NA	NA	0.09	0.02	80	NA	NA	0.08	0.02
Art	404	NA	NA	0.03	0.02	371	NA	NA	0.03	0.02
Social groups	31	NA	NA	0.06	0.06	31	NA	NA	0.06	0.06
Lacks knowledge	7	NA	NA	0.06	0.05	7	NA	NA	0.06	0.05
Fortune	28	NA	NA	0.25	0.19	27	NA	NA	0.24	0.2

Note: Words are the original words obtained, including different forms of a word (e.g., plural and singular) while preprocessed words collapse across these by lemmatizing, deleting symbols, among others (see development data section for preprocessing procedures). High and Low refers to the number of words for each direction of the dictionary, when available. Valence (Pos: Positive, Neg: Negative) was obtained from SentiWordNet.

Dictionaries, as well as formalizing the procedures in an R package (https://github.com/gandalfnicolas/SADCAT) made available for researchers to create other dictionaries of psychological constructs using this semi-automated approach. An in-depth explanation of the expansion procedures is offered in the Supplement.

After the first round of dictionary words expansion, we explored unaccounted-for words to identify some potential additional topics and used WordNet to expand on these topics. A few specific unaccounted-for responses were added manually to a catch-all dictionary with words that denoted lack of knowledge (e.g., "I don't know" or "?"). We note that some items in the dictionaries contain multiple words, as WordNet includes some multiple-word entries (e.g., for disease names). Unlike traditional stereotype content models (e.g., Fiske et al., 2002), which focus on two or three dimensions believed to be primary, the goal here was to create dictionaries with high coverage for stereotype content, including dimensions that may be used less often than the theoretical dimensions, resulting in this data-driven step. For a follow-up on this topic, see Nicolas et al. (under review). Additional versions of the dictionaries (e.g., shorter versions) are discussed in the Supplement.

8.2 | Results

The final dictionaries contained 14,449 words across 28 dictionaries. Final dictionaries varied in length from seven (lack of knowledge) to 2,402 (morality) preprocessed words (see Table 2 for descriptives, and online repository for full dictionaries). Differences in length may to a small degree reflect biases in WordNet or seed list, but most likely reflect differences in the semantic generality of the dimensions (see Fellbaum, 1998). For example, morality encompasses a wider set of related constructs in the WordNet network than concepts

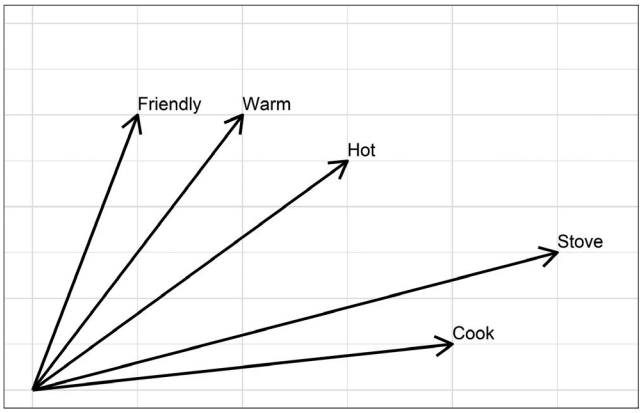

FIGURE 2 Hypothetical two-dimensional word space, with vectors representing different words. Cosine similarity is measured as the angle between vectors, and as shown, words used in similar contexts such as *friendly* and *warm* are more similar to each other than to words used in other contexts (e.g., *stove* and *cook*)

related to uncertainty. Some of these differences in semantic generality may be further explored for the generation of hypotheses on the linguistic nature of different stereotype contents.

For each word, we also obtained its SentiWordNet (Baccianella et al., 2010) valence. This metric indicates how positive or negative each word is, based on human coding. This differs from direction, which is specific to each dictionary (e.g., aggressive is low valence, i.e., it is evaluated as a negative trait, but high direction for the assertiveness dictionary, i.e., it indicates an assertive trait). For words with multiple senses, the direction and valence were averaged across senses.

The final dictionaries accounted for 77% of the development data responses, indicating a significant improvement in coverage from the WordNet expansion. Given the potential for overfitting to these specific data, in a confirmatory study (described later in the first validation study), the dictionaries accounted for 84% of the responses. Thus, the dictionaries can account for the vast majority of stereotype-relevant responses in traditional groups explored in stereotyping research.

8.3 | Summary

In this study we created the final dictionaries by expanding them into semantically related words in an automated fashion. The final dictionaries covered multiple dimensions of stereotyping, from Morality and Ability, to physical features. These expanded dictionaries accounted for over three-quarters of the stereotypes of salient American social groups, suggesting they would be useful in minimizing missing data in the analysis of stereotypes from text.

9 | STUDY 4: INTERCORRELATIONS AND INTERNAL RELIABILITY

In this study we explore how some of the dictionaries intercorrelate and how internally reliable is each dictionary. Internal reliability in the context of dictionaries refers to the semantic internal consistency of the dictionaries. That is, are words within a dictionary semantically similar to each other?

9.1 | Methods

To numerically represent text and calculate semantic similarities we used word embeddings, which are numerical vector representations of words derived from models trained on large corpora of natural language text (see Bengio et al., 2003; see Figure 2). The specific word embeddings used here are Word2Vec's model pretrained on Google News (Mikolov et al., 2013) and Glove's model pretrained on the Common Crawl (Pennington et al., 2014; presented in Supplement). These vector representations encode each word's position in a multidimensional semantic space, derived from word co-occurrences in large corpora of text. The background model computations are beyond our scope but a brief explanation of the intuition behind training these models is provided in the Supplement.

If we then take two of these vectors representing two words in our dictionaries, we can measure their semantic similarity by calculating their cosine similarity (e.g., Kenter & De Rijke, 2015). Cosine similarity can be interpreted similarly to a Pearson correlation, with larger numbers indicating more similarity. In theory, cosine similarity can range from −1 to 1, but for word embeddings the values tend to be high for words in a common domain, such as words used to describe people, as in our dictionaries. Using this metric, we may obtain high similarity for words such as *river* and *ocean*, as they are semantically related, and lower scores for words such as *river* and *stove*. Others have recently noted the novel technique described here for internal reliability as a measure of semantic coherence (Garten et al., 2018), but not directly for the purpose of evaluating dictionary quality. In order to obtain correlation measures for the different dictionaries, we computed their numeric representation using the previously described word embeddings. Specifically, we obtained the vectors for each dictionary's word and averaged them, which resulted in a vector representation for the dictionary, allowing us to calculate cosine similarities between them.

Alternatively, one can use traditional Cronbach's Alphas to assess reliability. We used the previously described measures of similarities as the inter-item correlations and applied the formula for alpha (as described in Chakrabartty, 2018).

9.2 | Results

We present a subset of dictionary intercorrelations (using cosine similarities derived from Word2vec's pretrained model) in Table 3. Additional similarities are included in the Supplement. These results show general patterns such as higher similarity between Beliefs and Warmth (vs. Competence) that replicate previous laboratory findings (e.g., association between Progressive-Conservative beliefs and Warmth; Koch et al., 2020) in a natural language corpus.

TABLE 3 Correlations between our dictionaries

	Warmth	Competence	Beliefs	Health	Status	Work	Emotion	Family	Deviance	Social Groups	Geography	Appearance
Warmth	1.000											
Competence	.896	1.000										
Beliefs	.767	.716	1.000									
Health	.547	.552	.493	1.000								
Status	.813	.826	.699	.470	1.000							
Work	.666	.603	.608	.467	.596	1.000						
Emotion	.865	.838	.651	.514	.745	.524	1.000					
Family	.573	.474	.584	.463	.519	.648	.482	1.000				
Deviance	.735	.760	.605	.449	.681	.456	.750	.401	1.000			
Social Groups	.577	.460	.545	.388	.442	.528	.454	.564	.463	1.000		
Geography	.514	.469	.640	.393	.523	.552	.414	.483	.402	.478	1.000	
Appearance	.690	.651	.606	.584	.593	.655	.625	.565	.589	.541	.494	1.000

Note: All values are cosine similarities, with larger numbers indicating higher similarities. Values can theoretica[illegible]nd high values for words in a similar domain (e.g., person descriptions, as here), thus interpretation should be comparative.

These results also hint at other theoretical insights derived from the semantic relatedness of these stereotype contents in natural language. For example, Deviance-related words were more closely associated with Competence than Warmth, suggesting that, at least in the natural language corpora used here, descriptions about targets' distinctiveness tend to be more semantically related to the Abilities and Agency domains than the Morality and Sociability domains. We present this as a hypothesis-generating secondary finding derived from the development of instruments through our automated procedure.

A straightforward inferential test for the internal reliability of dictionaries is to check whether the average pairwise similarity between words from the same dictionary is larger than the average pairwise similarity between words from different dictionaries. Indeed, using the pretrained Word2vec's model similarities such a test reveals that words within a dictionary are more co-similar ($M = 0.177$) than words between dictionaries ($M = 0.097$), $t(29.14) = -8.69$, $p < .001$, $d = 1.67$. This large effect size denotes the semantic consistency of the WordNet network, and therefore our dictionaries.

we could do this on a dimension basis. average pairwise similarity of random words in the examples corpus average pairwise similarity

The results for Cronbach's Alphas also indicated high internal reliabilities (> .9) for most dictionaries, with the exceptions of the "lacks knowledge" dictionary (.37) and "fortune" dictionary (.85), which had very few items found in the word embeddings model. We do note that alpha has some limitations as a measure of internal consistencies of dictionaries which are discussed in the Supplement.

9.3 | Summary

Our dictionaries showed remarkable internal consistency. Words within a dictionary were much more similar in meaning than words from different dictionaries. This pattern was also reflected in very high Cronbach's Alpha scores. This suggests that indeed the dictionary words cluster together as necessary for a construct's indicators. Appropriate internal reliability allows for further explorations of validity in the following studies.

10 | STUDY 5: VALIDITY AS RELATED TO EXISTING INSTRUMENTS

In order to obtain estimates of convergent and discriminant validity in the context of existing dictionaries, we compare our dictionaries to the recently developed Communion (akin to Warmth) and Agency (akin to Competence) dictionaries (Pietraszkiewicz et al., 2018), as well as some of the LIWC dictionaries (Pennebaker et al., 2015). We expected to find that our Warmth and Competence dictionaries correlate with the Communion and Agency dictionaries, while some of our additional dictionaries correlate with relevant LIWC dictionaries (e.g., our Religion dictionary with LIWC's Religion dictionary). We also expected lower correlations with theoretically unrelated dictionaries, such as LIWC's Numbers dictionary (including words such as *dozen* or *nine*).

TABLE 4 Correlations between our dictionaries and corresponding existing measures

	Communion	Agency	Power	Religion	Health	Body	Work	Numbers
Warmth	**.57**	.48	.52	.56	.36	.43	.42	.12
Competence	.49	**.61**	.57	.45	.41	.43	.50	.18
Status	.51	.57	**.66**	.51	.34	.38	.48	.28
Religion	.34	.22	.37	**.79**	.31	.36	.29	.08
Health	.24	.24	.31	.29	**.64**	.51	.28	.16
Appearance	.24	.28	.36	.43	.36	**.75**	.26	.14
Work	.42	.36	.59	.48	.39	.46	**.53**	.19

Note: The first column shows a subset of our dictionaries and the subsequent ones indicate the comparison dictionaries from the Agency/Communion dictionaries and LIWC. In bold, in the diagonal, are the congruent dictionaries we expected to show the highest scores (row-wise). Indeed, this is what the results indicate. *Numbers* serves as an irrelevant comparison, confirmed by the low similarity of this dictionary to all other dictionaries. All values are cosine similarities, with larger numbers indicating higher similarities. Values can theoretically range from −1 to 1.

10.1 | Method

We adapted the procedure employed by Pietraszkiewicz and colleagues for this study, as well as using their Communion/Agency dictionaries.[1]

The Communion/Agency and LIWC dictionaries contain word stems that need to be expanded into sense-appropriate words (e.g., *commun** into *communion*). In order to do this, we adapted Pietraszkiewicz et al.'s (2018) procedure of using the 2,500 most frequent words in a large English corpus (Google Web Trillion Word Corpus; Brants & Franz, 2006) as possible expansions.

In order to obtain correlation measures for the different dictionaries, we computed their numeric representation using the previously described word embeddings. Specifically, we obtained the vectors for each dictionary's word and averaged them, resulting in a vector representation for the dictionary. With this information, we could obtain the cosine similarity between different dictionaries. Most of the dictionaries we developed had no appropriate existing comparison, but those that did include: Warmth and Communion (Pietraszkiewicz et al.); Competence and Agency (Pietraszkiewicz et al.); Status and Power (LIWC); Religion and Religion (LIWC); Health and Health (LIWC); Appearance and Body (LIWC); Work and Work (LIWC); and we included LIWC's "numbers" dictionary for an irrelevant comparison to all dictionaries.

10.2 | Results

As expected, corresponding dictionaries had higher similarities than non-corresponding dictionaries. Results based on word2vec embeddings are shown in Table 4. As shown, our dictionaries showed the highest similarity to their corresponding theoretical constructs measured by similar instruments. For example, Competence is more similar to Agency (cosine similarity = .61) than other dimensions and Warmth is most similar to Communion (.57). As a comparison, our dictionaries are dissimilar to theoretically irrelevant dictionaries, such as LIWC's *numbers* (on average around .1). Additional co-similarities between other dictionaries which also support our conclusions are presented in the Supplement.

10.3 | Summary

To summarize this first validation study, we found evidence that our dictionaries correlate with existing, theoretically relevant, dictionaries, providing evidence for convergent validity and situating our dictionaries in relation to current measures. Similarly, our dictionaries showed much lower similarity to theoretically irrelevant constructs, such as number words.

11 | STUDY 6: VALIDITY AS RELATED TO RATING SCALES

Given that scales are the traditional and most commonly used way of gathering information in psychology, we next tested how our dictionaries related to scale stereotype ratings of social groups. For each social group, in addition to seven open-ended responses, we collected participants' Likert-type ratings on stereotype content dimensions of warmth, competence, and beliefs. We planned to test how well the scale ratings were predicted by our sociability, morality, ability, assertiveness, beliefs, and status dictionaries, all of which have been linked to these dimensions in the literature.

11.1 | Method

Participants (N = 251) were recruited through Amazon Mechanical Turk (M_{age} = 33.3, 52% female; 76% White, 10% Black, 5% Hispanic, 4% Asian).

[1]We identified one additional set of Agency and Communion dictionaries (Hart, Sedikides, Wildschut, Arndt, Routledge, & Vingerhoets, 2011), but these were not fully reported in a research article establishing their reliability and validity, and are not as recent as the Pietraszkiewicz and colleagues set we use for comparison. The Agency/Communion dictionaries showed convergent validity with, for example, related dictionaries (e.g., LIWC's *Affiliation* dictionary for Communion) and social groups' stereotypes in a media source.

TABLE 5 Prediction of scales by dictionaries' direction and direction by prevalence interaction

Outcome	Predictor	Beta	t	*df*	*p*	Marg. R^2
Warmth	Warmth	0.36	11.99	756.99	<.001	.14
Warmth	Morality	0.422	11.5	591.88	<.001	.182
Warmth	Sociability	0.337	9.37	471.24	<.001	.12
Competence	Competence	0.302	10.15	832.41	<.001	.115
Competence	Ability	0.242	6.46	618.71	<.001	.068
Competence	Assertiveness	0.294	8.52	593.62	<.001	.1
Morality	Morality	0.391	11.1	593.9	<.001	.161
Sociability	Sociability	0.329	8.25	477.87	<.001	.116
Ability	Ability	0.274	7.14	611.61	<.001	.091
Assertiveness	Assertiveness	0.265	6.94	591.31	<.001	.081
Beliefs	Beliefs	0.155	2.3	224.3	.022	.027
Beliefs	Politics	0.191	2.62	181.51	.01	.04
Beliefs	Religion	0.285	2.12	43.95	.039	.091
Politics	Politics	0.23	3.05	185.79	.003	.056
Religion	Religion	0.175	1.41	42.12	.167	.038

Note: Outcomes are scales and predictors are the dictionary direction. Models with religion as variable either needed further simplification (e.g., deletion of random intercepts for group), or were not computable, due to the low number of responses related to religion, such as in the case of interaction effects. Marginal R^2 are provided for the models.

In an initial block, participants saw a sample of four social groups, from the same social groups as the development data. The instructions read: "Please indicate how the following people are viewed by society. Please note that we are not interested in your personal beliefs, but in how you think these people are viewed by others." They were also told to use one word per box, two maximum, and then saw the prompt "As viewed by society, what are the characteristics of a person who is…" followed by the social group and seven boxes for responses. These responses were preprocessed in the same way as those from the development data.

In a second block, they saw the same groups, but rated them on scales. The prompt read "Please indicate how the following people would be viewed by society. Please note that we are not interested in your personal beliefs, but in how you think these people are viewed by others." This was followed by "To what extent would most individuals in our society view a person who is (social group) as…" and a 1 (not at all) to 5 (extremely) scale for the items "Friendly/Sociable", "Trustworthy/Moral", "Self-confident/Assertive", "Competent/Skilled", "Wealthy/High-status", "Politically conservative". "Religious". These items corresponded to the facets of sociability, morality, assertiveness, ability, status, politics, and religion. To form indexes, "Friendly/Sociable" and "Trustworthy/Moral" were combined for warmth (alpha = .76), "Self-confident/Assertive" and "Competent/Skilled" were combined for competence (alpha = .86), and "Politically conservative" and "Religious" were combined for beliefs (alpha = .7). After these blocks, participants completed demographic questions.

Analyses were mixed-effects models with participants and social groups as random factors, and observations were each participant's responses to a group (i.e., averaging across the seven responses for the text data). In terms of relevant variables, note that while scales measure only direction (e.g., low to high competence in a 5-point scale), our theory-driven dictionaries measure as separate variables both prevalence and direction. Prevalence refers to the number of words related to the dimension (e.g., out of a participant's seven responses, more competence-related words indicate higher prevalence of competence). Direction refers to the antonymy dimensional end of the word. Words high on a dimension (e.g., *friendly* for Warmth) were coded as 1 for that dimension's direction, and words low on the dimension (e.g., *unfriendly* for Warmth) were coded as −1. If direction was unknown it was coded as 0, and if the response was not in the dictionary, it was coded as missing. Thus, dictionary direction variables ranged from −1 to 1. If the dictionary is valid, dictionary direction should predict scale ratings: the higher the direction score, the higher the scale score. In the main text we present only the direction results, and in the Supplement, we include results including prevalence, as well as a comparison with the previously introduced Communion and Agency dictionaries (Pietraszkiewicz et al., 2018). Both these results support the incremental validity of these dictionaries.

11.2 | Results

We found the expected patterns of results, with all dictionary direction indicators predicting scale ratings. For example, responses coded as high warmth significantly predicted higher scale ratings on warmth dimension (r = .36, p < .001). Responses coded as high competence significantly predicted higher scale ratings

TABLE 6 Estimated values and pairwise comparisons between congruent and incongruent response options

	Sociability	Morality	Assertiveness	Ability	Status	Beliefs	Health	Work	Body	Family	Emotions	Geography
Sociability	**3.69**	2.92**	2.8**	2.7**	2.61**	2.58**	2.56**	2.56**	2.49**	2.73**	3.46^	2.47**
Morality	3.11*	**3.42**	2.84**	2.77**	2.78**	2.72**	2.52**	2.75**	2.47**	2.62**	3.04**	2.48**
Assertiveness	2.94**	2.84**	**3.4**	3.14*	2.66**	2.6**	2.65**	2.87**	2.54**	2.6**	3.27	2.47**
Ability	2.87**	2.73**	3.15**	**3.77**	2.74**	2.57**	2.55**	2.82**	2.57**	2.63**	2.87**	2.47**
Status	2.86**	2.73**	3.1**	3.14**	**3.52**	2.6**	2.63**	2.88**	2.5**	2.7**	2.88**	2.51**
Beliefs	2.71**	2.9**	2.74**	2.68**	2.77**	**3.81**	2.47**	2.71**	2.58**	2.68**	2.82**	2.79**
Health	2.4**	2.54**	2.45**	2.47**	2.42**	2.32**	**4.14**	2.48**	3.18**	2.52**	2.72**	2.46**
Work	2.63**	2.55**	2.69**	3.24**	2.86**	2.54**	2.62**	**3.86**	2.56**	2.54**	2.62**	2.62**
Body	2.69**	2.59**	2.65**	2.67**	2.7**	2.57**	2.7**	2.6**	**3.74**	2.56**	2.67**	2.56**
Family	2.84**	2.71**	2.65**	2.64**	2.76**	2.52**	2.53**	2.62**	2.59**	**3.89**	2.86**	2.55**
Emotions	3.09**	2.78**	3.03**	2.69**	2.65**	2.5**	2.69**	2.52**	2.56**	2.62**	**4.01**	2.43**
Geography	2.6**	2.58**	2.54**	2.49**	2.76**	2.76**	2.44**	2.56**	2.58**	2.71**	2.55**	**4.25**

Note: Each row is a dictionary and each column a response option. Values are coefficient for the response in the specified dictionary. *p*-values refer to the pairwise comparison between that baseline (the congruent response score) and the column response (e.g., the morality response for the sociability dictionary, 2.92, is significantly smaller than the sociability response for the sociability dictionary, 3.69). *p*-values for each outcome control for family-wise multiple comparisons by using the Dunnet method for 11 tests. For sociability dictionary, emotion and sociability are different at $p = .009$ when not adjusting multiple comparisons. For the assertiveness dictionary, emotion and assertiveness are at $p = .116$ when not adjusting for multiple comparisons.

*$p < .01$, **$p < .001$, ^NS only when controlling for multiple testing. Bold (diagonal), congruent scores, no pairwise comparisons with themselves, so no significance testing is provided, all are different from zero.

on competence dimension ($r = .30$, $p < .001$). See Table 5 for all direction results. We note however that the religion direction indicator was not significant for predicting the religion item. This was probably due to the low rate of religion-related open-ended responses, which greatly lowered the useable data for models with this variable. Also as expected, informal observations of cross-dictionary models (e.g., competence direction predicting Warmth scales) showed smaller and/or non-significant results compared to models for congruent dimension dictionaries (see Supplement). Additional secondary and potentially hypothesis-generating observations include the finding that scaled Warmth was better predicted by the Morality (vs. Sociability) spontaneous stereotypes (in line with models arguing for the priority of the Morality facet; e.g., Ellemers, 2017; Goodwin, 2015).

11.3 | Summary

Results from this study support the validity of our main Warmth and Competence dictionaries. We found that these dictionaries applied to open-ended stereotypes of social groups predicted how these social groups were rated using traditional numerical scales. This finding suggests that the dictionaries indeed capture judgments of warmth and competence in the context of social groups.

12 | STUDY 7: VALIDITY AS RELATED TO HUMAN JUDGMENT

A third test of validity used human ratings of thematic identification to study the extent to which the dictionary words reflect human semantic judgments. Specifically, we expected to show that human coders appropriately identify words from a dictionary as belonging to it.

12.1 | Method

Participants ($N = 245$) were recruited through Amazon Mechanical Turk ($M_{age} = 33.3$; 61% male; 78% White, 9% Black, 6% Asian, 2% Hispanic). Participants saw 13 blocks, each of which presented a random sample of six words of a dictionary. Instructions asked participants to identify the common theme of the six words and to rate on a scale from 1 (*Not at all*) to 6 (*Extremely*) how well they fit into a condensed list of our dictionaries, in lay terms: sociability/friendliness, morality/trustworthiness, confidence/autonomy, ability/skill, socioeconomic status, political or religious beliefs, health, work/professions, body properties/parts/appearance, familiarity/family, feelings/emotions, and geography. Participants were told to consider words from both directions (e.g., both morality and immorality) to refer to the same theme and were asked to base their responses on the objective meaning of the words rather than personal opinion.

TABLE 7 ICCs and models for natural language validity of the dictionaries

Dimension	ICC	Df	Direction
Sociability	0.46 [0.36, 0.55]	349	.34*
Morality	0.33 [0.20, 0.44]	243	.26*
Ability	0.51 [0.41, 0.59]	416	.39*
Assertiveness	0.28 [0.14, 0.39]	253	.29*
Status	0.41 [0.30, 0.51]	403	.24*
Beliefs	0.65 [0.58, 0.70]	151	.09

Note: Direction results indicate the Pearson correlation between the coders ratings and the dictionary direction scores.
*$p < .001$.

Validity in this case is indexed by whether human coding of the content of a dictionary's words matches the construct they are intended to measure, and to a higher extent than constructs they are not intended to measure. Analyses consisted of a series of mixed models (participants as random intercepts), one for each dictionary. Thus, for example, the morality model was based on data from the block which showed items from the morality dictionary, and so on for all other dictionaries. Because each block had 12 questions, one for each dictionary label (e.g., morality/trustworthiness, confidence/autonomy), each of these questions (or labels) became a level in a contrast-coded predictor. The response to each question was the outcome variable, allowing us to statistically compare the mean response for each question, for each dictionary. Thus, for example for the morality model, using words from the Morality dictionary, we could statistically compare if coders had higher ratings for the morality/trustworthiness item than for the other 11 items.

12.2 | Results

Analyses largely supported the expected patterns. Given the large number of tests, results are summarized in Table 6. In general, the congruent score for each dimension was significantly higher than the score for all other incongruent dimensions. For instance, human coders rated words randomly sampled from the Ability dictionary to refer more to the concept of Ability ($M = 3.77$) than the concepts of Morality ($M = 2.73$) or geography ($M = 2.47$), $ps < 0.001$. Exceptions were only for the Sociability and Assertiveness dictionaries, where scores on the emotion response option were not significantly lower than scores on the congruent dimensions. Possibly, sociability and assertiveness are simply more highly correlated to emotional words; for example, the Stereotype Content Model posits that stereotypes trigger emotions (Fiske et al., 2002), such as others' positive or negative emotions telling us about their friendliness, or because approach and avoidance emotions (see Elliot et al., 2013) relate to assertiveness. However, future studies could further explore this issue, and understand the overlap between these dictionaries when making inferences. In other words, at least on this metric of coder

identifiability, the dictionaries for sociability and assertiveness should be expected to also reflect emotional content.

12.3 | Summary

To reiterate, Study 7 provided evidence that human coders were able to identify the dimension that the dictionaries meant to measure from small samples of words from the dictionary. This provides evidence for the validity of the dictionaries as they are correctly identified by human coders.

13 | STUDY 8: VALIDITY IN NATURAL LANGUAGE

In previous tests of validity, we have isolated the content words to test their semantic validity in relation to the construct of interest. This isolated use of the words is also useful in an experimental setting where researchers may ask for single-word responses. However, much of the use given to dictionaries is on natural language data in longer formats (e.g., social media posts; Nicolas et al., 2019), where the words are not isolated but rather parts of sentences. Thus, it is important to test the validity of the dictionaries in the context of longer texts, as this is one of the most relevant uses of the instrument. In order to do this, we collected obituaries from the web and analyzed them using our dictionaries. Subsequently, we predicted human coders' ratings of the obituaries along multiple dimensions, in order to evaluate the dictionaries' validity.

13.1 | Method

We collected 500 obituaries from various newspaper websites indexed by obituaries.com. Obituaries made an appropriate sample for several reasons. First, they are largely person descriptions, making them the most likely type of text for the dictionaries' use. Second, they were likely to include multiples of the dimensions measured by the dictionaries (e.g., sociability, beliefs, status, health), allowing for validation of most dictionaries using the same sample. Finally, their format is highly standardized, allowing for removal of irrelevant information for more efficient coding and analysis.

To code the obituaries with the dictionaries, we used a similar strategy to the one used in Study 6. Specifically, we obtained a direction score (ranging from −1 to 1) by matching all the words in each obituary to each dictionary and averaging them. For the human coding, we recruited two coders and asked them to code the dimensions for which we had a directional variable using the following scale, illustrated with the sociability dimension: Based on the text, how unsociable–sociable do you believe the person described is? (1—not at all to 5—a lot; or NA if the text does not provide enough information to rate the target's sociability). This scale was used for sociability, morality (immoral-moral), ability (low-high ability), assertiveness (unassertive-assertive), status (low-high status), and beliefs (progressive/non-religious-conservative/religious). Both coders rated all 500 obituaries on all dimensions, allowing us to calculate their interrater reliability. The Intraclass Correlation Coefficients (for average fixed raters) are also reported in Table 7. The scores of the coders were averaged.

To test for validity, we ran a set of linear models predicting the human numerical direction code from the dictionary direction. Additional results including prevalence scores and comparison with the benchmark Communion and Agency dictionaries are presented in the Supplement and further establish the incremental validity of these and additional secondary dictionaries.

13.2 | Results

All the results are presented in Table 7. We found that, regardless of the model used, the dictionaries' validity was supported. Specifically, we found that human perceptions of whether a dimension was high or low in a text correlated with the dictionaries' coding of this direction. For instance, an obituary including descriptions of an individual such as "member of the national honor society" and "vice president of the … club" received very high scores on Socioeconomic Status from both the dictionary and human raters, while a text describing an individual as working in the food services industry received low scores on this metric from both the dictionary and the human coder. On average, this translated, for example, to a correlation of .24 between the Status dictionary and human coding, $p < .001$. The exception for the expected pattern was the Beliefs' direction indicator not reaching statistical significance (although other indicators were significant, see Supplement), potentially as a result of a lower rate of Beliefs-related words resulting in more missing data in the model.

We note that a limitation of obituary data is that the person descriptions tend to include mostly positive words, which could potentially impact the estimates presented here. Nonetheless, as indicated by the significant direction results, our dictionaries were still able to capture the obituaries' subtle valence variations (valence correlates with direction, see Supplement).

13.3 | Summary

In naturalistic data obtained from the internet, a potentially vast application for these dictionaries, we find that indeed the dictionaries showed validity in predicting human coding of obituaries. Our dictionaries were able to capture subtle variations in person evaluations in text along multiple dimensions of stereotype content.

14 | DISCUSSION

In this article we created novel stereotype content dictionaries that have excellent coverage, reliability, and validity. To do this, we used a

novel approach that is more automated than existing human-coded approaches and based on standardized sources such as WordNet and word embeddings models. Furthermore, we provided guidelines for the evaluation of text analysis instruments, including coverage, reliability, and validity.

14.1 | Summary of current studies

The field of stereotype content is ever-growing but suffers from a deficit of studies exploring open-ended stereotype responses, and a lack of access to online text data due to the limitations of current instruments. The current studies used the following steps to develop dictionaries for the measurement of stereotype content:

1. **Creating Seed Dictionaries** (Study 1): We identified 341 words for the literature-relevant constructs of sociability, morality/trustworthiness, ability, assertiveness/dominance, status, and political and religious beliefs, as well as indicators of their direction (i.e., high or low on the dimensions). In our case, this was a fully theory-driven step, but it was complemented by data-driven dictionaries in the following steps.
2. **Seed Dictionaries Coverage** (Study 2): We collected development data in which participants provided open-ended stereotypes about social groups. We tested how many of their responses were covered by our seed dictionaries. Seed words accounted for only about 20% of participants' stereotypes, an unacceptable level that prompted us to expand the dictionaries.
3. **Expansion and Final Dictionaries Coverage** (Study 3): We used WordNet to expand the seed words into fuller dictionaries. We also identified additional seed words from unaccounted-for responses and expanded those as well. The final version of the instrument has 28 dictionaries and 14,449 words. We also obtained the valence for all these words, in addition to direction for most of the dictionaries. The expansion of words resulted in over 80% coverage. We considered this coverage to be acceptable, and it was a considerable improvement on the existing items in the literature.
4. **Internal Reliability Testing** (Study 4): We obtained the pairwise similarities between all words in our dictionaries using word embeddings. These metrics indicated that words within dictionaries were more semantically similar than words between dictionaries, suggesting the expected internal consistency.
5. **Validity testing** (Studies 5–8): We found evidence for the validity of our dictionaries across four separate metrics. First, we established that an important subset of our dictionaries showed convergent and divergent validity based on pre-existing dictionaries. Second, for the theory-driven dictionaries, we used open-ended data to predict scaled warmth, competence, and beliefs ratings. These results showed that our dictionaries predicted social groups' predicted warmth, competence, and beliefs. In a third study, we presented human coders with subsets of words from each dictionary and asked them to rate how much the subsets referred to different contents. Participants were able to place the words in the expected dictionary with great accuracy. In a final study, we validated the dictionaries using natural language text that is likely to be the target of the instrument in non-experimental settings. Thus, we successfully created and validated high-coverage dictionaries in an area that lacked such instruments.

Text data is vital for new and renewing fields of psychology. Developments in machine learning fields such as natural language processing have opened the door for psychologists to tap into these so-far underused sources of information. Being able to identify constructs of interest in text and create instruments for their measurement will generate opportunities to expand the science by allowing us to ask new questions or extend previous findings to a wider variety of contexts of potential higher ecological validity. In fact, preliminary versions of these dictionaries have already been used in research including the study of spontaneous stereotyping (Nicolas et al., under review), how social opinions may leak through non-verbal gestures (Lakshmi, Fiske, & Goldin-Meadow, in prep.), and how stereotypical biases in investment decisions from social interaction verbal data (Hu & Ma, 2020). Boghrati and Berger (under review) used the preliminary dictionaries to study a quarter of a million songs over 50 years and found that women were less likely to be associated with competence traits in song lyrics, with relative improvements over decades, and with variations across genres. In the future, we expect these dictionaries to be useful in the study of first impressions from faces, schematic processing (e.g., does stereotyping a target along one dimension activate another?), and dual process theories (e.g., recent research suggests that stereotypes extracted from online data resemble more implicit vs. explicit attitude processes; Kurdi et al., 2019). In practical terms, the dictionaries may also be useful, for example, in recognizing hate and discriminatory text online or biases in machine learning models (e.g., see Caliskan et al., 2017). Finally, we believe that the dictionary creation method used here and made accessible through the SADCAT package (https://github.com/gandalfnicolas/SADCAT) will be useful in the creation of other needed dictionaries for psychological constructs in social and personality psychology (e.g., group entitativity, psychological essentialism beliefs). This is particularly the case given the increasing understanding that the digital behavioral footprint, including text data, must play a central (yet well-considered) role if we are to continue advancing our science (e.g., Boyd et al., 2020).

In this article we have created novel dictionaries for the measurement of stereotype content, and described their properties, coverage, reliability, and validity. We have also provided a tutorial on how to create semi-automated dictionaries for the measurement of other psychological constructs in text data. Previous approaches in psychology heavily rely on multiple human judges to create the totality of the dictionaries over multiple iterations of group discussion and subjective decisions, which is resource-intensive and time-consuming, and risks introducing selection biases based on a specific group of judges who might differ from other judges. On the other hand, the approach provided here largely automates the process,

greatly reducing the time and resources necessary for the creation of the dictionaries. (We provide all the R code used here for readers to be able to implement the procedures for creating dictionaries of their own.)

The use of WordNet in the development of dictionaries has multiple other advantages. Given WordNet's multi-sense network it is possible to obtain valence scores for specific senses using SentiWordNet. While most sentiment analyses rely on the words, SentiWordNet provides different valence scores for each sense. This is important for many psychologically relevant words that share meaning with less psychologically relevant words, such as *warmth*, a central concept in the field of stereotype content illustrated here, which has multiple other meanings, including of course physical warmth. This advantage reduces noise in sentiment analyses when the context of interest is known. In addition, knowing the sense of a word allows for superior translation to other languages using tools such as Babelnet (Navigli & Ponzetto, 2012). Babelnet allows translation from WordNet senses into their corresponding sense in other-language WordNets. Translation can depend on context, and this is facilitated by WordNet's structure, otherwise requiring manual translation by fluent speakers of the target language. Given the neglect of cross-cultural research, using WordNet and Babelnet to study text in multiple languages can provide a fruitful avenue for the generalizability of psychological findings. In fact, Spanish translations of the dictionaries using Babelnet on the corresponding senses are included in the R package as a preliminary instrument (pending further validation). Others have used automated methods to translate, for example, the LIWC dictionaries (van Wissen & Boot, 2017), as well as multi-lingual research using methods such as topic modeling and the Meaning Extraction Method to study issues ranging from self-schemas to sexual assault experiences (see Chung & Pennebaker, 2019; Ikizer et al., 2019; Rodríguez-Arauz et al., 2017).

15 | LIMITATIONS AND FUTURE DIRECTIONS

Dictionaries, like any instrument, have limitations, particularly when used in long-format text data routinely found in most online sources such as social media. When dealing with sentences and paragraphs, a simple word-counting approach misses some of the sentence-level structure, potentially resulting in more noisy estimates. For example, negations, modifiers, and sarcasm are missed by dictionaries in a word-counting approach (however, our last validation study demonstrated that our dictionaries are valid measures of stereotype in long text, outperforming existing instruments). Additionally, context and domain-dependence may be issues when applying these dictionaries in text from different domains from those validated. Thus, validation context must be kept in mind when deploying dictionaries (see Van Atteveldt & Peng, 2018). Some words included in our dictionaries may not show face validity in certain contexts, and sense disambiguation may help. For example, our Morality dictionary includes the word *setup* ("an act that incriminates someone on a false charge", see WordNet Online: https://wordnet.princeton.edu/). However, in a context where other senses of setup (e.g., "equipment designed to serve a specific function") are more common/appropriate, this may be problematic, particularly for frequent words. Combining dictionaries with modern word embeddings that incorporate some degree of sense disambiguation (e.g., Universal Sentence Encoder, Cer et al., 2018; these embeddings are included in the R package), in a way similar to that described in the Supplement, may be helpful for these purposes. Finally, we validated our dictionaries in four different ways, but the focus for both development and validation was on their use as an instrument in explicit target impression descriptions (and not, for example, implied content, measurement of speaker traits or writing style, etc.).

Among other limitations, dictionaries provide categorical measures of a word's semantic association with an overarching topic. However, words differ in their prototypicality for the specified construct (e.g., in a dictionary for sociability, the word *sociable* may be more prototypical than the word *extroverted*) and may belong to multiple constructs simultaneously to different extents (e.g., the word *extroverted* may be classified as related both to sociability and to assertiveness to different extents). To address some of the limitations of dictionaries, they can combine with additional natural language processing methods (e.g., see Supplement; Garten et al., 2018). Finally, a multi-method approach is recommended to study text data, when possible, to balance out potential biases in the dictionary creation process (e.g., biases in WordNet or other training data used for the creation of the models; e.g., Caliskan et al., 2017), to incorporate domain expertise from human coders, among other benefits of robustness checks.

16 | CONCLUSION

In this article, we created and validated novel stereotype content dictionaries that accounted for over 80% of the stereotypes of a representative sample of social groups. We also provide guidance and examples on how to import natural language processing methods, specifically WordNet and word embeddings, into the automation, creation, and evaluation of psychological text instruments. Text data open the possibilities to ask novel questions about behavior in the laboratory and beyond and may provide a way to improve both psychological theory and practice. We hope that the procedures outlined here greatly facilitate the use of text data to complement traditional approaches in social psychology and the study of stereotypes.

DATA ACCESSIBILITY STATEMENT

All data and code for the analyses presented here are also available at https://osf.io/yx45f/.

CONFLICT OF INTEREST

The authors declare no conflicts of interest with respect to the authorship or the publication of this article.

ETHICAL STATEMENT

All validity studies with human subjects were approved by the University ethics committee, and adhered to the ethical guidelines specified in the APA Code of Conduct and the US Federal Policy for the Protection of Human Subjects (including informed consent, right to withdraw, and debriefing).

ORCID

Gandalf Nicolas https://orcid.org/0000-0001-8215-1758

REFERENCES

Abele, A. E., Hauke, N., Peters, K., Louvet, E., Szymkow, A., & Duan, Y. (2016). Facets of the fundamental content dimensions: Agency with competence and assertiveness—Communion with warmth and morality. *Frontiers in Psychology*, *7*, 1810. https://doi.org/10.3389/fpsyg.2016.01810

Abele, A. E., Uchronski, M., Suitner, C., & Wojciszke, B. (2008). Towards and operationalization of fundamental dimensions of agency and communion: Trait content ratings in five countries considering valence and frequency of word occurrence. *European Journal of Social Psychology*, *38*(7), 1202–1217. https://doi.org/10.1002/ejsp.575

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)* (pp. 2200–2204).

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137–1155. Available at: http://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf

Bergsieker, H. B., Leslie, L. M., Constantine, V. S., & Fiske, S. T. (2012). Stereotyping by omission: Eliminate the negative, accentuate the positive. *Journal of Personality and Social Psychology*, *102*(6), 1214–1238. https://doi.org/10.1037/a0027717

Boghrati, R., & Berger, J. (under review). Quantifying 50 years of Misogyny in Music.

Boyd, R. L. (2017). Psychological text analysis in the digital humanities. In S. Hai-Jew (Ed.), *Data analytics in digital humanities* (pp. 161–189). Springer International Publishing. https://doi.org/10.1007/978-3-319-54499-17

Boyd, R. L., Pasca, P., & Lanning, K. (2020). The personality panorama: Conceptualizing personality through big behavioural data. *European Journal of Personality*, https://doi.org/10.1002/per.2254

Boyd, R. L., & Pennebaker, J. W. (2017). Language-based personality: A new approach to personality in a digital world. *Current Opinion in Behavioral Sciences*, *18*, 63–68. https://doi.org/10.1016/j.cobeha.2017.07.017

Brants, T., & Franz, A. (2006). *Web 1T 5-gram Version 1 LDC2006T13. DVD*. Linguistic Data Consortium.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Carnaghi, A., Maass, A., Gresta, S., Bianchi, M., Cadinu, M., & Arcuri, L. (2008). *Nomina sunt omina*: On the inductive potential of nouns and adjectives in person perception. *Journal of Personality and Social Psychology*, *94*(5), 839–859. https://doi.org/10.1037/0022-3514.94.5.839

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y. H., Strope, B., & Kurzweil, R. (2018). *Universal sentence encoder*. arXiv preprint arXiv:1803.11175.

Chakrabartty, S. N. (2018). Cosine similarity approaches to reliability of Likert scale and items. *Romanian Journal of Psychological Studies*, *1*(6), 3–16. https://ssrn.com/abstract=3202379.

Chung, C. K., & Pennebaker, J. W. (2019). Textual analysis. In H. Blanton, J. M. LaCroix, & G. D. Webster (Eds.), *Frontiers of social psychology. Measurement in social psychology* (pp. 153–173). Routledge/Taylor & Francis Group.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology*, *92*(4), 631–648. https://doi.org/10.1037/0022-3514.92.4.631

Decter-Frain, A., & Frimer, J. A. (2016). Impressive words: Linguistic predictors of public approval of the US congress. *Frontiers in Psychology*, *7*, 240. https://doi.org/10.3389/fpsyg.2016.00240

Dewey, J. (1910). How we think. Courier Corporation. Available from: https://pure.mpg.de/rest/items/item_2316308/component/file_2316307/content

Dupree, C. H., & Fiske, S. T. (2019). Self-presentation in interracial settings: The competence downshift by White liberals. *Journal of Personality and Social Psychology*, *117*(3), 579–604. https://doi.org/10.1037/pspi0000166

Ellemers, N. (2017). *Morality and the regulation of social behavior: Groups as moral anchors*. Routledge.

Elliot, A. J., Eder, A. B., & Harmon-Jones, E. (2013). Approach–avoidance motivation and emotion: Convergence and divergence. *Emotion Review*, *5*(3), 308–311. https://doi.org/10.1177/1754073913477517

Fellbaum, C. (Ed.) (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, *27*(2), 67–73. https://doi.org/10.1177/0963721417738825

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*, 878–902. https://doi.org/10.1037/0022-3514.82.6.878

S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.). (2010). *Handbook of Social Psychology* (Vol. 1). John Wiley & Sons.

Fiske, S. T., Nicolas, G., & Bai, X. (in press). Stereotype content model: How we make sense of individuals and groups. In P. A. M. Van Lange, E. T. Higgins, & A. W. Kruglanski (Eds.). *Social psychology: Handbook of basic principles* (2nd ed). Guilford.

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. https://doi.org/10.3758/s13428-017-0875-9

Gendron, M., Roberson, D., & Barrett, L. F. (2015). Cultural variation in emotion perception is real: A response to Sauter, Eisner, Ekman, and Scott (2015). *Psychological Science*, *26*(3), 357–359. https://doi.org/10.1177/0956797614566659

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science*, *24*(1), 38–44. https://doi.org/10.1177/0963721414550709

Hu, A., & Ma, S. (2020). *Human Interactions and Financial Investment: A Video-Based Approach*. Available at SSRN: https://ssrn.com/abstract=3583898

Ikizer, E. G., Ramírez-Esparza, N., & Boyd, R. L. (2019). # sendeanlat (# tellyourstory): Text analyses of tweets about sexual assault experiences. *Sexuality Research and Social Policy*, *16*(4), 463–475. https://doi.org/10.1007/s13178-018-0358-5

Iliev, R., Dehghani, M., & Sagi, E. (2015). Automated text analysis in psychology: Methods, applications, and future developments. *Language and Cognition*, *7*(2), 265–290. https://doi.org/10.1017/langcog.2014.30

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2014). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, *33*(2), 125–143. https://doi.org/10.1177/0261927X13502654

Katz, D., & Braly, K. (1933). Racial stereotypes of one hundred college students. *The Journal of Abnormal and Social Psychology*, *28*(3), 280. https://doi.org/10.1037/h0074049

Kenter, T., & De Rijke, M. (2015). Short text similarity with word embeddings. In *Proceedings of the 24th ACM international on conference on information and knowledge management* (pp. 1411–1420).

Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., & Alves, H. (2016). The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology*, *110*(5), 675–709. https://doi.org/10.1037/pspa0000046

Koch, A., Imhoff, R., Unkelbach, C., Nicolas, G., Fiske, S. T., Terache, J., Carrier, A., & Yzerbyt, V. (2020). Warmth is a personal matter: Consensus reconciles the Agency-Beliefs-Communion (ABC) model with the Stereotype Content Model (SCM). *Journal of Experimental Social Psychology*, *89*, 1–12. https://doi.org/10.1016/j.jesp.2020.103995

Kteily, N. S., & Bruneau, E. (2017). Darker demons of our nature: The need to (re) focus attention on blatant forms of dehumanization. *Current Directions in Psychological Science*, *26*(6), 487–494. https://doi.org/10.1177/0963721417708230

Kurdi, B., Mann, T. C., Charlesworth, T. E., & Banaji, M. R. (2019). The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences*, *116*(13), 5862–5871. https://doi.org/10.1073/pnas.1820240116

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, *23*(2), 254–277. https://doi.org/10.1093/pan/mpu019

Maass, A. (1999). Linguistic intergroup bias: Stereotype perpetuation through language. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 31, pp. 79–121). Academic Press. https://doi.org/10.1016/S0065-2601(08)60272-5

Maass, A., Salvi, D., Arcuri, L., & Semin, G. R. (1989). Language use in intergroup contexts: The linguistic intergroup bias. *Journal of Personality and Social Psychology*, *57*(6), 981–993. https://doi.org/10.1037/0022-3514.57.6.981

Maks, I., Izquierdo, R., Frontini, F., Agerri, R., Azpeitia, A., & Vossen, P. (2014). *Generating Polarity Lexicons with WordNet propagation in five languages*. Proceedings of LREC2014, Reykjavik.

Meshi, D., Tamir, D. I., & Heekeren, H. R. (2015). The emerging neuroscience of social media. *Trends in Cognitive Sciences*, *19*(12), 771–782. https://doi.org/10.1016/j.tics.2015.09.004

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

Miller, G. (1951). *Language and communication*. McGraw-Hill.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*, 217–250. https://doi.org/10.1016/j.artint.2012.07.001

Nerbonne, J. (2003). Computer-assisted language learning and natural language processing. In R. Mitkov (Ed.), *The Oxford handbook of computational linguistics* (pp. 670–698). Oxford University Press.

Nicolas, G., Bai, X., & Fiske, S. T. (under review). *A spontaneous stereotype content model: Taxonomy, properties, processes, and prediction*.

Nicolas, G., Bai, X., & Fiske, S. T. (2019). Exploring research methods blogs in psychology: Who posts what about whom, with what effect. *Perspectives on Psychological Science.14*(4), 691–704.

Nicolas, G., & Skinner, A. L. (2017). Constructing race: How people categorize others and themselves in racial terms. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (2nd ed., pp. 607–635). Elsevier Science. https://doi.org/10.1016/B978-0-08-101107-2.00025-7

Nicolas, G., Skinner, A. L., & Dickter, C. L. (2018). Other than the sum: Hispanic and Middle Eastern categorizations of Black-White mixed-race faces. *Social and Personality Psychological Science*, *10*(4), 532–541. https://doi.org/10.1177/1948550618769591

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092. https://doi.org/10.1073/pnas.0805664105

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Available at http://liwc.net/howliwcworks.php

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

Pietraszkiewicz, A., Formanowicz, M., Gustafsson Sendén, M., Boyd, R. L., Sikstrom, S., & Sczesny, S. (2018). The big two dictionaries: Capturing agency and communion in natural language. *European Journal of Social Psychology*, *49*(5), 871–887. https://doi.org/10.1002/ejsp.2561

Roberts, S., & Rizzo, M. (in press). The Psychology of American Racism. *American Psychologist*. http://dx.doi.org/10.1037/amp0000642

Rodríguez-Arauz, G., Ramírez-Esparza, N., Pérez-Brena, N., & Boyd, R. L. (2017). Hablo Inglés y Español: Cultural self-schemas as a function of language. *Frontiers in Psychology*, *8*, 885. https://doi.org/10.3389/fpsyg.2017.00885

Sato, S., Gygax, P. M., & Gabriel, U. (2013). Gender inferences: Grammatical features and their impact on the representation of gender in bilinguals. *Bilingualism: Language and Cognition*, *16*(4), 792–807. https://doi.org/10.1017/S1366728912000739

Semin, G. R., & Fiedler, K. E. (1992). *Language, interaction and social cognition*. Sage Publications Inc.

Slatcher, R. B., & Pennebaker, J. W. (2006). How do I love thee? Let me count the words: The social effects of expressive writing. *Psychological Science*, *17*(8), 660–664. https://doi.org/10.1111/j.1467-9280.2006.01762.x

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Todorov, A. (2017). *Face value: The irresistible influence of first impressions*. Princeton University Press.

van Atteveldt, W., & Peng, T. Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2–3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

van Wissen, L., & Boot, P. (2017). An electronic translation of the LIWC dictionary into Dutch. In *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference* (pp. 703–715). Lexical Computing.

Wojciszke, B., Baryla, W., Parzuchowski, M., Szymkow, A., & Abele, A. E. (2011). Self-esteem is dominated by agentic over communal information. *European Journal of Social Psychology*, *41*(5), 617–627. https://doi.org/10.1002/ejsp.791

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Nicolas G, Bai X, Fiske ST. Comprehensive stereotype content dictionaries using a semi-automated method. *Eur J Soc Psychol*. 2021;51:178–196. https://doi.org/10.1002/ejsp.2724