A Spontaneous Stereotype Content Model: Taxonomy, Properties, and Prediction

Gandalf Nicolas[12a], Xuechunzi Bai[2], & Susan T. Fiske[2]

[1] Rutgers University – New Brunswick, NJ, United States

[2] Princeton University, NJ, United States

In press at the *Journal of Personality & Social Psychology*.

This is a non-final, non-copy-edited version of the paper.

Abstract

The Spontaneous Stereotype Content Model (SSCM) describes a comprehensive taxonomy, with associated properties and predictive value, of social-group beliefs that perceivers report in open-ended responses. Four studies (N = 1,470) show the utility of spontaneous stereotypes, compared to traditional, prompted, scale-based stereotypes. Using natural language processing text analyses, Study 1 shows the most common spontaneous stereotype dimensions for salient social groups. Our results confirm existing stereotype models' dimensions, while uncovering a significant prevalence of dimensions that these models do not cover, such as Health, Appearance, and Deviance. The SSCM also characterizes the valence, direction, and accessibility of reported dimensions (e.g., Ability stereotypes are mostly positive, but Morality stereotypes are mostly negative; Sociability stereotypes are provided later than Ability stereotypes in a sequence of open-ended responses). Studies 2 and 3 check the robustness of these findings by: using a larger sample of social groups, varying time pressure, and diversifying analytical strategies. Study 3 also establishes the value of spontaneous stereotypes: compared to scales alone, open-ended measures improve predictions of attitudes toward social groups. Improvement in attitude prediction results partially from a more comprehensive taxonomy as well as a construct we refer to as stereotype representativeness: the prevalence of a stereotype dimension in perceivers' spontaneous beliefs about a social group. Finally, Study 4 examines how the taxonomy provides additional insight into stereotypes' influence on decision making in socially relevant scenarios. Overall, spontaneous content broadens our understanding of stereotyping and intergroup relations.

Keywords: Stereotype content, social cognition, intergroup relations, text analysis, natural language processing

A Spontaneous Stereotype Content Model: Taxonomy, Properties, and Prediction

Categorizing people into groups organizes society and aids in planning behavior (Allport, 1979; Bodenhausen et al., 2012). Categorization may use numerous social stimuli, from faces (Todorov, 2012) to descriptions of behaviors (Wojciszke, 1994). Furthermore, categorization of others into groups reliably activates associated stereotypes (Quinn et al., 2003). But the dimensions along which perceivers evaluate social groups and other social agents remain contested. Decades of research into stereotyping suggest one, two, or three core dimensions of their contents (Fiske et al., 2002; Koch et al., 2016; Osgood et al., 1957; for a review, see Fiske et al., 2016), and controversy remains (Abele et al., 2021). Yet, social reality is complex. A deeper understanding of social perception may thus require greater nuance than current low-dimensional theories examine. Starting closer to the data, the current paper proposes a comprehensive model of the content spontaneously applied to social groups. It documents a taxonomy of spontaneous stereotype contents, associates previously understudied properties, and demonstrates enhanced predictive utility.

The Stereotype Content Model

In the long stereotyping research tradition, two content dimensions have been arguably fundamental, cross-culturally stable, and evolutionarily plausible: Warmth and Competence[1]. According to the stereotype content model (SCM; Fiske et al., 2002), inferring Warmth and Competence serves adaptive functions for human beings. Specifically, when interacting with unknown others, social actors need to determine whether the targets are friends or foes (warm,

[1] Throughout the paper we capitalize the names of stereotype content dimensions to distinguish them from related words.

moral, trustworthy, friendly), and whether they can act on their intentions to help or harm (competent, skilled, agentic, assertive).

Social targets can vary independently on the Warmth and Competence dimensions (Fiske et al., 2002). For example, some social targets appear both warm and competent, such as the middle class or White people in the U.S. Other groups are judged to be both untrustworthy and incompetent, such as homeless people. However, social targets can also seem high on Warmth but low on Competence (e.g., elderly people), as well as low on Warmth but high on Competence (e.g., rich people). Depending on the historical moment, various ethnic and national groups land into these stereotypic quadrants, e.g., in the U.S., Asian people as competent but not warm; Hispanic people as neither competent nor warm; Canadians as both.

The content matters because individuals' focus on other people's character predicts a myriad of outcomes, from emotional responses to interpersonal behaviors (Cuddy et al., 2007). Besides helping and harming, other behaviors associated with Warmth and Competence considerations include impression management (Dupree & Fiske, 2019), interactions across societal and organizational hierarchies (see Fiske et al., 2016), and hiring and performance evaluations (Cuddy et al., 2011).

Considerations of Warmth and Competence extend even beyond other humans, to values and beliefs with respect to animals (e.g., Goodwin, 2015; Sevillano & Fiske, 2016) and organizations (e.g., Malone & Fiske, 2013), informing descriptive ethical issues ranging from veganism and ecological conservation to corporate responsibility.

Current Controversies

Despite its proven utility, the development of the SCM proceeded in an entirely theory-driven manner, working from the assumption that Warmth and Competence would be a good fit

for evaluative dimensions, building on the larger previous person perception literature (e.g., Asch, 1946). As a result, it focused on a subset of the possible dimensions that perceivers may use to make sense of others. But perhaps the taxonomy of stereotypes is much more complex. Many lines of research have studied different stereotype contents, in a non-unified manner. For example, some research has examined intersectional group membership stereotypes, that is, beliefs about a group's members also belonging to other social groups (e.g., Schug et al., 2015). Other research has examined stereotypes about geographic origin (e.g., Lee et al., 2009) or about beauty and physical traits (e.g., Nicolas, Skinner, & Dickter, 2019). And just as the SCM stereotype dimensions have a myriad of associated behaviors and consequences for targets, so can these less-frequently-studied dimensions (e.g., see Nicolas et al., 2017). For example, Geography stereotypes (e.g., related to foreignness) about Asian Americans can result in interracial tension and discrimination (Lee et al., 2009), Deviance stereotypes about ethnic minorities predict social distance in domains of marriage and friendship (Hagendoorn & Hraba, 1989), and Emotions and Health stereotypes can lead to misdiagnoses and discrimination in healthcare settings (Boysen et al., 2006; Neighbors et al., 1989). However, these studies often focus on a single dimension for a limited subset of relevant social groups. This paper aims to unify these prior endeavors by studying multiple dimensions used widely across a society's most salient social groups.

A recent exception to these examinations of very specific social groups, which initially challenged the SCM, is the Agency-Beliefs-Communion (ABC) model of stereotype content (Koch et al., 2016; Koch et al., 2020). The ABC model examined stereotypes of a large representation of social groups in a data-driven manner, allowing for the emergence of dimensions other than Warmth and Competence. In fact, focusing on intergroup similarity in a

spatial arrangement task, the ABC model suggested that socioeconomic Status (which the ABC model refers to as Agency) and progressive-traditional Beliefs were the best-fitting dimensions of stereotype content. Although the SCM proposed Status as a structural antecedent of Competence, and thus acknowledges its role in stereotype content, the SCM did not model Status as a content dimension. Furthermore, the Beliefs dimension was completely unaccounted for by the SCM. However, the ABC model is still considerably low-dimensional for the complexity of social perception (c.f., personality taxonomies with five, six, and more dimensions, e.g., Saucier & Goldberg, 1998).

In addition to the ABC proposal, other social cognition models have advanced additional modifications to the SCM. In particular, several authors have argued that Warmth should be subdivided into the more specific facets of Morality and Sociability, and that Competence should be subdivided into Ability and Assertiveness (e.g., Abele et al., 2016). These different facets of the Warmth and Competence dimensions predict unique outcomes, suggesting that future developments in the field will benefit from taking a closer look at more specific attributions. For example, a growing breadth of research has indicated that Morality is primary amongst the facets of social perception across multiple metrics (Brambilla et al., 2021). Previous research suggests that people report that their groups' Morality is more important to them than their groups' Competence or Sociability, and highly-identified group members ascribe more Morality to the group than do low-identified members (but this is not the case for other dimensions; Leach et al., 2007). Additionally, global impressions of others relate more closely to Morality than Sociability (Goodwin et al., 2014), and Morality information has priority when learning about unknown others (Brambilla et al., 2011). These findings provide a more nuanced picture of "the big two" of Warmth and Competence. Nonetheless, taking a step back, here we present a more

comprehensive taxonomy, sketching a framework for understanding more complex dynamics of stereotyping.

Spontaneous Content

The importance of understanding stereotype content and the current diversity of proposed stereotype dimensions demands a unified taxonomy of contents for a representative sample of societal groups. We believe that traditional metrics such as scales, although useful at exploring knowledge of stereotypes along predefined dimensions, cannot uncover the whole gamut of contents that perceivers possess about a diverse sample of social groups. More novel approaches, such as the ABC's spatial arrangement method, capitalize on abstract measures of similarity to derive models of how perceivers organize groups along content dimensions. However, these methods still need to be correlated with a predetermined set of scale-measured dimensions for interpretation and are thus limited to evaluating only these preselected dimensions (see Koch et al., 2016). In addition, spatial arrangement cannot determine the percentage of stereotype content that can be classified into the dimensions the model identifies (i.e., its coverage), a criterion that can be used to evaluate a taxonomy's comprehensiveness.

In contrast to previous methods, here we propose that free-response, open-ended stereotypes of social groups may best systematically reveal the complex contents that are spontaneously available to perceivers upon encountering a target. Free response tasks have been pivotal in recent attempts to revise and improve upon well-established theories and findings. For example, partially due to reliance on forced-choice tasks, previous research on emotion has posited the existence of universal basic emotions that were closely tied to specific physical representations and were independent of language. However, studies on spontaneous emotion perception reveal a more psychological constructionist perspective, wherein emotion perception

depends on linguistic, cultural, and idiosyncratic factors (Gendron et al., 2014). Additionally, studies have used free-response methods to show how the widespread use of forced-choice tasks in racial categorization research has resulted in biased estimates of categorization rates (Nicolas, Skinner, & Dickter, 2019). Specifically, when not constrained to categories such as "Black", "White", or "Multiracial" to categorize Black-White mixed-race faces (as has been the norm in the field; see Nicolas & Skinner, 2017; Skinner & Nicolas, 2015), participants categorize these faces into alternative categories, such as "Hispanic" and "Middle Eastern". This finding suggests that free response tasks can uncover previously neglected perceptions that may more closely align with real-world perceptions.

In the stereotyping literature, however, free response tasks have rarely been used, and certainly not to the ends and extent examined here. For example, classical studies on stereotype content (e.g., Katz & Braly, 1933) use open-ended responses only as an initial means to obtain more traditional measures (e.g., scales or checklists), rather than as the focus of analysis itself. As a result, they end up focusing on only a subset of the possible dimensions, preselected by the investigator, and obtaining information on recognition (i.e., knowledge) rather than recall (i.e., spontaneous content). Other studies have looked at the open-ended responses themselves (e.g., Niemann, et al., 1994), but focused only on a subset of responses and did not fully characterize the taxonomy and its associated properties. Furthermore, none of the existing studies have used a large, representative sample of social groups, thus potentially being applicable to only a specific subset of social categories. Avoidance of free responses may give researchers an incomplete, at times oversimplified, understanding of how people think about others.

Methodology Advances

Social cognition research has failed to thoroughly address spontaneous stereotype content, in part due to psychology's reliance on theory-driven numerical measures and the complexities of analyzing text data. For example, many studies using text analyses rely on human coders, resulting in codes that reflect the researcher's instructions, do not result in easily interpretable categories (particularly with large amounts of data), and are often prohibitively expensive.

However, by incorporating developments in the computer science and machine learning subfield of natural language processing, the analysis of these kinds of data becomes more manageable. For example, resources such as Wordnet (Fellbaum, 1999) make it possible to create reliable, valid stereotype content dictionaries (Nicolas et al., 2021). Dictionaries are word lists that facilitate coding for a construct of interest by matching the words in the list with text responses. For example, a dictionary validated for the coding of Sociability content may include words such as "friendly", "sociable", and "amicable", among others. If a participant then responds that a target is "smart and friendly," researchers can use the dictionaries to code the response as including Sociability-related content, because one of the words used by the participant matches a word in the Sociability dictionary.

The stereotype content dictionaries developed by Nicolas et al. (2021) accounted for 84% of participants' stereotypes about a small sample of social groups used in the development of the dictionaries. Furthermore, the dictionaries were internally reliable, and dictionary-guided coding of spontaneous stereotypes predicted traditional scale-rated evaluations of social groups. These dictionaries appear in this study as one of the methods to code participants' responses.

More recent natural language analysis advances such as word embeddings are also useful in the study of spontaneous stereotyping. Word embeddings are numerical vector representations

of words in a multidimensional space, based on their co-occurrences in large text corpora (e.g., large news articles archives; see Supplement for more information). Notably, word embeddings allow for a quantitative analysis of participants' text responses. Word embeddings appear in both data- and theory-driven approaches that are elaborated later.

Current Studies

Despite a few studies looking at unprompted stereotypes of specific social groups (e.g., ethnic and racial groups; Katz & Braly, 1933) and content analysis of individual impressions (Fiske & Cox, 1979; Park, 1986), no systematic investigation has examined the content of spontaneous stereotypes across a representative sample of social groups, including gender, racial, and occupational groups, among others. Free responses have the potential to advance psychological theories of perceivers' perceptions and evaluations of themselves and others. This provides an opportunity for interdisciplinary research that integrates insights and methods from fields such as linguistics and computer science. Specifically, we use a task asking participants to list their spontaneous thoughts about a series of social groups, presented one at a time. These responses are then quantitatively analyzed for content dimensions, as well as response order and reaction times, among other measures.

Thus, the current research uses a free-response task and computer-aided coding to study spontaneous stereotype content. Throughout the paper, we use the term "spontaneous" to indicate that participants arrived at the content of their responses without such content being explicitly elicited (e.g., the participant may evaluate a group as "warm" in the free-response task, but Warmth as a content dimension is never elicited or primed). Compare this to traditional scales, where participants are explicitly provided with the content dimensions along which to evaluate targets (e.g., "How warm is group X?"). On the other hand, because we explicitly ask the

participant to provide characteristics of the target, the term spontaneous here does not mean that the process of evaluating the targets occurs automatically (c.f., spontaneous trait inferences; Uleman, 1987).

The current research has several aims. First, it will allow us to revisit and improve existing theories of social cognition by proposing a working taxonomy of stereotype content. Currently, multiple models propose distinct stereotype dimensions, from Warmth and Competence (Fiske et al., 2002) to Status and Beliefs (Koch et al., 2016; 2020). Given the variety of dimensions revealed by different methods, and the lack of basic discovery-driven research using free responses, these studies fill a gap that may help clarify the content of social cognition. As previously discussed, different dimensions predict distinct interpersonal discriminatory behaviors and organizational policy decisions (Fiske & Tablante, 2015). Clarifying which dimensions perceivers use to represent social groups will help us better address some of the presently most relevant social and ethical issues.

Second, the current approach permits exploring critical properties of spontaneous stereotype contents, for example, how representative a dimension is in a perceiver's mental mapping of a social group. Stereotype representativeness may be differentiated from the direction of scores on the dimensions. For example, farmers and Christians may be rated as similarly highly warm and highly competent (direction) using scale averages, but if a perceiver uses mostly Warmth-related words to describe Christians but mostly Competence-related words to describe farmers, then these groups differ on which dimension is most representative of the group's stereotypes. More representative dimensions may be more predictive of attitudes, decision making, and behavior (c.f., Fazio et al., 1986), a possibility we examine in the current paper.

Third, the method uncovers biases in terms of direction and valence of the dimensions. An overall negativity bias has appeared for Sociability and Morality (Fiske, 1980), such that people, for example, pay more attention and give more weight to negative (vs. positive) information about a target's Warmth. However, different dimensions might align with different ends of the negative-to-positive continuum. For example, positive Ability and negative Morality information respectively are more diagnostic, and thus weighted more heavily in impressions of others (e.g., Skowronski & Carlston, 1989), perhaps because negative Ability and positive Morality behaviors are judged as more common (e.g., Mende-Siedlecki et al., 2013). Thus, the valence of Ability and Morality in spontaneous representations may reflect these biases. Studies using scales in similar contexts (e.g., Fiske et al., 2002) are not designed to find such valence/direction asymmetries.

Finally, the current research potentially offers additional insight into various facets of dimensional primacy. Different models of person perception content disagree on which dimensions are primary (e.g., the SCM proposes Warmth and Competence, the ABC model proposes Beliefs and Status). However, the models also differ on how they conceptualize and operationalize primacy. Recent theoretical integration attempts (Abele et al., 2021) have identified how dimension primacy may vary depending on whether it is established based on subjective weight (e.g., information gathering interest; see Nicolas et al., in press), pragmatic diagnosticity (e.g., which dimension is more salient or readily available from target features), processing speed (e.g., which dimension is recognized faster), among others (e.g., predictivity of attitudes and behavior). Thus, the question of dimension primacy is complex. The spontaneous measures we introduce allow for the examination of multiple aspects of primacy, including measures distinct to the method, such as spontaneous prevalence of use of the dimension and

response order. These facets of primacy may differ. For example, while Sociability stereotypes may be more prevalent in stereotypes, and more predictive of attitudes towards targets, they may be provided later in a free response list if the group labels provide more readily available information about Ability or Status (e.g., because the labels themselves contain objective information about these dimensions, such as in “rich”, “poor”, or “homeless” people). In general, we examine dimension primacy from the lenses of prevalence (related to the concept of spontaneous representativeness described above) and response times/order (related to time-based accessibility; see e.g., Fazio et al., 2000).

In a nutshell, we introduce the Spontaneous Stereotype Content Model (SSCM), which proposes an initial comprehensive taxonomy of spontaneous stereotype content. Besides recovering more of the nuance and complexity of social reality, the SSCM also sheds new light on stereotype properties and enhances the prediction of general attitudes and decision making as compared to prior low-dimensional models.

In what follows, Study 1 uses both data- and theory-driven codings of open-ended data (cluster analyses and dictionary classifications, respectively) to uncover spontaneous stereotypes’ taxonomic structure, general properties (e.g., valence), and accessibility (through response order). Study 2 uses a speeded version of the previous task to explore the robustness of the previous study’s findings, and to analyze stereotype accessibility through response times. Study 3 tests the robustness of the model using a variety of alternative methods, including dimension embedding coding and participant self-coding. In addition, Study 3 introduces spontaneous representativeness as a property that provides novel insights into perceptions of social groups. Finally, Study 4 examines the predictive value of the extended taxonomy in various decision-making scenarios.

Study 1:

Initial Cluster & Dictionary Coding of Spontaneous Stereotypes

Study 1 aimed to provide a first look at the content of spontaneous stereotypes. We obtained a large sample of societal groups salient to Americans and asked American online participants to provide the characteristics of the targets that they spontaneously thought about. Using natural language processing methods we examine the prevalence, valence, direction, and accessibility of spontaneous stereotypes.

Method

Participants

Participants were 400 workers recruited through Amazon Mechanical Turk. Participants' mean age was 36, more men (54%, 46% women), and mostly White (78%, 7% Asian, 7% Black, 4% Hispanic, 3% Multiracial). Excluding participants based on an attention measure did not significantly affect the results.

For our initial study, we identified the sample size required to adequately power a between-subjects *t*-test to detect a small-to-medium effect, $d = 0.4$. Despite our design being within-subjects, we used this initial heuristic given the complexity of estimating power for generalized mixed models with crossed random effects and the lack of previous studies using these methods from which to draw an expected effect size. Thus, we chose what we considered to be a conservative and accessible test to estimate a sufficient sample size. Later, we were able to compute more precise power analyses via simulation, which indicated that our sample size achieved over 90% power to detect small ($d = 0.2$) effects given our model specifications. Subsequent studies used these simulations to estimate sample sizes more accurately before data collection. Power analysis for Study 1 was calculated using G*Power (Faul et al., 2009).

Materials and Procedure

To select a sample of salient social groups to be evaluated, we looked to previous literature asking American participants to spontaneously list societal groups (Fiske et al., 2002; Koch et al., 2016). Then, we chose the subset of the most salient group labels (i.e., those mentioned by most participants), resulting in a set of 43 social groups that we use as targets in Study 1 (e.g., people who are "homeless", "vegan", "CEOs", "drug addicts", "undocumented immigrants", "elderly", "Christian"; see Supplement for full list). Group labels reflect the language most frequently used by the participants (based on the literature studies selected).

For the main task, participants saw a random sample of 6 of these groups, presented in random order, and listed 6 characteristics that they spontaneously thought about in relation to the targets. The number of social groups sampled from the total was selected to balance power and survey length (as long surveys could lower data quality via decreased interest and attention; the same logic is applied in all studies designs). For congruence with previous research and reduced social desirability biases, participants were further informed that their responses would be completely anonymous and that they need not personally believe the characteristics listed accurately describe the groups (see Devine, 1989; Fiske et al., 2002). They read, "We are interested in any characteristics, traits, or descriptions of the groups that come up to your mind." A final instruction asked for each response/characteristic to be single words if possible, and a maximum of two words if necessary ("for example, an adjective and a noun").

After participants provided open-ended responses for all targets, they saw these targets again and responded to a series of measures on how society views the targets in a 1 (*not at all*) through 5 (*extremely*) scale. The scales measured Warmth's facets of Sociability (α = .76; items: "friendly" and "sociable") and Morality (α = .91; items: "trustworthy" and "honest"),

Competence's facets of Ability (α = .89; items: "competent" and "skilled") and Assertiveness (α = .81; items: "confident" and "assertive"), as well as Beliefs (α = .82; items: "traditional" and "conservative") and Status (α = .9; "wealthy" and "high-status"). We also asked participants to rate how society views the targets in general attitude/valence (i.e., global evaluations, which are used for subsequent analyses of predictive power), from 1 (*very negatively*) to 5 (*very positively*), as an exploratory measure.

In a final block, participants completed a series of demographics and a question about ingroup membership in any of the social groups they rated (for exploratory purposes). An attention question was also included.

Analysis Strategy

To code the large number of open-ended responses that participants provided we made use of two different dimensionality-reduction approaches. First, we borrowed from modern natural language analysis techniques to obtain a data-driven cluster structure of content. Then, we corroborated these findings in a more confirmatory approach, coding responses through recently developed dictionaries of stereotype content covering multiple semantic dimensions (Nicolas et al., 2021). Both of these approaches allowed us to define an initial taxonomy of spontaneous stereotype content as well as measure how comprehensive the taxonomy is (i.e., how many of the responses it accounts for, a measure which is not possible through traditional approaches such as scales).

Cluster analysis. We start by presenting data-driven results, based on a cluster analysis of the word embedding representations of participants' responses. Word embeddings are numeric vector representations of words, which allow for quantitative analyses.

The specific word embeddings used in this paper are from a Universal Sentence Encoder model (USE; Cer at al., 2018; 600 billion words), a Fasttext model (Bojanowski et al., 2017; 600 billion words), a Glove model (Pennington et al., 2014; 840 billion words), all of which were trained on the Common Crawl (a vast sample of world wide web content), and a Word2vec model (Mikolov et al., 2013; 100 billion words) trained on Google News data. For Study 1' cluster analysis we focused on USE embeddings, which are the most flexible and recently developed from the embeddings discussed (see Supplement for more information). In subsequent studies we averaged the results from the multiple word embeddings to diminish the role of distinctive biases from different models (e.g., due to being trained on different data sources; however, these decisions made little difference, see online repository).

The word embeddings encode semantic relatedness from large corpora of text based on word co-occurrences (i.e., how often two words appear close to each other) by comparing the similarity of the context in which two words appear. Put differently, words that often co-occur with the same set of words tend to be more semantically related to each other. For example, both "liberal" and "democrat" tend to co-occur with words such as "political" or "government", and are thus encoded by similar word embeddings, whereas "liberal" and "short" do not necessarily co-occur often with the same context words and thus their word embeddings are more dissimilar. Using word embeddings, we can get a numeric similarity score (called cosine similarity) between pairs of words (as in Study 1), or between a word and a set of words (as in Study 3's "dimension embeddings"). To illustrate, for pairs of words, "liberal" and "democrat" would get high word embeddings similarity scores, while "liberal" and "short" would get lower scores.

To identify the underlying dimensions in the participants' responses, we first selected words that were provided at least 5 times across participants. We then computed a cosine

similarity matrix from the USE word embeddings of every pair of responses. Next, we performed K-Means clustering analysis to examine which set of words clustered together (see a hierarchical clustering solution in the Supplement). To determine the number of clusters for the K-Means algorithm, we used several metrics provided by the R package NBclust (Charrad & Ghazzali, 2014). Based on convergence across metrics, we decided on a model with 40 clusters (we expected to reduce the number of clusters analyzed by combining similar clusters based on their subjective labeling). See more explorations with different numbers of clusters in the Supplement. Finally, to interpret the cluster results, we obtained the words within each cluster that were most similar to the cluster's centroid, as these should be most prototypical. Then, based on these representative words, two of the authors labeled the clusters, and each was allowed to use one or two dimension labels per cluster. The authors reviewed each other's independent labels until they agreed on one or two dimensions that best fit each cluster.

Stereotype content dictionaries. A second analysis of the data made use of stereotype content dictionaries. These dictionaries consist of lists of words associated with different stereotype contents and have high response coverage (accounting for over 80% of open-ended stereotypes in previous tests), as well as high reliability and validity (Nicolas et al., 2021).

To code participants' responses using the dictionaries, we first preprocessed them (i.e., transformed to lower case, cleared of symbols, and lemmatized [removal of inflectional endings]) such that they matched the format of the dictionaries. Then, we matched the preprocessed responses to all the dictionaries described in Nicolas et al. (2021) and available at www.github.com/gandalfnicolas/SADCAT, and coded each response as either a 0 (absent) or 1 (present) in 14 variables, one per dictionary (see Table 1). These dictionaries include dimensions reported in the literature—such as Sociability, Morality, Ability, Assertiveness, Beliefs, and

Status—as well as many other potentially relevant dimensions (see Table 2 for the names of all dictionaries/dimensions). We also had Warmth and Competence dictionaries which were simple combinations of their facets' dictionaries (Warmth = Sociability + Morality, Competence = Ability + Assertiveness; if a word was in both facets, it was only counted once). A single response could be coded into more than one dictionary. Responses not included in any dictionaries were recoded into a single, separate variable, to quantify coverage.

To simplify the analyses, we summed over each participant's six responses for each dimension (see Table 1). Thus, the outcome response rate variable could range from 0 to 6. Given that we had a count outcome, we used Poisson or negative binomial (if overdispersed and when convergence allowed) mixed models, with participants and targets as random intercepts (models with random slopes did not converge). For presentation of results in tables and figures, we transform these values to percentages.

In addition to coding whether a response was included in a dictionary or not (**prevalence**), we had variables indicating whether the word was low (-1), neutral (0), or high (1) on the dictionary (i.e., the **direction**). For example, "friendly" would be high on the Sociability dictionary, "unfriendly" would be low. For the Beliefs dimension, direction is more arbitrarily defined (following Koch et al., 2016): high words indicate more conservatism/religiousness (e.g., "religious"), while low words indicated more liberalism/secularism (e.g., "democrat"). **Traditional scales** included are also a measure of direction.

We also obtained a **valence** measure using a composite of sentiment dictionaries available through R (see Nicolas et al., 2021). The valence scores ranged from -1 (negative) to 1 (positive). For example, words such as "attractive" (.96) and "righteous" (.94) are scored as more positive, while words such as "unfortunate" (-.97) and "perverted" (-.96) are scored as more

negative. We also computed valence per dimension: for example, if a response was coded as being about Morality, we coded its valence score separately as an indicator of the negativity/positivity of Morality content.

Given that the direction and valence indicators were continuous data, we averaged across all 6 responses and analyzed each using linear mixed models. We note here and throughout that valence and direction correlate highly for most dimensions. For example, being warm (high Warmth) is positive, while being cold (low Warmth) is negative. However, this is not always the case. For example, high Assertiveness could involve more positive traits, such as "confident" or "hard-working", but also more negative traits, such as "aggressive" or "dominant". In addition, the coding method for direction is more theory-driven: which words fall into the high or low poles of each dimension were selected based on the person perception and stereotyping literature. For example, which words refer to high Morality and which to low Morality were selected from items measuring these constructs, and then expanded using synonymy and other semantic relations (see Nicolas et al., 2021). On the other hand, valence scores are based on automatic sentiment analyses, which are often trained on different domains (e.g., product reviews), so they may be noisier. For completeness, we present both metrics.

For analyses of response order accessibility, we used multilevel logistic regressions to predict whether each response was in a dictionary or not. We included trial number (participants provided 6 responses per group so this ranged from 1 to 6) in an interaction with the dimension label and had random factors for participants and groups. Thus, response order analyses examine change in content across the responses a participant gives for each social group.

These variables: Dictionary prevalence, dictionary direction, and dictionary valence are used in most studies, so in Table 1 we illustrate an example coding of a participant's responses to

a social group across these variables. We also show how they are aggregated for the mixed model analyses, where each observation is a participant's response to a specific social group. Additionally, Table 1 includes an illustration of the dimension embedding similarity coding used in Study 3.

All analyses were run using the R packages *lme4* (Bates et al., 2015) and *lmerTest* (Kuznetsova et al., 2017; ANOVAs estimated with Satterthwaite degrees of freedom). For contrasts, we use estimated marginal means and pairwise comparisons with their appropriate multiple comparison corrections using the R package *emmeans* (Lenth, 2016). Effect sizes are calculated ignoring the multilevel structure (see http://jakewestfall.org/blog/index.php/category/effect-size/).

Table 1

Example Sociability and Beliefs Coding of Three of a Participant's Responses to a Social Group

		Dictionary Prevalence		**Dictionary Direction**		**Dictionary Valence**		**Dimension embedding similarity**	
Response Order	**Response**	Sociability	Beliefs	Sociability	Beliefs	Sociability	Beliefs	Sociability	Beliefs
1	friendly	1	0	1	NA	0.73	NA	0.73	0.59
2	religious	0	1	NA	1	NA	-0.03	0.61	0.82
3	liberal	0	1	NA	-1	NA	0.1	0.6	0.75
Aggregated		1	2	1	0	0.73	0.035	0.65	0.72

Note. Dictionary prevalence indicates whether the word is related to the dimension (1) or not (0). Dictionary direction codes for whether the response is high (1), medium (0), or low (-1) on the dimension. For Beliefs, direction ranges from liberalism/non-religiosity (low) to conservatism/religiosity (high). Dictionary valence codes for whether the word is more negative or more positive (ranging from -1 to 1). Direction and valence require that the response is coded into the corresponding dictionary, else they are treated as missing data. Dimension embedding similarity (Study 3) indicates the degree of semantic similarity of the response to the dimension, ranging from low (-1) to high (1) similarity. For analyses we sum over responses for prevalence and average over responses for all other variables.

Results

Taxonomy – Clustering

We start by presenting the K-Means clustering results. Table 2 includes words that illustrate the dimensions and were used in labeling the clusters and a table presented in the Supplement (S7) provides all the top associational words for each of the 40 clusters.

Results suggested several clusters related to well-known dimensions—8 related to Morality, 7 to Sociability, 6 to Assertiveness, 5 to Ability, 4 to Status, and 2 to Beliefs—but also clusters related to less-well-studied dimensions— 4 to intersectional Social Group membership, 3 to Health, 2 to Appearance, 2 to Deviance, and 2 to Emotion. We note that some of the clusters were mixtures of two of these dimensions, and that we were unable to label two of the clusters[2] (see "Other" column, for example in Supplemental table S7). One additional cluster was related to general positive valence words. The existence of multiple clusters for the same dimensions also highlights the point that the dimension boundaries used in this paper can always break down further, and that this is one of multiple possible taxonomies.

Beyond counting the clusters, we can look further into how these clusters describe the current data. Given that we used only words mentioned at least 5 times (to remove more idiosyncratic responses and for ease of interpreting the clusters), these analyses account for only ~74% of the total responses but should provide an initial idea of the content distribution to be expanded by the dictionaries. A chi-square test comparing these dimensions' response rates was significant, $\chi 2(12) = 4512.5$, $p < .001$, meaning that some dimensions were mentioned more frequently than others (specific estimated means and pairwise contrasts appear in the

[2] Mixture and unidentifiable clusters are limitations of unsupervised learning with the word embeddings used here (e.g., because the embeddings do not differentiate multiple senses of a word). For example, in the "Other" cluster of Table S7, the physical senses of words such as "break," "open," and "broke" are semantically related (captured by the k-means), but their person-descriptive senses are not, leading to difficulty in interpretability.

Supplement). As shown in Table 2, when combining across clusters for the same dimension, participants' responses were most often categorized into the Sociability, Morality, Assertiveness, and Ability dimensions[3].

Taxonomy – Dictionaries

Next, we analyzed the data using the stereotype content dictionaries. An initial test of coverage indicated that our dictionaries accounted for 86.4% of the participants' responses. A chi-square test comparing these dimensions' response rates was significant, $\chi 2(14) = 5869.8$, $p < .001$, suggesting differences in the prevalence of the various dimensions (specific estimated means and pairwise contrasts appear in the Supplement).

Breaking this down (see Table 2 and Figure 1), we again find that the facets of Warmth and Competence are the most common dimensions. However, also in line with the previous analysis, we find substantive prevalence of dimensions not included in prominent stereotype content models (e.g., Health, Deviance). We include an "Other" category which grouped less frequent content related to family relations, fortune, insults, art, science, and philosophy.

[3] Alternative models for both k-means and dictionary analyses (Supplement) moved the rankings around slightly (e.g., Ability rising to the most common dimension), but the general pattern remains.

Table 2

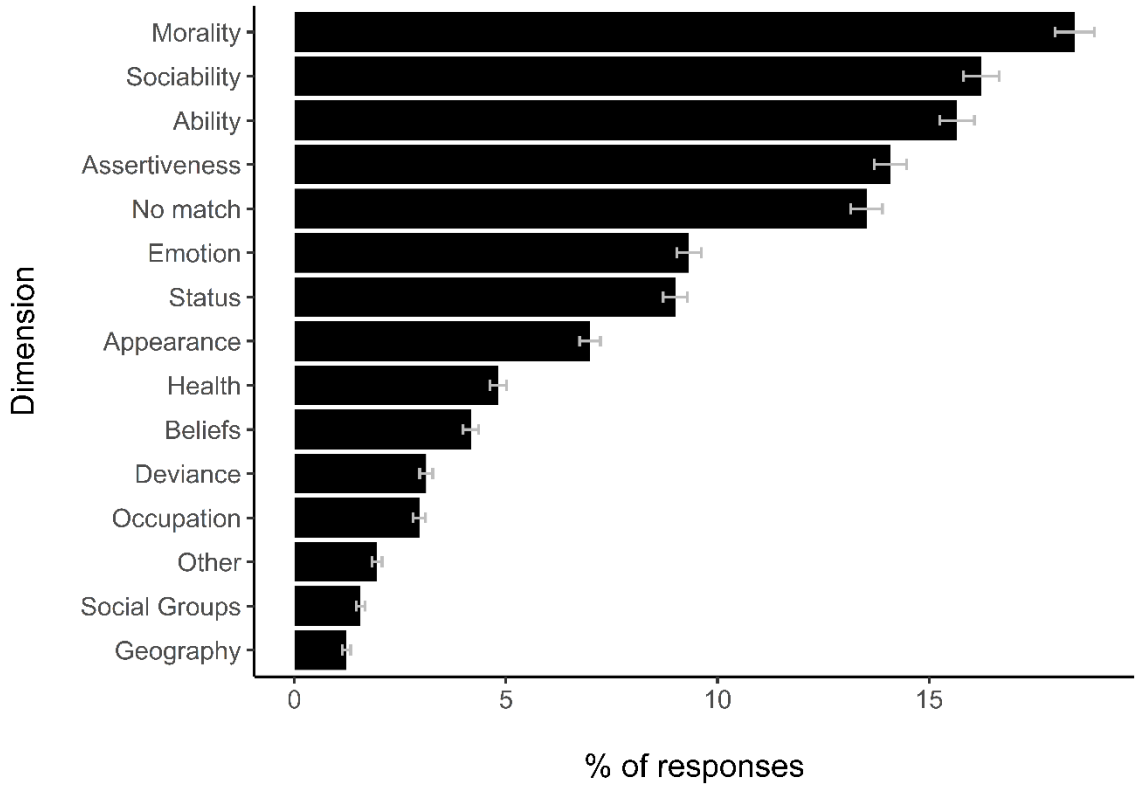
Study 1 Prevalence, Valence, and Direction by Coding Method

Clustering		Dictionary				
Dimension	**Prevalence**	**Dimension**	**Prevalence**	**Valence**	**Direction**	**Example words**
Sociability	14.6[a]	**Morality**	18.5[a]	-.09*[cd]	-.21*[d]	greedy, honest, immoral
Morality	13[b]	**Sociability**	16.2[b]	-.01[b]	-.08[c]	friendly, fun, mean
Assertiveness	12.3[b]	**Ability**	15.7[bc]	.14*[a]	.31*[a]	smart, skillful, uninformed
Ability	10.9[c]	**Assertiveness**	14.1[cd]	-.03[b]	.3*[a]	determined, confident, lazy
Status	9[d]	**No match**	13.5[d]	-.12*[cd]		mr. rogers, 1234, meow
Emotion	5.7[e]	**Emotion**	9.3[e]	-.1*[cd]		sad, happy, anxious
Deviance	5.1[e]	**Status**	9[e]	.01[b]	.15*[b]	wealthy, poor, needy
Health	4[f]	**Appearance**	7[f]	-.12*[cd]		fat, small, attractive
Social groups	3.7[fg]	**Health**	4.8[g]	-.22*[e]		healthy, disabled, sick
Appearance	3.1[g]	**Beliefs**	4.2[g]	-.09[bcd]	.02[bc]	religious, liberal, traditional
Beliefs	2.4[h]	**Deviance**	3.1[h]	-.07[bcd]		different, normal, unique
General valence	2.4[h]	**Occupation**	3[h]	-.03[bc]		doctor, police, unemployed
Other	2.3[h]	**Other**	2[i]	-.19*[de]		daughter, food, science
		Social groups	1.6[ij]	-.1[cde]		man, Black, old
		Geography	1.2[j]	-.19*[de]		foreign, Mexican, country

Note. K-means clustering and Dictionary-coded results shown separately, with dimensions sorted by prevalence within each method. Prevalence values indicate estimated percentage of responses about the dimension. Values for Valence range from -1 (negative) to +1 (positive) and for Direction from -1 (low) to +1 (high). Values with different superscript letters within a column are significantly different from each other ($p < .05$). Valence and Direction values marked by an asterisk are significantly different from zero ($p < .05$). Example words for each dimension are drawn from the most prevalent words for the dimension and may include words from any valence/direction.

Figure 1

Dictionary-coded Spontaneous Stereotype Content Prevalence



Note. The x-axis shows the estimated percentage of responses coded into a dimension. Error bars extend +/-1 standard errors.

Valence and Direction

We also looked at the general valence of the responses and found that they did not significantly differ from neutral on average, M = -0.037, $t(43.6)$ = -0.88, p = 0.385. However, the valence of the responses for different dimensions differed significantly, $F(14, 10760)$ = 29.71, p < .001 (see Table 2; see Supplement for additional information). We found strong positivity of Ability content (i.e., higher prevalence of positive words such as "smart", instead of negative content such as "ignorant"). On the other hand, Morality content was significantly negative (e.g., "immoral", "greedy").

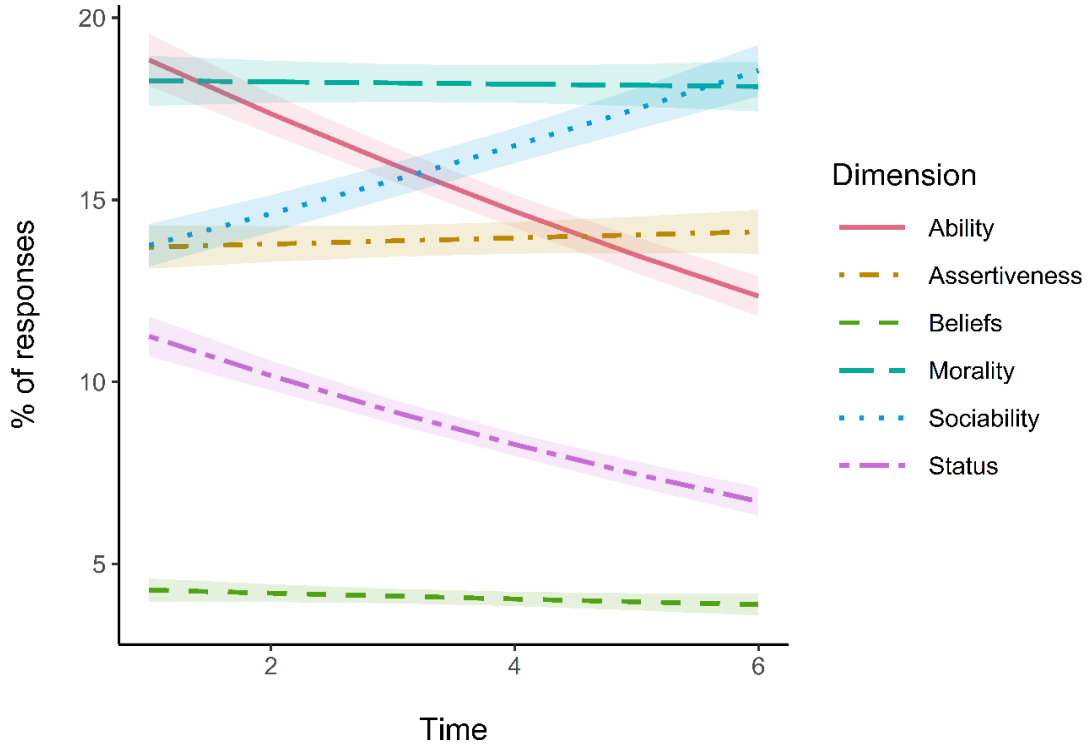
Direction and valence were often correlated, but not necessarily: while direction and valence correlate highly for Sociability, Morality, Ability, and Status (rs > .73, ps < .001), the correlation was more moderate for Assertiveness (r = .41, p < .001), which includes both positive (e.g., "confident") and negative (e.g., "aggressive") words in the high senses, and non-significant for Beliefs (r = .03, p = .452), for which direction here ranges from progressive/non-religious to traditional/religious beliefs. Thus, we also analyzed directional patterns for a subset of dictionaries for which the coding was available and present these results in Table 2.

Accessibility – Response order

Given participants' six responses per group, we investigated the ordering of content over time as an indicator of accessibility. An examination of these patterns showed that Morality, Assertiveness, and Beliefs remained relatively stable across time (ps > .05). On the other hand, Ability, b = -0.10, and Status, b = -0.11, words became less frequent as more responses had been provided, and Sociability-related words became more frequent, b = 0.07, all ps < .001 (see Figure 2 and Supplement for further details).

Figure 2

Response Order Effects for the Theoretical Dimensions/ Facets



Note. Smaller numbers for the x-axis indicate responses provided earlier in the series of responses. The y-axis shows the dimension's prevalence. Error ribbons extend +/-1 standard errors.

Discussion

Study 1 provided an initial look at the spontaneous stereotype content of a representative sample of salient social groups. An entirely data-driven approach showed contemporary stereotype content theories may not be able to account for the full taxonomy of content spontaneously associated with social groups. For example, response clusters associated not only with well-known dimensions such as Morality, but also with other non-big-two dimensions such as Emotions and Health attributions.

A more in-depth analysis based on dictionary-coding of the responses found evidence for further dimensions of content. For example, participants often mentioned the Appearance and

Health of the targets, or their Deviance (familiarity/uniqueness), Occupation, and Geographic origin/nationality. Although stereotype research has previously studied these associations in relation with specific social groups, our taxonomy explores these dimensions across a large sample of salient societal groups and incorporates them into a general model of content. This has relevant implications as it provides a comparative framework where multiple dimensions and how they influence each other can be understood in the context of salient societal groups.

We also found the expected evidence for the high use of dimensions described by dominant general stereotype content models, such as Beliefs (ABC model), and Warmth and Competence (SCM model). In fact, the four most common dimensions (dictionary coded) were the facets from SCM dimensions, in order: Morality and Sociability (Warmth), and Ability and Assertiveness (Competence). Thus, these results support Warmth as the primary dimensions in terms of how frequently it is spontaneously used when thinking about social groups (c.f., Abele et al., 2021). In addition, by focusing on the facets rather than the overarching dimensions, we find that indeed these facets show dissociable patterns (Abele et al., 2016), with significantly different prevalence in stereotypes as well as differences in their average valence. For example, while stereotypes about Sociability were mostly neutral in valence, Morality stereotypes were significantly negative. On the other hand, we found evidence for a positivity bias for Ability but not Assertiveness. We found no evidence of spontaneous responses being more susceptible to social desirability bias than scales (e.g., while scale-rated Warmth was above the mid-point, it was below the mid-point for spontaneous metrics, although admittedly the midpoint may be an arbitrary threshold for subjective evaluations).

Finally, Study 1 results have relevance to questions of temporal accessibility primacy (see Abele et al., 2021). Previous research suggest that Warmth takes temporal precedence over

Competence (e.g. Abele & Bruckmüller, 2011). However, our analysis at the level of facets reveals additional nuance: Although both had similar overall prevalence, Ability words appeared earlier in the list of responses (higher temporal accessibility) while Sociability words were provided later (lower temporal accessibility). We further explore this finding in the next studies.

Study 2: Dictionary Coding of Speeded Spontaneous Stereotypes of 87 Groups

Study 2 used a speeded version of Study 1's task, tapping into less controlled associations, as well as allowing a measure of response time. To further improve generalizability, we also expanded the pool of social groups that participants evaluated.

We expected to largely replicate Study 1 results but note the following distinctions: given that we are only asking for one response in Study 2, and that Ability words occurred earlier and Sociability words occurred later in the sequence of 6 responses in Study 1, we expected Ability responses to rank higher in prevalence and Sociability responses to rank lower in this study. Additionally, we expected Ability-related responses to be provided faster than Sociability-related responses.

Method

Participants. Participants were 250 U.S. Mechanical Turk workers. Participants' mean age was 35, more men (54%, 45% women), and mostly White (74%, 10% Asian, 7% Black, 5% Multiracial). In the previous study, many comparisons (e.g., between facets of a dimension) were closer to small effects (Rate Ratios of ~1.22; Olivier et al., 2016). Given the complexity of power analyses for generalized mixed models, we used a simplified simulation approach through the R package simr (Green & MacLeod, 2015). Because now we had existing data from which to draw an estimated effect size, for Study 2 we took the smallest significant effect size from Study 1 main analysis (i.e., Sociability vs. Morality prevalence difference, *Rate Ratio* = 1.14), and

powered for that effect size using simulations. Such an analysis indicated that 250 participants provided 100% power (95% CI [99.6%, 100%]).

Materials and procedure. We obtained from the literature a sample of 87 social groups. As in Study 1, we selected targets from previous studies in which American participants created lists of groups in their society (Fiske et al., 2002; Koch et al., 2016). In Study 2 we included not only the most mentioned social groups from these studies, but also other group labels that were mentioned less frequently but that were still prevalent and that have been used in previous research (e.g., Koch et al., 2016). Groups added in this study include people who are "celebrities", "libertarians", and "transgender". This was done to improve generalizability (see Supplement for full list). Each participant saw 30 groups, in random order, and provided only one response per group in each trial, using the same instructions as in Study 1 but prompting participants to respond as quickly as possible. Participants saw each group twice across two blocks (once in each block) to perform preliminary explorations of reliability and shared/idiosyncratic responding (unreported; additional repetitions also reduce measurement error; see Martinez et al., 2020). We recorded response times for each response at the time of the first key press.

After participants provided open-ended responses for all targets, they indicated which groups, from those shown, they belonged to, and completed scales measuring their self-rated political orientation and socioeconomic status. These measures were for exploratory purposes. Finally, participants completed demographic items and were debriefed.

Analysis strategy. Having provided evidence of convergence and incremental value of dictionaries (vs. K-Means clustering), we focused on dictionary coding for Study 2. All coding details are the same as in Study 1.

Analysis of response times in mixed models is complex. As a result, we tried different models, as they converged, to evaluate consistent patterns. We computed a response time variable (in ms) excluding responses below 200 ms and above 10,000 ms. We report here the results from a model using a Gamma distribution and an identity link (see Lo & Andrews, 2015) to test for differences in response times between dimensions. Other models (different distributions and cutoffs) are included in the Supplement for robustness examinations.

Results

Taxonomy – Dictionaries

A coverage test indicated that the dictionaries accounted for 85% of the responses. A chi-square test comparing the response rates for these dimensions was significant, $\chi 2(14) = 6277.5$, $p < .001$, again suggesting differences in dimensional prevalence (specific estimated means and pairwise contrasts appear in the Supplement). As in Study 1, the facets of Warmth and Competence were the most prevalent dimensions (see Table 3). However, in line with the response order findings from Study 1, Ability-related words were more prevalent while Sociability were less prevalent, since fewer responses were provided. Many other non-big-two dimensions were also significantly prevalent.

Table 3

Study 2 Prevalence, Valence, Direction, and Response Times (RTs)

Dimension	Prevalence	Valence	Direction	RTs
Ability	18[a]	0.11*[a]	0.24*[b]	2932[a]
Morality	15.3[b]	-0.17*[fg]	-0.27*[d]	3137[e]
No match	15.0[bc]	-0.07*[def]		
Assertiveness	13.1[cd]	0.03[bc]	0.37*[a]	3092[d]
Sociability	11.6[de]	-0.02[cde]	-0.09[c]	3113[f]
Status	11.3[e]	0.06*[ab]	0.3*[ab]	3067[c]
Beliefs	7[f]	-0.04[cde]	0.21*[b]	2974[b]
Appearance	6.8[f]	-0.07*[def]		
Emotion	6.4[f]	-0.11*[ef]		
Deviance	3.9[g]	-0.08*[d-g]		
Health	3.7[g]	-0.2*[g]		
Occupation	3.5[g]	-0.01[b-e]		
Social groups	2.7[h]	0.02[bcd]		
Other	2.4[h]	-0.12*[ef]		
Geography	1.5[i]	-0.09[d-g]		

Note. Dimensions sorted by prevalence. Prevalence values indicate estimated percentage of responses about the dimension. Values for Valence range from -1 (negative) to +1 (positive) and for Direction from -1 (low) to +1 (high). Response times (RTs) are mean latencies measured in milliseconds. Values with different superscript letters within a column are significantly different from each other ($p < .05$). Valence and Direction values marked by an asterisk are significantly different from zero ($p < .05$).

Valence and Direction

In terms of general valence, as in Study 1, responses were not significantly different from neutral on average, $M = -0.02$, $t(96.9) = -0.66$, $p = 0.513$. Looking at the dimensions' valence individually (Table 3), Ability words tended to be positive and Morality words tended to be negative, as in Study 1. Results for direction also largely replicated Study 1 (see Table 3).

Accessibility – Response times

Across all the models we ran (with slight differences on significance of particular pairwise comparisons and ranking), responses related to Ability, Beliefs, and Status tended to

have faster response times, while Sociability, Assertiveness, Morality, and other content responses were slower (see Table 3 and Supplement).

Discussion

Study 2 replicated and extended the findings from Study 1. Specifically, using a larger sample of social stimuli, the same patterns of previously underestimated dimensions emerged as relevant spontaneous stereotypes, in addition to traditional dimensions. Study 2 also found that, similarly to Study 1, Ability as a dimension was more time-accessible than Sociability. Specifically, Ability-related words were provided faster than Sociability-related words. Although there was some variability depending on the model specifications on whether this result achieved statistical significance, Ability responses were provided more quickly than responses associated with other dimensions. In combination with the response order results from Study 1, this finding suggest that if Ability content comes to mind when thinking about a social group, this content is retrieved earlier and faster from memory (on average and compared to most other dimensions).

So far, we have established an initial descriptive model of spontaneous stereotype content that allows for an integration of current stereotype content models. However, questions of robustness to the coding method and practical impact remain. In the next study we revisit prevalence findings using additional coding approaches and examine spontaneous stereotypes' relevance in evaluations of social targets.

Study 3:

Coding Method Robustness and Predictive Value of Spontaneous Stereotypes of 87 Groups

Study 3 took elements from both Studies 1 and 2 to address some limitations and pending issues and to extend the basic findings of this paper.

In particular, we revisited the non-speeded task with a more extensive sample of social groups. We also tested robustness by using additional coding methods: Dimension embeddings similarity and participant self-coding. Dimension embeddings similarity measured the semantic relatedness between the participants' responses and each of the taxonomy dimensions. Compare this to Study 1's more data-driven approach, in which we conducted a cluster analysis based on the semantic relatedness between the word embeddings of participants' responses only. Thus, the current study's use of dimension embeddings allows for a direct robustness check on all the dimensions previously examined through dictionary coding. In addition, participants coded their own responses into the theoretical dimensions, allowing us to compare the (relatively more systematic/consensual) semantic structure revealed by the previous methods with participants' subjective understanding of the dimensional alignment of their responses.

Subsequent analyses in Study 3 focus on the predictive value of spontaneous stereotypes. Global attitude evaluations are an important predictor of real-world outcomes for social targets (e.g., Wallace et al., 2005). Previous studies (e.g, Goodwin et al., 2014; Wojciszke et al., 1998) have used global evaluations of targets to measure the relative weight of different dimensions on attitudes toward social groups.

In Study 3, we used global evaluations of social targets to determine whether spontaneous stereotype content provides insights above and beyond traditional measures of stereotype knowledge (particularly scales). We hypothesized that stereotype representativeness (i.e., the degree to which a perceiver spontaneously associates a stereotype content dimensions with a social group) would interact with stereotype scale-measured direction to improve predictions of global evaluations of social targets. This hypothesis follows from the attitude literature suggesting that the strength of association between concepts and targets, which

representativeness is an indicator of (c.f. Higgins, 1996), moderates the effect of the attitude on outcomes (e.g., Fazio et al., 1986).

Finally, given that we have highlighted the high prevalence of understudied dimensions (e.g., Health and Deviance stereotypes), we ran analyses to explore the predictive advantage of incorporating all the information contained in spontaneous responses (as encoded by word embeddings), above and beyond dimensions from existing stereotype content models. These analyses serve as initial evidence for the predictive benefits of the expanded taxonomy, an issue we explore again in more depth in Study 4.

Method

Participants

Participants were 402 Mechanical Turk workers. Participants' mean age was 35, more men (56%, 43% women), and mostly White (71%, 9% Asian, 9% Black, 7% Multiracial). Using the same power analysis as before, we determined that for this study's design, 400 participants provided 99% power for a small effect as described previously.

Materials and Procedure

We used the same sample of 87 social groups as in Study 2. In the first of three blocks, each participant saw 6 groups, one at a time and in random order, and provided three responses per group, using the same instructions as in Study 1.

In a second block, they saw each of their open-ended responses and were asked to code them into a subset of the dictionary responses, namely the dimensions that were based on theoretical models, or "No match". The question indicated: "Below are the responses you gave for people who are [group label]. Please choose the category, from those provided, that best fits what you meant by your response," followed by, for each of the three responses sequentially,

"Which of the following characteristics fits best what you meant by [response]?" Coding choices were (information on parentheses not shown): "Traditional/Conservative" (high Beliefs), "Progressive/Liberal" (low Beliefs), "Confident/Assertive" (high Assertiveness), "Not confident/Not assertive" (low Assertiveness), "Friendly/Sociable" (high Sociability), "Unfriendly/Unsociable" (low Sociability), "Competent/Skilled" (high Ability), "Incompetent/Unskilled" (low Ability), "Wealthy/High-Status" (high Status), "Poor/Low-Status" (low Status), "Trustworthy/Honest" (high Morality), "Untrustworthy/Immoral" (low Morality), and "No match".

In a third block, participants saw the same groups and rated them on the same scale items as in Study 1 (i.e., now rating the groups instead of the individual responses), ranging from 1 (*not at all*) to 5 (*extremely*) for each dimension. All scales had Cronbach's alphas ≥ .8. In addition, participants rated global evaluations of the target using a 5-point scale (1 – *Very negatively* to 5 – *Very positively*) to indicate how, in general, society views members of each group. Finally, they indicated their demographic information.

Analysis Strategy

Responses self-coded by the participants were analyzed in the same way as dictionary-coded responses: with Poisson or negative binomial mixed models (random factors for subjects and groups).

Dimension embeddings similarity. To obtain dimension embeddings similarities we use word embeddings following slightly different steps than in Study 1. Specifically, we first obtained dimension embeddings by averaging the word embeddings of highly prototypical words for each theoretical dimension (obtained from the literature; see Nicolas et al., 2021). These prototypical words were direction-balanced, including equal numbers of high and low senses.

For dimensions for which we did not have a small set of highly prototypical words from the literature (Appearance, Emotion, Deviance, Social Groups, Occupations, Health, Geography), we instead used all the words in the dimension's dictionary to compute their dimension embeddings[4]. Thus, dimension embeddings encode the average semantic information of words from each dimension.

Subsequently, we obtained the cosine similarity between the word embeddings of each participant's response and each dimension embedding. Thus, for each response, we had a score for its semantic similarity to each of the available dimensions, with scores that could range from -1 (less similar) to 1 (more similar). See Table 1 for an example of dimension embeddings similarity coding.

Because only prototypical subsets of words were used for many dimension embeddings, analyses using dimension embeddings similarity are relatively independent from the dictionaries. In other words, the dictionaries and the dimension embeddings rely on different content indicators for the dimensions, allowing us to further diversify the robustness tests. We used linear mixed models for these analyses with participants and targets as random intercepts (after averaging across the three responses provided by each participant for each group).

Predictive models. To examine the predictive value of spontaneous content information, we used linear mixed models with participants and targets as random factors and focused on the Warmth and Competence dimensions only. In particular, we were interested in the role of spontaneous representativeness (operationalized here as a dimension's dictionary-coded prevalence). We predicted the global evaluations from the scale-measured Warmth and

[4]Because comparisons between embeddings constructed from prototypical vs. whole dictionaries may not be the most appropriate, we conduct an exploratory analysis using researcher-selected subsets of prototypical words for all dimensions, presented in the Supplement. Using this approach results in some differences, but the ranking remains stable for most dimensions.

Competence direction scores, the dictionary-coded Warmth and Competence representativeness, and their interaction. We hypothesized that representativeness and direction would interact, such that a scale's direction is more predictive of attitudes when the dimension is more spontaneously representative. Standardized (without accounting for multilevel structure) Beta coefficients are provided.

Additionally, we used regularized regression through the R package glmnet (Friedman et al., 2010; not accounting for random effects, for simplification and accessibility of complex methods) to test whether adding all the spontaneous information encoded by the word embeddings improved predictions of global evaluations above and beyond the theoretical dimensions. Regularized regression serves to reduce overfitting given the large number of predictors. As a baseline model, we predicted global evaluations from the scale ratings for Warmth, Competence, Beliefs, and Status. For the comparison model, we predict global evaluations from all 512 dimensions of the USE word embeddings (recall each response's USE word embedding is a numerical vector with length of 512 encoding its semantic information), in addition to the baseline model predictors. We compare these models' R^2 and expected the model incorporating spontaneous information about multiple dimensions, beyond those established by the literature, to outperform the baseline model.

Results

Given the response-order patterns and the results from Study 1 (6 responses per group) and Study 2 (1-2 responses per group), we expected prevalence, valence, direction, and response order results for Study 3 (with 3 responses per group) to fall somewhere in between these studies. Indeed, we confirmed this hypothesis and replicated the overall patterns shown in the previous studies. We present these results in the Supplement and instead focus on exploring the

robustness of the taxonomy using different coding approaches, as well as examining predictive value.

Taxonomy – Dimension Embeddings

We analyzed the data using dimension embeddings. The results from this analysis indicated significant differences in the prevalence of the dimensions, $F(12, 30856) = 3225.6$, $p < .001$, and were very similar to those from the dictionaries (see Table 4).

Because some dimension embeddings tended to have high correlations, additional models performed on different dimensional words (e.g., larger set of less prototypical words) and orthogonalized/residualized embeddings of the theoretical dimensions (c.f., Oh et al., 2019) are presented in the Supplement. These models show similar patterns, with some differences in ranking for specific dimensions.

Taxonomy – Self-coding

We examined how participants coded their own responses into the theoretical dimensions. A chi-square test comparing the prevalence of these dimensions was significant, $\chi 2(5) = 720.12$, $p < .001$. The results (see Table 4) indicate one big difference to results from other coding approaches: Participants coded their responses as being about Morality less frequently than the dictionaries and dimension embeddings. Additionally, because participants were only able to code their responses into 6 of the ~ 14 dimensions identified in the taxonomy, "No match" self-codings were the most common.

Direction – Self-coding

In terms of the self-coded dimension direction, participants tended to use high-direction words for all dimensions. However, Morality still had a significantly lower direction than all

others (see Table 4). Direction results for dimension embeddings appear in the Supplement, largely replicating these results.

Table 4

Study 3 Prevalence and Direction by Coding Method

Dimension Embeddings Similarity		Self-coding		
Dimension	**Prevalence**	**Dimension**	**Prevalence**	**Direction**
Ability	0.473[a]	**No Match**	0.69[a]	
Sociability	0.471[a]	**Ability**	0.61[a]	0.41*[a]
Morality	0.463[b]	**Sociability**	0.46[b]	0.17*[b]
Assertiveness	0.456[c]	**Beliefs**	0.35[c]	0.14*[b]
Status	0.439[d]	**Status**	0.34[c]	0.16*[b]
Deviance	0.418[e]	**Assertiveness**	0.32[c]	0.47*[a]
Beliefs	0.413[f]	**Morality**	0.24[d]	-0.03[c]
Emotion	0.408[g]			
Social groups	0.382[h]			
Appearance	0.365[i]			
Occupation	0.354[j]			
Health	0.317[k]			
Geography	0.289[l]			

Note. Dimension embeddings similarity and self-coded results shown separately, with dimensions sorted by prevalence within each method. Prevalence values indicate estimated percentage of responses about the dimension. Values for Direction range from -1 (low) to +1 (high). Values with different superscript letters within a column are significantly different from each other ($p < .05$). Direction values marked by an asterisk are significantly different from zero ($p < .05$).

Spontaneous Representativeness as a Predictor

To test the predictive value of spontaneous stereotypes we examined whether an interaction between a dimension's scale-measured direction and its spontaneous representativeness significantly predicted global evaluations of targets. As expected, we found that indeed Warmth direction and Warmth representativeness interacted to predict global

evaluations, $B = .03$, $t(2365.40) = 2.01$, $p = .044$. Competence direction and representativeness also interacted to predict global evaluations, $B = .04$, $t(2353.03) = 2.90$, $p = .004$. These effects largely replicate the patterns from exploratory analyses of Study 1 and a replication of Study 3 (see Supplement), as evidence of robustness. In the supplement we also examine alternative models (e.g., using dimension embeddings similarity) which provide additional evidence for the predictive value of spontaneous content.

Breaking down the representativeness by direction interaction, we found the expected pattern: the higher the representativeness of a dimension, the larger the impact of its direction on global evaluations. The estimate for Warmth direction was $B = 0.34$ in the lowest Warmth representativeness group (i.e., zero open-ended responses being about Warmth), while it was $B = 0.43$ for the highest Warmth representativeness group (i.e., three open-ended responses being about Warmth; see Figure 3). The estimate for Competence direction was $B = 0.18$ in the lowest Competence representativeness group, while it was $B = .32$ for the highest Competence representativeness group.

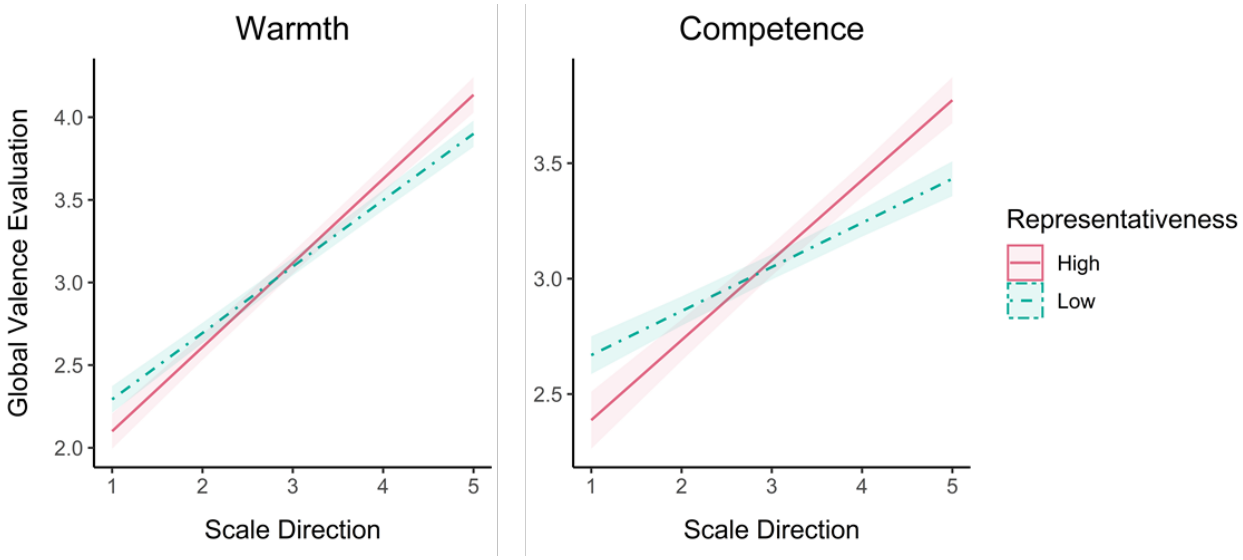
Figure 4 shows the direction and representativeness scores, respectively, for a sample of groups in the study. Note how some groups highly similar on direction for both Warmth and Competence tend to have one dimension more representative than the other. For example, although doctors and nurses are scale-rated very similarly (as highly warm and competent), when people think about these groups, spontaneously they think more about nurses' Warmth (and less about their Competence), and more about doctors' Competence (and less about their Warmth). In other words, Competence is more representative of doctors, Warmth is more representative of nurses, highlighting an example of the kind of unique information that can be gleaned from spontaneous representativeness. To put differently, traditional scale measures suggest doctors

and nurses are perceived similarly, but this spontaneous approach indicates that is not entirely right – the former are perceived primarily as more competent and the latter are perceived as primarily warmer, and it is doctors' Competence and nurses' Warmth that have a higher weight on their corresponding positive global evaluations.

Other groups showing a similar pattern include Black and Asian people being rated similarly average, but Warmth being more representative of stereotypes about Black people and Competence of stereotypes about Asian people. On the other hand, targets who are obese or poor are amongst the groups rated the lowest on both Warmth and Competence direction, yet neither content is as spontaneously representative for these groups as it is for other groups. Instead, these groups are on average higher (vs. other groups) on spontaneous content about alternative dimensions such as Appearance, Health, and Emotions. In many cases, the representativeness of non-big-two dimensions is higher than for Warmth and Competence, which suggests that for many groups, understanding perceptions along alternative dimensions of the taxonomy may be as (or more, depending on context) important as understanding their Warmth or Competence. Also note (in Figure 4) direction-representativeness asymmetries for "hackers", "children", "students", and "rich", among others.

Figure 3

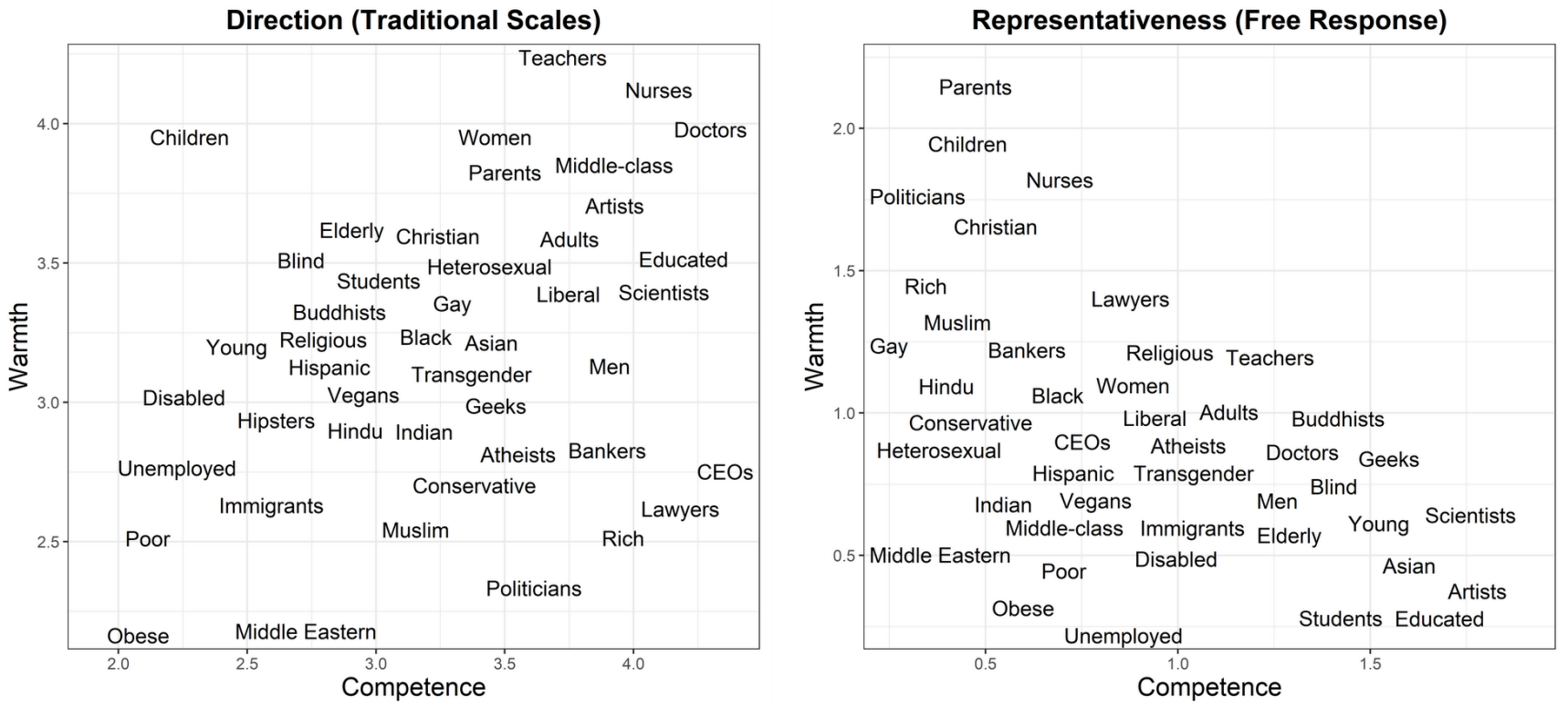
Scale Direction by Spontaneous Representativeness Interaction with Global Valence Evaluation as Outcome



Note. Low representativeness stands for zero Warmth/Competence-related words, high stands for three Warmth/Competence-related words. Error ribbons extend +/-1 standard errors.

Figure 4

Comparison of Warmth and Competence Scale-Measured Direction and Spontaneous Representativeness for a Sample of Social Groups



Note. Direction ranges from "less" to "more" warm or competent. Representativeness ranges from "fewer" to "more" of the social group's spontaneous stereotypes being *about* Warmth or Competence (regardless of direction). The negative relation between Warmth and Competence representativeness reflects the fact that more responses related to one dimension results in fewer opportunities for responses related to the other dimensions.

Word Embeddings Encoding all Spontaneous Information as Predictors

To examine the role of dimensions beyond Warmth and Competence, we used a final exploratory analysis to check whether adding information about all the semantic content in the open-ended responses improved upon a model including only dominant theoretical dimensions. For this baseline model (predicting global evaluations from scaled Warmth, Competence, Beliefs, and Status), $R^2 = .475$. The comparison model expanded this baseline model by adding all 512 dimensions of word embeddings and resulted in an $R^2 = .545$. Therefore, this exploratory model comparison shows an R^2 increase of .07 by incorporating spontaneous information beyond prominent models' dimensions. For context, the effect size of the added accounted-for variance is higher than some estimates of median effect size in social psychology (R^2 *increase* ~ .032; Richard et al., 2003). This result provides preliminary evidence of predictive value of the additional dimensions identified in our taxonomy.

Discussion

Study 3 further built on our previous findings, using a wider variety of methods to detail the taxonomy of spontaneous stereotypes. Using a new semantic metric (dimension embeddings similarity), we replicated the main findings from previous studies using dictionaries: high prevalence of Warmth and Competence facets, along with significant prevalence of additional less-studied stereotype dimensions, such as Health and Deviance[5].

Similar patterns are suggested by the self-coding responses, where participants were only able to code responses into a subset of established dimensions (e.g., Warmth, Competence facets). "No match" responses were ~13% and ~23% of responses coded by the dictionaries and the participants, respectively. Responses dictionary-coded as belonging to dimensions not

[5] An additional study conducted with participants from Spain and Colombia is included in the Supplement, with results replicating most patterns.

included in the self-coding options (e.g., Appearance, Deviance) accounted for > 36% of self-coded "No match" responses. So, if provided in the self-coding scales, many of these responses could have been self-coded into alternative dimensions of the taxonomy. We note, however, that part of the point we advance is that forced-choice tasks inherently change responding patterns (as seen for Morality codings; see also Nicolas, Skinner, & Dickter, 2019), even if all the response options are presented and the forced-choice task is used to code open-ended responses. In other words, indirect semantic coding (e.g., dictionary coding) of spontaneous responses can provide information that is unavailable through methods that rely, in one way or another, on participants choosing from an explicitly provided list of response options. Additionally, given how cumbersome it becomes for participants to rate many dimensions across multiple groups, dictionaries and word embeddings shift that burden to automated coding procedures, making research into these nuances more accessible. Moreover, dictionaries and word embeddings are more reproducible, and their validity more easily measured (vs. self-coding which depends on the participant sample and coding instructions).

On the other hand, dictionary and word/dimension embeddings codings have limitations as well. For example, dictionary coding results in all-or-nothing coding of words into a dimension, and thus fails to capture that words relate to content to different degrees. Embeddings better capture the continuous nature of semantic relations. However, word embeddings are also more influenced by cultural biases embedded in the training data used to compute them. For example, words referring to social groups and identities (e.g., "gay"), may be coded as not only highly semantically related to the Social Groups dimension, but also as relatively highly related to the Morality dimension. Such results may reflect the fact that moralizing language is often used to discuss social groups in word embeddings training data (e.g., the Common Crawl;

Luccioni & Viviano, 2021) and that social group labels are often associated with cultural biases (e.g., the use of "gay" as a general negative term; Nicolas & Skinner, 2012). The dictionaries, developed through a literature search (Nicolas et al., 2021) and lexical expansion based on more vetted data (Wordnet; Fellbaum, 1998) may be much less susceptible to these biases, although by no means eliminate them. Thus, our approach supports multimethod analyses for the study of stereotypes, integrating spontaneous metrics with more traditional scales (along with other less explicit methods), and balancing the strengths and limitations of the various methods.

Study 3 also provided a complementary look at the relevance of spontaneous stereotypes. In particular, we found that information encoded on spontaneous responses improved predictions of global evaluations significantly above scales. Our results suggest that spontaneous representativeness interacts with traditional measures of stereotype direction to increase their effect. To illustrate, the degree to which a social group's high or low Warmth matters for evaluations of the target depends on how representative Warmth stereotypes are of the target. Thus, knowing not only a perceiver's evaluation along a stereotype dimension's direction, but also the representativeness of that dimension is important for attempts to predict and understand social behavior. For example, although doctors and nurses shared similar Warmth and Competence scale scores, they diverged sharply in terms of which dimension was more representative of each group.

Finally, we found that incorporating information about the full taxonomy of content also improves predictions of global evaluations. Specifically, in models controlling for Warmth, Competence, Beliefs, and Status scores, semantic information about additional content

dimensions, such as those we propose in the taxonomy led to a considerable increase in explained variance[6]. We further explore the value of the expanded taxonomy in the next study.

Study 4:

Predictive Value of a More Comprehensive Taxonomy

We have reviewed in the literature how the dimensions incorporated into our taxonomy, beyond Warmth and Competence, have an impact on a variety of outcomes (e.g., Hagendoorn & Hraba, 1989; Lee et al., 2009). In addition, in the previous study we provided evidence that models incorporating all semantic information, including about these alternative dimensions, improved the prediction of general attitudes. But to further drive home the point that a more comprehensive taxonomy can provide needed richness to problems relevant to social group perception, in this study we explore how these alternative dimensions, controlling for Warmth and Competence, can further our understanding of socially relevant decision making.

Drawing from the attitude literature (e.g., principles of compatibility; Ajzen & Fishbein, 1977) on increased predictive value for higher congruence between targets and elements of evaluation, we expected that decision making scenarios more directly related to the taxonomic dimension would be predicted by these dimensions even after controlling for big two evaluations. Thus, this study highlights how the full taxonomy may not only be useful in understanding general evaluations of social groups, but also allows greater insight when specific application contexts are considered.

Method

Participants

[6] Certainly, due to the black-box nature of the embeddings, additional non-dimensional information captured by the models (e.g., linguistic patterns such as noun vs. verb use; Carnaghi et al., 2008) may also partially contribute to the change in explained variance.

Participants were 418 Prolific[7] workers. Participants' mean age was 33, more men (51%, 47% women), and most were White (69%, 13% Black, 6% Asian, 6% Hispanic). This study was powered as Study 3.

Materials and Procedure

This study included many of the measures from previous studies, while adding items designed to test the predictive value of the dimensions we incorporate into the taxonomy. Each participant was assigned a random sample of three groups (out of 72 of the Studies 2 and 3 targets, with some small variations in labeling[8], see Supplement). In a first block, participants were asked to imagine they were in a decision-making position for a series of scenarios. We expected that dimensions in the taxonomy, beyond Warmth and Competence, would predict decision making in these real-life relevant contexts across multiple groups.

For each scenario participants rated (1- *Not at all* to 5- *A lot*) how much they would prioritize each of the groups they saw: "in terms of early eligibility for a covid vaccine" (Vaccination priority; hypothesized importance of Health stereotypes), "for government programs that make psychological and emotional counseling more accessible and available to them" (Counseling programs; hypothesized importance of Emotion stereotypes), "for programs aimed at ensuring they feel included in their community or workplace" (Inclusion interventions; hypothesized importance of Social groups and Deviance stereotypes), "for programs aimed at ensuring they are not unfairly stopped by immigration officials" (Immigration policing; hypothesized importance of Geography stereotypes), "for programs aimed at detecting/preventing discrimination in facial recognition technologies (Face recognition

[7] Prolific offers similarly highly representative samples as Mechanical Turk, and the quality of the responses was equal or superior to that of the previous studies.

[8] Changes were made given the smaller number of groups participants would see and to remove irrelevant or relatively outdated labels. For example, given the U.S. decision-making context, we removed the "American" label.

discrimination; hypothesized importance of Appearance stereotypes)", "for programs aimed at detecting/preventing discrimination based on LinkedIn profiles" (Hiring discrimination; hypothesized importance of Appearance stereotypes). We had no hypotheses about the direction of the relationships, just that the corresponding dimension would be predictive and thus useful in understanding the real-world-relevant decisions at hand.

In a second block, participants completed the open-ended task, seeing each group one at a time and providing four open-ended responses for each. Then, participants saw the groups again, one at a time, and rated them on 13 items (1- *Not at all* to 5 - *Extremely*) measuring direction across multiple of the taxonomy dimensions ("Sociable", "Moral", "Intelligent", "Confident", "Conservative", "Wealthy", "Healthy", "Angry", "Sad", "American", "Physically attractive", "Having recognizable features", "Unique/Different from most people"). Note that for dimensions without a general direction (e.g., it is unclear what being high or low on Appearance is), we selected items based on evaluations of semantic subdimensions (e.g., attractiveness and feature distinctiveness/recognizability for Appearance; see Nicolas et al., 2021). Finally, participants completed demographics questions.

Analysis Strategy

To analyze the results, we used linear mixed models with participants and targets as random factors. We predicted the corresponding decision-making item from the hypothesized dimension as the primary analysis. Because the main aim of this study was to further demonstrate the utility of measuring the multiple dimensions from the taxonomy, we used not only measures of spontaneous representativeness (coded with dictionaries), but also measures of direction (using scales, which tend to have larger effect sizes than dictionary direction). We run models with both predictors simultaneously, in line with our argument that both

representativeness (prevalence) and direction are relevant measures. We furthermore run models controlling for the spontaneous prevalence and scale-based direction of the well-established dimensions of Warmth and Competence, to bolster the case for a comprehensive approach. Standardized (without accounting for multilevel structure) Beta coefficients are provided.

Results

Across all decision-making scenarios we found robust evidence for the predictive value of the dimensions tested, including in terms of incremental validity above the comparison model (Warmth and Competence; Fiske et al., 2002). In addition to the dimensions expected to relate to the scenarios, many others were predictive of decision-making (controlling for each other), in line with the potentially unexpected utility of exploring multiple content dimensions. Table 5 shows the main results for each scenario with a relevant dimension from the taxonomy.

Table 5

Select Hypothesized Dimensions Related to the Decision-Making Scenarios

Decision-making scenario	**Predictor dimension**	**Variable**	**Beta**	***p***
Vaccination priority	Health	Representativeness	.051	.023*
		Direction	.018	.489
Counseling programs	Emotions	Representativeness	.053	.036*
		Direction	.058	.028*
Inclusion interventions	Deviance	Representativeness	-.022	.365
		Direction	.197	< .001*
Immigration policing	Geography	Representativeness	.054	.027*
		Direction	-.06	.022*
Face recognition discrimination	Appearance	Representativeness	-.048	.072
		Direction	.181	<.001*
Hiring Discrimination	Social groups	Representativeness	-.043	.024*

Note. Controlling for Warmth and Competence did not impact statistical significance. For the Emotions scale, the "Sad" item is presented ("Angry" was not significant), For the Appearance scale, the "Having recognizable features" item is presented ("Beauty" was not significant). For the Deviance scale, the "Unique" item is presented. Social groups did not have an associated direction/scale item.

Among others, we found that higher Health and Geography stereotypes respectively predicted higher prioritization for immigration policing protection programs and higher COVID-19 vaccination prioritization. Direction was also predictive. For example, higher Deviance direction (as measured through scales) predicted higher prioritization for inclusion programs. Additional results are presented in the Supplement (e.g., we found some interactive patterns between representativeness and direction, such that direction was more predictive when representativeness was higher).

For additional context, when controlling for the relevant predictor dimensions, Competence stereotypes (either in terms of direction or representativeness) were non-significant across most scenarios. Warmth (again controlling for each scenario's relevant predictor

dimensions) was more variable, ranging from not significant to being the only significant result with effect sizes multiple times those of the predictor dimension (full model information is available in the online repository). These patterns illustrate that various dimensions from the full taxonomy may be more informative than big two dimensions in specific contexts.

Discussion

Study 4 presented additional evidence in favor of the relevance of most of the dimensions incorporated into our taxonomy. While Study 3 provided overarching evidence that information from spontaneous metrics with the full taxonomy improved predictions of general attitudes towards groups, Study 4 further illustrated how a more comprehensive dimensional evaluation can expand the impact of stereotype research. Specifically, dimensions such as Health, Emotion, Deviance, Social Group membership, Geography, and Appearance stereotypes were significant predictors of decision-making scenarios relevant to multiple social issues, even when controlling for the big two of Warmth and Competence. This suggests that these dimensions indeed add information when exploring a representative sample of social groups and not just when studied in isolation for smaller subsets of groups (e.g., Boysen et al., 2006; Neighbors et al., 1989). This pattern of results suggests that the comprehensive taxonomy may prove useful in exploring stereotypes in specific contextual settings (e.g., healthcare, ethical artificial intelligence, immigration policy), either as an initial exploratory step to identify relevant dimensions, or as a more nuanced approach to understand dynamics between multiple dimensions or situational and goal moderators.

Notably, many the taxonomy dimensions mattered both as scales as well as when coded from spontaneous metrics. This not only further highlights the role of spontaneous representativeness beyond Warmth and Competence but can also make measuring a complex

taxonomy simpler. Specifically, for this study we had to use multiple scale items to measure a large number of dimensions, potentially making it too burdensome for participants to rate multiple groups (we had each participant rate only three groups). On the other hand, for the open-ended measures participants need only provide at least one response and the measurement of the dimensions is done afterwards via automated and standardized coding by the researcher. In fact, through word embeddings and spontaneous measures, researchers can capture an "abstract" stereotype about a target (i.e., a numerical representation of all the semantic information contained in the text responses), which summarizes the general perception of the target. This may be useful, for example, for studies on multiple categorization and information integration, to explore how the abstract stereotypes of single groups are integrated into intersectional targets. It may also be useful for modern analysis methods used, e.g., in social neuroscience, such as representational similarity analysis (see also Freeman et al., 2018), which complements analyses of specific dimensions with more holistic approaches.

In the Supplement we present a close replication of this study, for which most (but not all) effects replicate, upholding the conclusion that the extended taxonomy improves predictions of decision making.

General Discussion

The current paper introduced the Spontaneous Stereotype Content Model (SSCM), which describes a comprehensive taxonomy and associated properties of stereotypes that are reported by perceivers through open-ended responses. In a field with a growing number of models with competing proposed dimensions of stereotype content, we sought to provide a unifying taxonomy that allowed for higher dimensionality. In doing so, we found evidence of high prevalence of well-established dimensions such as Warmth and Competence (SCM; Fiske et al.,

2002). However, our results also suggested, in line with recent proposals, that understanding facets of these dimensions (Sociability and Morality for Warmth, Ability and Assertiveness for Competence; see Abele et al., 2016) can yield additional insights, as they can vary independently. Additionally, we found support for dimensions that have been more recently proposed as fundamental, such as socioeconomic Status and political-religious Beliefs (Koch et al., 2016), although these dimensions were less prevalent than Warmth and Competence. Moreover, we found support for the prevalent use of dimensions that have not been explicitly included in general models of stereotype content across social groups. For example, many of the stereotypes referred to Emotions, Appearance, Health, Geography, Deviance, Occupations, and intersectional Social Group associations. Although these dimensions have been studied to differing degrees in relation to individual person perception and specific social group subsets (e.g., based on ethnicity), our model integrates them into a unified taxonomy that accounts for over 80% of the stereotypes spontaneously attributed to a representative sample of salient societal groups. A richer account of the complexity of social reality and perception thus emerged. Notably, we found support for the proposed taxonomy using a variety of methods, including dictionary coding, word embeddings similarity, and participants' self-coding. This taxonomy allows a common language and structure to integrate and evaluate dimensions based on their prevalence and properties.

The SSCM also provides initial insights into some general properties of spontaneous stereotype content. For instance, we found robust evidence for a positivity bias of the Ability dimension and a negativity bias of the Morality dimension, as has been found in other domains (e.g., Mende-Siedlecki et al., 2013; Skowronski & Carlston, 1989). The other Competence-related dimensions (Assertiveness and Status) also tended to be positively valenced, while

Sociability tended to be more neutral. These biases emerge only for specific facets of Warmth and Competence and are not reflected in traditional scale studies (e.g., Fiske et al., 2002). Spontaneous perceptions thus provide an additional way to examine valence in judgments.

Furthermore, the SSCM incorporates information about accessibility of the proposed taxonomy dimensions using methods facilitated by a spontaneous approach (e.g., response times and order). Using both response order and response times metrics we found robust evidence that Warmth's Sociability facet may be less immediately accessible (despite being highly prevalent overall over time, as shown in, e.g., Table 2). Competence's Ability facet tended to be more immediately accessible but becomes less prevalent over time.[9] Future studies can shed additional light on this topic, with potential implications for our understanding of the way in which different dimensions may be "primary." For example, the SCM has long viewed Warmth as primary, and with good reason as this dimension weighs more heavily in general impressions of social groups and has been shown to be important in a variety of contexts. In the SSCM, too, Warmth was consistently one of the most important dimensions, often surpassing Competence in prevalence. However, Ability's shorter response times and earlier responses (compared to Sociability) may suggest that facets of dimensions may be primary in different ways, such as in time-based accessibility and retrieval from memory, at least under some conditions (see Abele et al., 2021). However, there are, certainly, alternative explanations. For example, it is possible that participants were more reluctant to provide Warmth-related words due to social desirability (although in general we did not find a pattern of negative responses being provided at later times across dimensions, which does not support a large role of social desirability). Future research should attempt to further clarify this issue.

[9] However, in Colombia and Spain (Supplement), Ability remained stable over time.

Finally, spontaneous stereotypes also improve the predictive value of stereotypes on general evaluations of social groups. Part of this improvement is explained by a stereotype property that, to our knowledge, has not been studied before in a systematic manner in a general stereotype model (although it has parallels, e.g., in the literature on attitude/stereotype activation and strength, c.f., Krosnick et al., 1993; Higgins, 1996; Stangor & Lange, 1994; see the supplement for additional analyses and discussion). This property, spontaneous representativeness, refers to the prevalence of a specific stereotype dimension for a perceiver's mental representation of a group.

In the current paper we provided evidence that spontaneous representativeness interacts with dimensional direction to predict global evaluations of social targets. For example, our results suggest that people who evaluate Democrats as warm (e.g., on a scale from "cold" to "warm") *and* who spontaneously think about Democrats primarily in terms of their Warmth (vs. people who spontaneously think about other dimensions) hold more positive general attitudes towards Democrats. Many groups are evaluated similarly in terms of Warmth and Competence direction but very differently in terms of the relative representativeness of these dimensions. Doctors and nurses show this pattern on average (as do, e.g., farmers and Christians; Asian and Black people), where Competence is more representative of the former, and Warmth of the latter, despite being scale-rated very similarly in terms of semantic differentials. New hypotheses may be generated, such as that in a context where Competence is more valued, a group with higher Competence direction *and* representativeness will be seen more positively than a target whose high Competence is not as spontaneously representative. This property adds to our ability to predict evaluations of social groups. Future studies should also explore how spontaneous representativeness translates into behavior.

As such, the SSCM has implications for practical questions related to interpersonal, organizational, and societal discrimination and injustice. As reviewed previously, stereotypes predict multiple outcomes, from hiring evaluations to emotions to behavior toward animals and brands. However, the SSCM reveals that the impact of predictive stereotype models may be larger when incorporating spontaneous perceptions. Also, predictions derived from spontaneous measures may differ, arguably toward ecologically validity (c.f., Nicolas, Skinner, & Dickter, 2019), from stereotypes measured through researcher-defined scales. For example, some groups that have been traditionally understood in terms of Warmth and Competence, may be better understood as being perceived through the lens of alternative dimensions. To illustrate, although current general models highlight the primacy of low Warmth in stereotypes of people who are categorized as being homeless, stereotypes about Emotions ("sad", "desperate") were much more prevalent for this social group. Based on dominant models, interventions on behalf of this group to change perceptions could be leaving out relevant evaluations, potentially resulting in suboptimal initiatives and policies. For example, people who are homeless were rated second highest in priority for emotional and psychological counseling programs (Study 4), in line with their high prevalence of Emotions stereotypes, a decision-making result that may not be predicted from lower-dimensional stereotype models focusing on Warmth and Competence. Similarly, some groups are associated with extremely representative dimensions, which may end up overcoming evaluations and behaviors toward them. For example, politicians are average-high in terms of their Competence (direction), but it is their low Warmth that dominates their spontaneous stereotypes. Therefore, our proposed expanded taxonomy as well as spontaneous approach could be incorporated into the design of interventions and policies relevant to discrimination and social inequality.

We note that spontaneous representativeness patterns often differed from those of time- or response order- based measures of accessibility (c.f., Higgins et al., 1982), guiding our decision for treating it as a separate variable. For example, in Study 1, Ability was more accessible (in terms of response order), but Sociability was more prevalent. This is also descriptively illustrated for specific groups: Warmth was more representative of stereotypes about White and Catholic people, but Competence was more accessible; Competence was more representative of stereotypes about Hispanic and Atheist people, but Warmth was more accessible. In conjunction, these patterns suggest that time/order-based accessibility and prevalence-based representativeness are distinct properties that (perhaps independently) relate to the underlying associative strength between the target and the evaluative content (c.f., Higgins, 1996; see Supplement for additional information). But admittedly, additional research will be needed to further establish the relationship between these variables.

In general, our findings suggest that future research using a spontaneous content approach may obtain a more nuanced perspective of social group perceptions by complementing traditional scale measures with open-ended responses. Spontaneous responses may reveal unexpected content that would not have been obtained through researcher-determined scales (c.f., Nicolas, Skinner, & Dickter, 2019). Additionally, spontaneous responses may be more ecologically valid (vs. traditional scales), as they more closely reflect real-world person perception, where evaluative dimensions are explicitly provided for perceivers to make sense of the social world. For these and other reasons discussed (e.g., the ability to measure constructs such as spontaneous representativeness), spontaneous information may also improve social cognition models' ability to predict socially relevant outcomes.

Limitations and Future Directions

The current paper introduces the Spontaneous Stereotype Content Model as an initial iteration of a taxonomic structure of spontaneous stereotypes along with a set of associated properties and predictive value. As such, the studies have a number of limitations that need to be considered, and which pave the way for future studies of robustness, generalizability, and model refinement.

Some limitations have been discussed in previous sections. Others include the use of mostly online samples, which may limit generalizability or the potential for social desirability effects for some dimensions (although we found no general evidence of it from our valence measures). For these reasons, follow-up research could control for efforts to appear non-prejudiced, additional cross-cultural and cross-setting studies, and manipulations of instructions.

Given disputes about the nature and number of dimensions, we also expect future studies may suggest alternative taxonomic organizations. While, as simplified views, no specific taxonomy will capture the true structure of semantic content, we believe that the utility of different specifications can be empirically evaluated. We based the current taxonomy on decades of research from general stereotype content models, as well as data-driven dimensions labeled based on more isolated, but by no means obscure, stereotype contents. Nonetheless, calls for different configurations, such as the recently proposed subdivision of Competence into Ability and Assertiveness (Abele et al., 2016), will predictably arise, and may be incorporated into revised taxonomies if shown to be useful.

We note that the SSCM does not currently aim to provide universal principles of taxonomic hierarchies or stereotype properties. Instead, it provides a useful initial descriptive model to organize and understand patterns that could vary across cultures and time, as a function of which groups are salient in a society and their historical circumstances. Following on the

tradition of the SCM (Fiske et al., 2002), cross-cultural variation in aspects such as taxonomic hierarchy or stereotype representativeness may be understood in the light of societal-level variables such as income inequality, peace, conflict, and diversity (Bai et al., 2020; Durante et al., 2013; 2017). The SSCM complements theory-driven efforts to integrate competing conceptual frameworks of social evaluation (Abele et al., 2021).

The methods and metrics introduced here also provide an initial analytical approach for the study of spontaneous social cognition content, but they have weaknesses and are certain to evolve. The dictionaries may be revised to include more unaccounted-for words (although we find evidence that these < 15% remaining responses tended to be quite idiosyncratic: they tended to be different across studies and had few repeats). Word embedding models are also improving at a fast pace, as advances in natural language processing continue. Improved translation and multi-sense disambiguation may also be continually incorporated into the methods repertoire. Other machine-learning methods not used here, such as topic modeling (Blei et al., 2003), are also useful for finding patterns and reducing dimensionality in text data (e.g., Nicolas, Bai, & Fiske, 2019).

We envision multiple future directions for research building upon these methods and descriptive model. For example, spontaneous stereotypes may be particularly useful in the study of intersectional or multiply-categorizable targets, as they may reveal emergent stereotypes that are distinct from the algebraic combinations that are allowed by scale-metrics (e.g., averaging of the constituent stereotypes), and which may fall along dimensions that current models do not cover (see Kunda et al., 1990; Nicolas, Skinner, & Dickter, 2019). This topic is increasingly important as the world grows more diverse (e.g., Phinney & Alipuria, 2006). Our taxonomy also facilitates studies into stereotype compensation, which has mostly been studied around Warmth

and Competence, but may operate such that compensation occurs towards alternative dimensions (perhaps less valenced or more structural; c.f., Nicolas et al., in press). Finally, the combination of linguistic, machine learning, and social-cognitive nature of the methods and empirical findings makes this model particularly well-suited to increase social psychological insights into evermore relevant topics of fairness in machine learning (c.f., Caliskan et al., 2017). Current approaches to examine these linguistic biases rely on comparisons to implicit measures or direction-based stereotypes alone. However, machine learning models are trained on spontaneous language used in sources such as books, news, and the internet. As such, a psychological model of spontaneous stereotypes, which includes information about properties such as representativeness, will provide new ways forward to understand, and potentially address, issues relevant to bias in Artificial Intelligence.

Conclusion

The social world is complex, as is social perception. We introduce the Spontaneous Stereotype Content Model as an initial and comprehensive descriptive model to understand the structure, properties, and predictive value of spontaneous stereotypes. Drawing from advances in natural language processing to facilitate the quantitative analysis of text responses, the SSCM is integrative and generative, opening the way for a deeper understanding of the ways in which multiple dimensions of content organize people's mental representation of and behavior toward their society's members.

References

Abele, A. E., & Bruckmüller, S. (2011). The bigger one of the "Big Two"? Preferential processing of communal information. *Journal of Experimental Social Psychology, 47*(5), 935-948. https://doi.org/10.1016/j.jesp.2011.03.028

Abele, A. E., Ellemers, N., Fiske, S. T., Koch, A., & Yzerbyt, V. (2021). Navigating the social world: Toward an integrated framework for evaluating self, individuals, and groups. *Psychological Review, 128*(2), 290–314. https://doi.org/10.1037/rev0000262

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

Allport, G. W. (1979). *The nature of prejudice*. Reading, MA: Addison-Wesley. (Original work published 1954)

Ajzen, I., & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin, 84*(5), 888–918. https://doi.org/10.1037/0033-2909.84.5.888

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology, 41*(3), 258–290. https://doi.org/10.1037/h0055756

Bai, X., Ramos, M. R., & Fiske, S. T. (2020). As diversity increases, people paradoxically perceive social groups as more similar. *Proceedings of the National Academy of Sciences*, *117*(23), 12741-12749. https://doi.org/10.1073/pnas.2000333117

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003) Latent dirichlet allocation. *Journal of machine learning* research, 3, 993-1022. Available at: http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Bodenhausen, G. V, Kang, S. K., & Peery, D. (2012). Social Categorization and the Perception of Social Groups. *The SAGE Handbook of Social Cognition*, 318–336. https://doi.org/10.4135/9781446247631

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135-146. https://doi.org/10.1162/tacl_a_00051

Boysen, G. A., Vogel, D. L., Madon, S., & Wester, S. R. (2006). Mental health stereotypes about gay men. *Sex Roles: A Journal of Research, 54*(1-2), 69–82. https://doi.org/10.1007/s11199-006-8870-0

Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology, 41*(2), 135-143. https://doi.org/10.1002/ejsp.744

Brambilla, M., Sacchi, S., Rusconi, P., & Goodwin, G. P. (2021). The primacy of morality in impression development: Theory, research, and future directions. In Advances in Experimental Social Psychology (Vol. 64, pp. 187-262). Academic Press.

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.

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., & Charrad, M. M. (2014). Package 'nbclust'. *Journal of Statistical Software, 61*, 1-36.

Cuddy, A. J. C., 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

Cuddy, A. J., Glick, P., & Beninger, A. (2011). The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in Organizational Behavior, 31*, 73-98. https://doi.org/10.1016/j.riob.2011.10.004

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5–18. https://doi.org/10.1037/0022-3514.56.1.5

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

Durante, F., Fiske, S. T., Kervyn, N., Cuddy, A. J., Akande, A., Adetoun, B. E., ... & Barlow, F. K. (2013). Nations' income inequality predicts ambivalence in stereotype content: How societies mind the gap. *British Journal of Social Psychology*, *52*(4), 726-746. https://doi.org/10.1111/bjso.12005

Durante, F., Fiske, S. T., Gelfand, M. J., Crippa, F., Suttora, C., Stillwell, A., ... & Björklund, F. (2017). Ambivalent stereotypes link to peace, conflict, and inequality across 38 nations. *Proceedings of the National Academy of Sciences*, *114*(4), 669-674. https://doi.org/10.1073/pnas.1611874114

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Fazio, R. H., Williams, C. J., & Powell, M. C. (2000). Measuring associative strength: Category-item associations and their activation from memory. *Political Psychology, 21*(1), 7-25. https://doi.org/10.1111/0162-895X.00175

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50*(2), 229. https://doi.org/10.1037/0022-3514.50.2.229

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

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, *38*(6), 889. https://doi.org/10.1037/0022-3514.38.6.889

Fiske, S. T., & Cox, M. G. (1979). Person concepts: The effect of target familiarity and descriptive purpose on the process of describing others. *Journal of Personality, 47*(1), 136–161. https://doi.org/10.1111/j.1467-6494.1979.tb00619.x

Fiske, S. T., & Tablante, C. B. (2015). Stereotyping: Processes and content. In E. Borgida & J. A. Bargh (Eds.), *APA Handbook of Personality and Social Psychology, Volume 1: Attitudes and Social Cognition*. Washington, DC: American Psychological Association.

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

Fiske, S. T., Dupree, C. H., Nicolas, G., & Swencionis, J. K. (2016). Status, power, and intergroup relations: The personal is the societal. *Current Opinion in Psychology*, *11*, 44–48. https://doi.org/10.1016/j.copsyc.2016.05.012

Freeman, J. B., Stolier, R. M., Brooks, J. A., & Stillerman, B. S. (2018). The neural representational geometry of social perception. *Current Opinion in Psychology, 24*, 83–91. https://doi.org/10.1016/j.copsyc.2018.10.003

Friedman, J., Hastie, &., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software, 33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01.

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

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–68. https://doi.org/10.1037/a0034726

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029. https://doi.org/10.1037/a0015141

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493-498. https://doi.org/10.1111/2041-210X.12504

Hagendoorn, L. & Hraba, J. (1989). Foreign, different, deviant, seclusive and working class: Anchors to an ethnic hierarchy in the Netherlands. *Ethnic and Racial Studies, 12*(4), 441-468. https://doi.org/10.1080/01419870.1989.9993647

Higgins, E. T. (1996). *Knowledge activation: Accessibility, applicability, and salience.* In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (p. 133–168). The Guilford Press.

Higgins, E. T., King, G. A., & Mavin, G. H. (1982). Individual construct accessibility and subjective impressions and recall. *Journal of Personality and Social Psychology, 43*(1), 35–47. https://doi.org/10.1037/0022-3514.43.1.35

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

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. https://doi.org/10.1037/pspa0000046

Koch, A., Nicolas, G., Imhoff, R., Unkelbach, C., Terache, J., Carrier, A., Yzerbyt, V., & Fiske, S. (2020). Groups' warmth is a personal matter: Understanding consensus on stereotype

dimensions reconciles adversarial models of social evaluation. *Journal of Experimental Social Psychology*, *89*. https://doi.org/10.1016/j.jesp.2020.103995

Krosnick, J. A., Boninger, D. S., Chuang, Y. C., Berent, M. K., & Carnot, C. G. (1993). Attitude strength: One construct or many related constructs? *Journal of Personality and Social Psychology, 65*(6), 1132–1151. https://doi.org/10.1037/0022-3514.65.6.1132

Kunda, Z., Miller, D. T., & Claire, T. (1990). Combining social concepts: The role of causal reasoning. *Cognitive Science, 14*(4), 551–577. https://doi.org/10.1207/s15516709cog1404_3

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, 82(13). https://doi.org/10.18637/jss.v082.i13

Leach, C., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality vs. competence and sociability in the evaluation of in-groups. *Journal of Personality and Social Psychology, 93*, 234-249. https://doi.org/10.1037/0022-3514.93.2.234

Lee, S. J., Wong, N.-W. A., & Alvarez, A. N. (2009). The model minority and the perpetual foreigner: Stereotypes of Asian Americans. In N. Tewari & A. N. Alvarez (Eds.), *Asian American psychology: Current perspectives* (pp. 69-84). New York, NY, US: Routledge/Taylor & Francis Group.

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software, 69*(1), 1-33. https://doi.org/10.18637/jss.v069.i01

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1171. https://doi.org/10.3389/fpsyg.2015.01171

Luccioni, A. S., & Viviano, J. D. (2021). What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus. *arXiv*. preprint arXiv:2105.02732.

Malone, C., & Fiske, S. T. (2013). *The human brand: How we relate to people, products, and companies.* San Francisco, CA: Wiley/ Jossey Bass.

Martinez, J. E., Funk, F., & Todorov, A. (2020). Quantifying idiosyncratic and shared contributions to judgment. *Behavior Research Methods, 52*(4), 1428–1444. https://doi.org/10.3758/s13428-019-01323-0

Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013). Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *Journal of Neuroscience, 33*, 19406-19415. https://doi.org/10.1523/jneurosci.2334-13.2013

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

Neighbors, H.W., Jackson, J.S., Campbell, L., & Williams, D. (1989). The influence of racial factors on psychiatric diagnosis: A review and suggestions for research. *Community Mental Health Journal, 25*, 301–311. https://doi.org/10.1007/BF00755677

Nicolas, G., & Skinner, A. (2012). "That's so gay!" Priming the general negative usage of the word Gay increases implicit anti-gay bias. *The Journal of Social Psychology, 152*(5), 654–658. https://doi.org/10.1080/00224545.2012.661803

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). California: Elsevier Science. https://doi.org/10.1016/B978-0-08-101107-2.00025-7

Nicolas, G., Bai, X., & Fiske, S. T. (2019). Exploring research-methods blogs in Psychology: Who posts what about whom, and with what effect? *Perspectives on Psychological Science, 14*(4), 691-704. https://doi.org/10.1177/1745691619835216

Nicolas, G., Bai, X., & Fiske, S. T. (2021). Comprehensive stereotype content dictionaries Using a semi-automated method. European Journal of Social Psychology, 51(1), 178-196. https://doi.org/10.1002/ejsp.2724

Nicolas, G., de la Fuente, M., Fiske, S. T. (2017). Mind the overlap in multiple categorization: A review of crossed categorization, intersectionality, and multiracial perception. *Group Processes & Intergroup Relations, 20*(5): 621–631. https://doi.org/10.1177/1368430217708862

Nicolas, G., Skinner, A. L., & Dickter, C. L. (2019). Other than the sum: Hispanic and Middle Eastern categorizations of Black–White mixed-race faces. *Social Psychological and Personality Science, 10*(4), 532-541. https://doi.org/10.1177/1948550618769591

Nicolas, G., Fiske, S. T., Terache, J., Carrier, A., & Yzerbyt, V., Koch, A., Imhoff, R., & Unkelbach, C. (in press). Relational versus structural goals prioritize different social information. *Journal of Personality & Social Psychology*.

Niemann, Y. F., Jennings, L., Rozelle, R. M., Baxter, J. C., & Sullivan, E. (1994). Use of free responses and cluster analysis to determine stereotypes of eight groups. *Personality and Social Psychology Bulletin, 20(4)*, 379-390. doi:10.1177/0146167294204005

Oh, D., Buck, E. A., Todorov, A. (2019) Revealing hidden gender biases in competence impressions from faces. *Psychological Science, 30*(1), 65–79. https://doi.org/10.1177/0956797618813092

Olivier, J., May, W. L., & Bell, M. L. (2017). Relative effect sizes for measures of risk. *Communications in Statistics-Theory and Methods*, *46*(14), 6774-6781. https://doi.org/10.1080/03610926.2015.1134575

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957) *The measurement of meaning*. Urbana IL: University of Illinois Press.

Park, B. (1986). A method for studying the development of impressions of real people. *Journal of Personality and Social Psychology, 51*(5), 907. https://doi.org/10.1037/0022-3514.51.5.907

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).

Phinney, J. S., & Alipuria, L. L. (2006). *Multiple social categorization and identity among multiracial, multiethnic, and multicultural individuals: Processes and implications*. In R. J. Crisp & M. Hewstone (Eds.), *Multiple social categorization: Processes, models and applications* (p. 211–238). Psychology Press.

Quinn, K. A., Macrae, C. N., & Bodenhausen, G. V. (2003). *Stereotyping and impression formation: How categorical thinking shapes person perception*. In M. A. Hogg & J. Cooper (Eds.), Sage handbook of social psychology (pp. 87-109). Thousand Oaks, CA: Sage Publications.

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331

Saucier, G., & Goldberg, L. R. (1998). What is beyond the Big Five? *Journal of Personality*, *66*(4), 495–524. https://doi.org/10.1111/1467-6494.00022

Schug, J., Alt, N. P., & Klauer, K. C. (2015). Gendered race prototypes: Evidence for the non-prototypicality of Asian men and Black women. *Journal of Experimental Social Psychology, 56*, 121-125. https://doi.org/10.1016/j.jesp.2014.09.012.

Sevillano, V., & Fiske, S. T. (2016). Warmth and competence in animals. Journal of Applied *Social Psychology, 46*(5), 276–293. https://doi.org/10.1111/jasp.12361

Skinner, A. L., & Nicolas, G. (2015). Looking Black or looking back? Using phenotype and ancestry to make racial categorizations. *Journal of Experimental Social Psychology, 57*, 55–63. https://doi.org/10.1016/j.jesp.2014.11.011

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105*(1), 131-142. https://doi.org/10.1037/0033-2909.105.1.131

Stangor, C., & Lange, J. E. (1994). *Mental representations of social groups: Advances in understanding stereotypes and stereotyping*. In M. P. Zanna (Ed.), *Advances in experimental social psychology*, Vol. 26 (p. 357–416). Academic Press. https://doi.org/10.1016/S0065-2601(08)60157-4

Todorov, A. (2012). The social perception of faces. *The SAGE Handbook of Social Cognition*, 96–114.

Uleman, J. S. (1987). Consciousness and control: The case of spontaneous trait inferences. *Personality and Social Psychology Bulletin, 13*(3), 337–354. https://doi.org/10.1177/0146167287133004

Wallace, D. S., Paulson, R. M., Lord, C. G., & Bond Jr, C. F. (2005). Which behaviors do attitudes predict? Meta-analyzing the effects of social pressure and perceived difficulty. *Review of General Psychology, 9*(3), 214-227. https://doi.org/10.1037/1089-2680.9.3.214

Wojciszke, B. (1994). Multiple meanings of behavior: Construing actions in terms of Competence or Morality. *Journal of Personality and Social Psychology*, *67*(2), 222–232. https://doi.org/10.1037/0022-3514.67.2.222

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin, 24*(12), 1251–1263. https://doi.org/10.1177/01461672982412001