Article

Social Psychological and Personality Science
1-13

DOI: 10.1177/1948550616671998
spps.sagepub.com
SAGE

Real Men Don't Say "Cute": Using Automatic Language Analysis to Isolate Inaccurate Aspects of Stereotypes

Jordan Carpenter[1], Daniel Preotiuc-Pietro[1,2], Lucie Flekova[3], Salvatore Giorgi[1], Courtney Hagan[1], Margaret L. Kern[4], Anneke E. K. Buffone[1], Lyle Ungar[2], and Martin E. P. Seligman[1]

Abstract

People associate certain behaviors with certain social groups. These stereotypical beliefs consist of both accurate and inaccurate associations. Using large-scale, data-driven methods with social media as a context, we isolate stereotypes by using verbal expression. Across four social categories—gender, age, education level, and political orientation—we identify words and phrases that lead people to incorrectly guess the social category of the writer. Although raters often correctly categorize authors, they overestimate the importance of some stereotype-congruent signal. Findings suggest that data-driven approaches might be a valuable and ecologically valid tool for identifying even subtle aspects of stereotypes and highlighting the facets that are exaggerated or misapplied.

Keywords

big data, stereotypes, language analysis, person perception, social media

Social group is reflected in people's behaviors: Koreans are more likely than Afghans to speak Korean, social psychologists are more likely than airline pilots to write social psychology papers. The tension between the existence of stereotypes about groups and the existence of real group differences has long been a controversial topic in psychological research (e.g., Dovidio, Brigham, Johnson, & Gaertner, 1996; Eagly, 1995). In a series of studies, we take advantage of big data language analysis techniques to (1) quantitatively separate the accurate and inaccurate content of a variety of stereotypes and (2) directly assess the relation between perceived and actual group-based differences using the same behaviors. We examine four social groupings: gender, age, education level, and political orientation.

Stereotypes and Accuracy

A stereotype is an individual's set of beliefs and associations about a social group (Allport, 1954). Meta-analyses have demonstrated that stereotypes about demographic groups are often accurate in that people's perceptions of groups often correlate reasonably well with external criteria (Jussim, Crawford, & Rubinstein, 2015). However, stereotypes are dynamic and complex (Jussim, Cain, Crawford, Harber, & Cohen, 2009), and thus the content of stereotypes includes both accurate and inaccurate parts.

There has been conflicting research on whether stereotypes are generally inaccurate (McCauley & Stitt, 1978; Prothro & Melikian, 1955), the cognitive mechanisms behind inaccurate stereotypes (e.g., McCauley, 1995), and the types of content most likely to be inaccurate (Diekman, Eagly, & Kulesa, 2002). It is also difficult to define and measure the accuracy of stereotypes' content (Stangor, 1995).

Despite these difficulties, researchers often adopt a working definition of stereotype accuracy that relates to how well an individual's perception matches the actual traits of that group (Jussim & Zanna, 2005). This is often thought of in terms of differences in central tendency (e.g., members of Group X possess trait Z more than members of Group Y). However, observers can also be incorrect about the size of the variability within

[1] Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA
[2] Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA
[3] Ubiquitous Knowledge Processing Lab (UKP-TUDA), Department of Computer Science, Technische Universität Darmstadt, Darmstadt, Germany
[4] Melbourne Graduate School of Education, The University of Melbourne, Victoria, Australia

Corresponding Author:
Jordan Carpenter, Department of Psychology, University of Pennsylvania, Positive Psychology Center, 3701 Market Street, Philadelphia, PA 19104, USA.
Email: jordac@sas.upenn.edu

groups around the central tendency of the actual trait (Ryan, 2003). That is, the ''accuracy'' of a stereotypical assessment is a function of both the binary endorsement or denial of the existence of an accurate difference between two groups, and a realistic estimate of the level to which the trait varies within each group (e.g., Judd & Park, 1993).

Therefore, we aim to compare the extent to which a behavior *is represented* by a group to the extent that the exact same behavior *is believed to be represented*. One way of performing this comparison is to expose participants to real-world behaviors and ask them to use the behaviors to guess to what social group the actor belongs. Inaccurate stereotypes would be the behaviors associated with a higher proportion of raters making a particular miscategorization (i.e., believing a member of Group X is actually in Group Y). Importantly, this technique can highlight stereotypical associations that are exaggerated as well as those that are entirely incorrect. If words that people think are associated with Group X are indeed used by Group Y, then people make group-based judgments using incorrect theories. However, if inaccurate stereotypes about Group X actually are used more often by Group X, then people's judgments are based on beliefs that are exaggerated, overly salient, or misapplied.

This comparison must be performed across many people to account for the range of variance, and the behaviors must be nuanced and varied enough to capture subtle differences between groups. We thus require a large amount of data and an analytic approach that can capture fine effects. Also, we must isolate a single channel of information so that we can be sure people are basing their judgments solely on the behaviors in question and not on other cues. To address these challenges, we use language on social media for our analyses.

Language Use as a Representation of Group Differences

In psychology, group tendencies in language have most commonly been assessed using the linguistic inquiry and word count (LIWC), a set of theory-driven dictionaries that categorize words in psychologically meaningful ways (Pennebaker & Francis, 2001). LIWC can be used to reveal the ways that groups differ from one another psychologically. It has revealed group- and trait-based differences in spontaneous language use (Newman, Groom, Handelman, & Pennebaker, 2008; Sylwester & Purver, 2015). Notably, some of the most profound effects involve function words such as prepositions, pronouns, and articles, which people usually do not pay attention to or consciously control (Chung & Pennebaker, 2007). There are reliable gender differences: Women use first-person, singular pronouns and talk about social topics, while men use more articles and prepositions (Newman et al., 2008). Language also changes as a function of age: Authors express more positive affect and use the future tense more often as they age (). Similar group-level differences have also specifically been found in a specifically online context (Argamon, Koppel, Pennebaker, & Schler, 2007). However, although some research has examined gender stereotypes in language, suggesting that, for example, people believe that men swear more often than women (Haas, 1979), it is not known whether people have generally accurate lay theories about groups' linguistic differences.

With the recent availability of large-scale text data, new data-driven techniques demonstrate group-based differences in written expression across wide populations (e.g., Schwartz, Eichstaedt, Kern, et al., 2013). For instance, users who score themselves low in agreeableness are more likely to use swear words (Park et al., 2015), and self-identified teenagers are more likely than adults to use the word ''homework'' (Kern et al., 2014).

In computer science, there has been growing interest in automatically identifying author traits from their text using machine learning techniques, predicting gender (Burger, Henderson, King, & Zarrella, 2011[AQ1]), age (Rao et al., 2011; Sap et al., 2014), political orientation (Pennacchiotti & Popescu, 2011), and income (Preoţiuc-Pietro, Volkova, Lampos, Bachrach, & Aletras, 2015). Other studies have predicted language associations with *perceived* author traits for demographic characteristics such as gender (Nguyen et al., 2014), age (Nguyen, Gravel, Trieschnigg, & Meder, 2013), and other features. In these cases, researchers employed raters to categorize author profiles according to their perceptions of the author's traits. Thus, the language processing literature has been successfully used to predict both real and perceived behavior differences between groups. However, these methods are not centered on providing insight into the content or the accuracy of people's stereotypes.

The Current Studies

We use natural language processing techniques to automatically determine language cues associated with perceived group differences. We use a method introduced by Schwartz, Eichstaedt, Dziurzynski, et al. (2013)[AQ2], which consists of correlating words and groups of words with a specific outcome. In this case, the outcome is social group categorizations among naive observers. Participants will attempt to categorize authors based solely on the content of social media posts. This strategy directly tests the accuracy of stereotypes: Unless the posts contain direct and explicit self-identification, nonarbitrary observer classification *must* be based on theories of relative differences between groups (i.e., stereotypes).

We have three goals. First, we provide insight into the content of stereotypes in a manner that avoids directly asking participants to list their beliefs about social groups because people cannot or will not report all of their stereotypic associations (e.g., Greenwald & Banaji, 1995). Second, we isolate the language associated with rater misperception and in doing so highlight the aspects of stereotypes that are inaccurate. Third, we compare inaccurate stereotypes to actual group differences to determine whether people's stereotypes are entirely wrong or whether they are exaggerated.

Altogether, our goal is to illustrate the content of stereotype inaccuracy in a detailed and nuanced way. We reveal not only

what aspects of people's stereotypes are wrong but also *how* they are wrong.

Method

We present four studies of social classifications: gender, age, education, and political orientation. Linguistic and additional ''ground truth'' information was collected from online users, and a separate group of raters guessed author gender, age, education, or political orientation based on the language samples provided.

Data

Our materials are public posts from Twitter. Twitter is a popular (320 million active users, Twitter.com, 2015) platform allowing people to broadcast short messages (up to 140 characters). Twitter is an ideal source for this study, as it contains large volumes of spontaneous language use. Additionally, using stereotypes about group differences in linguistic expression is the only way to make categorical assessments of someone using online, word-based communication; this task is therefore externally valid and likely not to be confusing to participants.

In Study 1, ground truth labels of gender were obtained from Burger and colleagues (2011), who mapped Twitter accounts to their self-identified gender as mentioned in other user public profiles linked to their Twitter account. From a data set containing 67,337 users, we randomly created a gender-balanced sample of 3,000 authors.[1]

In Study 2, we used 796 Twitter users who self-reported their age in an online survey. The median age for these users was 23 years old. We split authors by this median, resulting in 400 authors younger than 24 and 396 authors 24 or older. Although forcing a continuous variable into a categorical split can be problematic (e.g., MacCallum, Zhang, Preacher, & Rucker, 2002), stereotypes are typically thought of in reference to social *groups*, and the median provides one way to create two groupings. We also had self-reported age in years for all authors ($M = 28.85$, $SD = 11.48$, range = 13–71). In our analysis, we considered both continuous and dichotomous age.

In Study 3, we used the data set introduced in Preotiuc-Pietro, Lampos, and Aletras (2015) to categorize authors by education. The self-identified occupations of 5,191 Twitter users were labeled according to the U.K. Standard Occupational Classification, an occupational taxonomy that groups occupations based on skill level (Elias et al., 2010). These users were then mapped to four broader groups, applicable to the United States, based on education requirements for each type of job. For the present study, we used 1,000 authors in three classes: ''advanced degree'' (334 users, master's or higher), ''college degree'' (333 users, bachelors and associates), and ''no college degree'' (333 users).[2]

In Study 4, we selected popular political figures unambiguously associated with U.S. liberal politics (Cory Booker, Bernie Sanders, Joe Biden, and John Kerry) or U.S. conservative politics (Ted Cruz, Marco Rubio, Rand Paul, and Ben Carson) as of August 2015 when the data were collected. We selected Twitter users who followed all four liberal political figures and none of the conservative figures for the liberal group or all of the conservative figures and no liberal figures for the conservative group. Our set of authors contained 1,250 conservatives and 1,250 liberals.

The four groups of authors (for gender, age, education, and political orientation) do not overlap and thus were each tested and analyzed separately. Sample sizes for Studies 2 and 3 were smaller due to availability, but they provide a sufficient amount of language data to produce a number of meaningful language correlates (Eichstaedt et al., 2016).

For each author, we randomly selected 100 tweets from the same 6-month interval using the Twitter Search API, Version 2.2.5. We filtered out non-English tweets using an automatic method (Lui & Baldwin, 2012) and eliminated duplicate tweets (e.g., created automatically by apps). URLs and @ mentions were removed, as they may contain sensitive information.

Rating Procedure

We set up four crowdsourcing tasks using Amazon Mechanical Turk for gender, age, education level, and political orientation. Each rater was presented with a random sample of 20 tweets from a single author's full 100-tweet battery. Figure 1 illustrates an example task from Study 4 (political orientation), showing what participants saw, with nine separate raters guessing each author's group membership. Using only these tweets as cues, raters were asked to guess the group of an author using a forced choice setup (e.g., binary male/female), using only these tweets as cues.[3] Nine raters made it, so that each of the 100 tweets was mathematically likely considered by at least one rater. Raters were encouraged to use any available cues to make their categorizations, and they were instructed to trust their instincts when not sure of a guess.[4]

Raters received a small compensation (US$0.02) for each rating they provided and could repeat the task as many times as they wished but never for the same author. They were also presented with a bonus (US$0.25) upon completing the first 20 ratings. Overall, 2,741 raters completed the task an average of 20.88 times (Study 1: 1,083 raters, 62% female; Study 2: 728 raters, mean age = 33.37 years; Study 3: 481 raters, 45% no college degree, 41% college degree, 4% advanced degree; Study 4: 943 raters, 59% liberal; some raters participated in two or more studies).

Linguistic Analyses

We first automatically extracted the relative frequency of all single words and phrases (sequences of two and three consecutive words) across the 100 tweet batteries of all authors (for methodology, see Kern et al., 2016). Using these features, we isolated the linguistic patterns associated with our constructs of interest: general and inaccurate stereotypes.

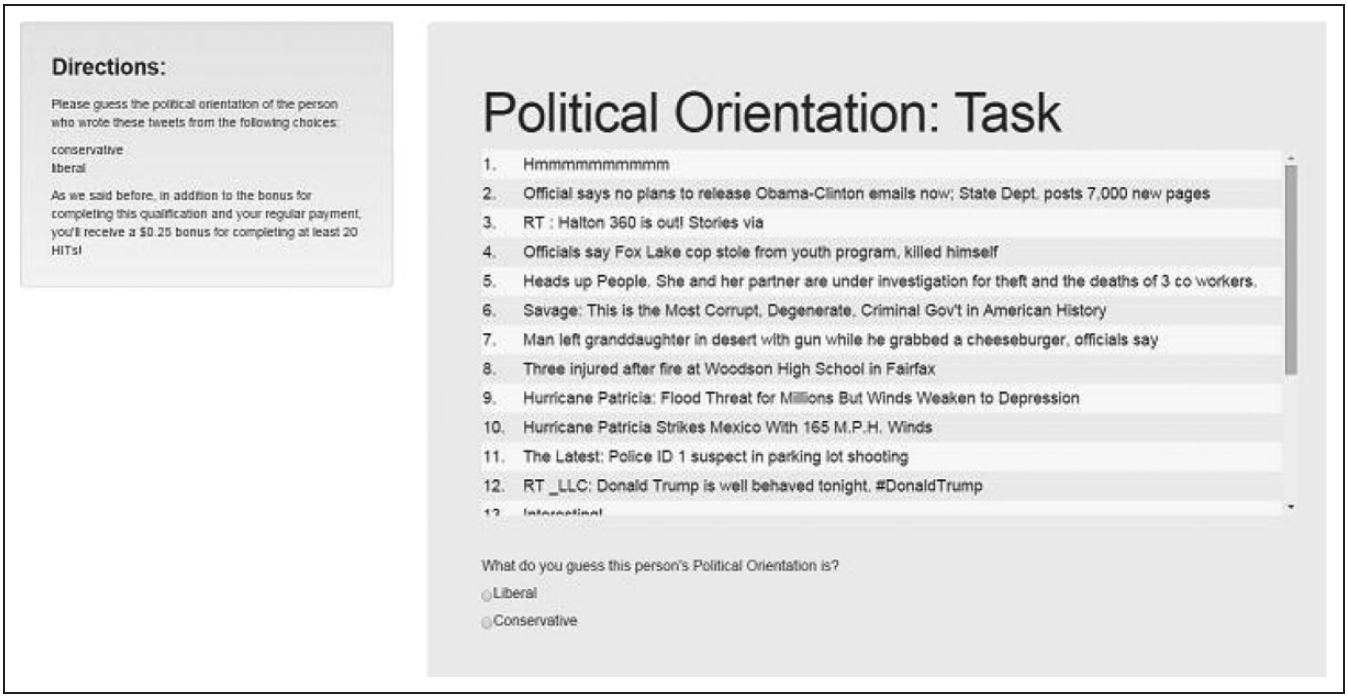

Figure 1. Sample categorization task.

Inaccurate stereotypes were assessed by correlating the words/phrases with the proportion of total raters who *believed* the author to be within each category among users *actually* within each ground truth category. For instance, inaccurate stereotypes for women were words/phrases *written by men* that led more raters to assess the author to be a woman.

Once we isolated the inaccurate stereotypes, we quantitatively examined their actual group associations by correlating them with overall *perceived* group membership (e.g., percentage of raters who thought the author was a woman) and with *actual* group membership (e.g., actually being a woman). We then examined the *z*-scored difference between the resulting correlation coefficients (Lee & Preacher, 2013; Steiger, 1980). This analysis tests the extent to which inaccurate stereotypes might be more strongly associated with perceptions than with reality.

Due to the large number of calculations, presentation of exact *p* values for each result was unmanageably unwieldy. We used the Simes *p* correction (Simes, 1986) and use $p < .001$ as a heuristic for indicating meaningful correlations.

Results

Study 1: Gender

Men were more likely to be categorized as women if they expressed positive emotion (''cute,'' ''wonderful,'' ''beautiful'') or used second-person pronouns. Women were more likely to be categorized as men if they talked about news (''war,'' ''news,'' ''media'') or technology (''google,'' ''mobile,'' ''tech''). Figure 2 presents ''overall stereotypes'' (top), which indicates the words that result in being categorized as female or male, regardless of the ground truth, and ''incorrect stereotypes'' (bottom), or the words associated with miscategorizing men as women or women as men. Although these words may not all contribute equally to participants' mistakes (e.g., pronouns and articles may simply not capture readers' explicit attention as much as nouns and verbs), as a whole they distinguish authors who were incorrectly categorized.

In general, raters were fairly accurate in assessing the gender of each author: 75.7% of categorizations were correct ($\chi^2 = 6357.47$, $p < .001$). Women were correctly identified 78.3% of the time and men 72.8% of the time. Still, there was substantial variation in the ratings of authors, with correct categorizations ranging from 0 (0%) to 9 (100%).

Tables 1 and 2 summarize the correlations between perceived gender and real gender for the 10 most misleading words and phrases for men and women. Typically, perceived and actual gender went in the same direction: Language features associated with perceived maleness were also associated with ground truth maleness. The same pattern was found for femaleness. However, across most words for both men and women, the association with perceived gender was significantly larger than the association with actual gender. For instance, raters' belief that men are more likely to write the word ''research'' was correct, but the diagnostic utility of the word research in determining gender was exaggerated. Thus, these cues do not represent stereotypical associations that are entirely inaccurate but rather are overestimated.

Study 2: Age

As illustrated in Figure 3, younger authors were exaggeratedly believed to be self-referential and casual, while

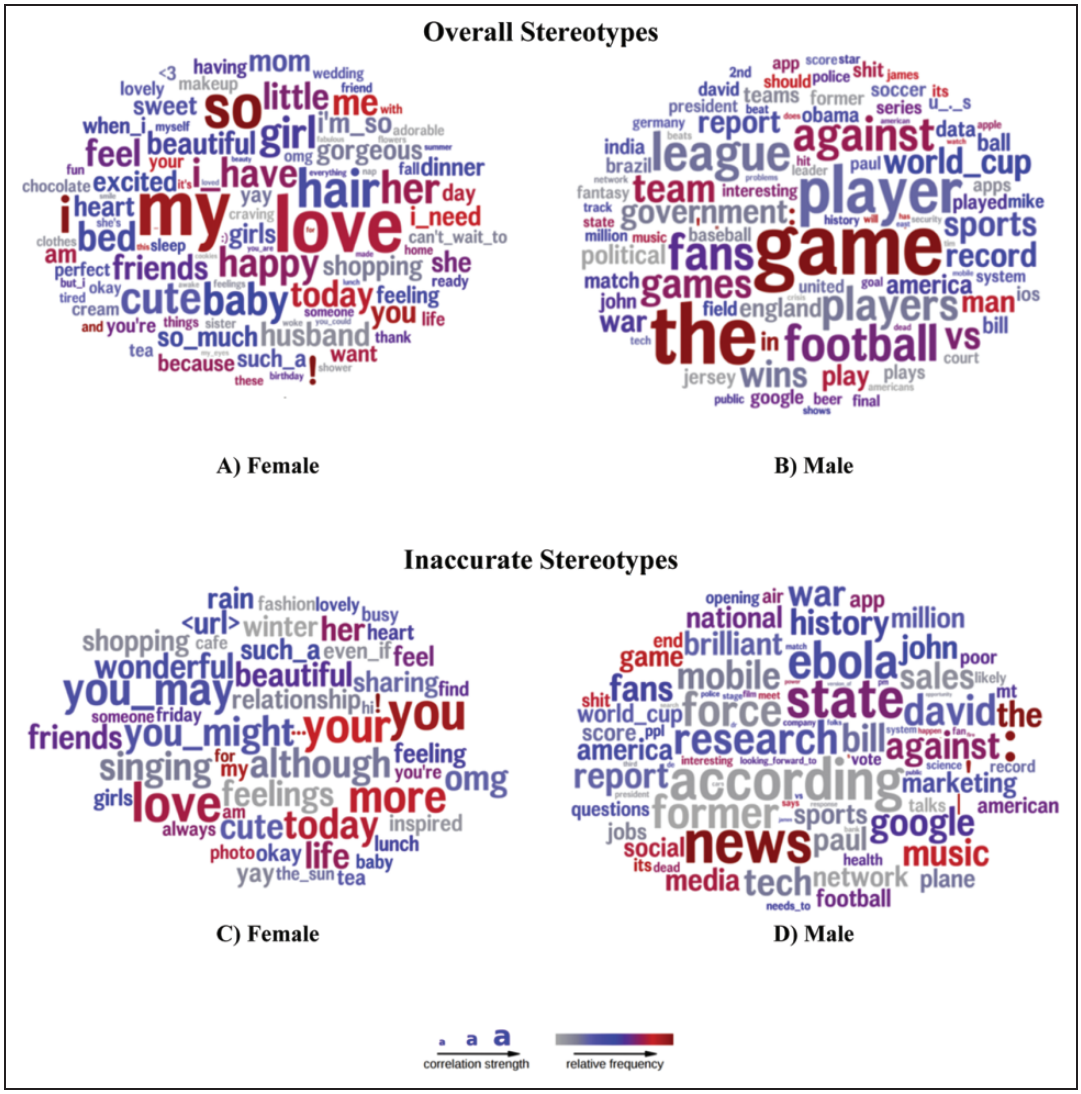

Figure 2. Words/phrases correlated with the ratio of total raters who categorized authors into each gender category. "Overall stereotypes" indicate words/phrases categorized as (a) female or (b) male, regardless of the ground truth. "Inaccurate stereotypes" indicate words (c) written by men but characterized as female or (d) written by women but characterized as male. Word size indicates strength of the correlation and word color indicates relative word frequency.

Table 1. Correlations for the 10 Words and Phrases Most Associated With Miscategorizing Men as Women.

Word or Phrase	Percentage of Raters Who Rated Author as a Man (r [95% CI])	Ground-Truth Maleness (r_{pb} [95% CI])	Z (p)
You	−.193 [−.227, −.158]	−.127 [−.165, −.089]	−4.80 (<.001)
Love	−.335 [−.354, −.290]	−.252 [−.286, −.218]	−6.27 (<.001)
More	−.109 [−.144, −.073]	−.058 [−.094, −.022]	−3.67 (<.001)
You may	−.109 [−.144, −.073]	−.060 [−.096, −.024]	−3.53 (<.001)
Your	−.154 [−.189, −.119]	−.118 [−.154, −.082]	−2.61 (.009)
Although	−.110 [−.144, −.074]	−.066 [−.102, −.030]	−3.17 (.002)
Today	−.206 [−.240, −.171]	−.144 [−.179, −.108]	−4.53 (<.001)
OMG	−.143 [−.176, −.106]	−.125 [−.159, −.089]	−1.30 (.194)
You might	−.094 [−.130, −.058]	−.058 [−.094, −.022]	−2.59 (.010)
Cute	−.239 [−.273, −.205]	−.204 [−.238, −.167]	−2.58 (.010)

Note. Z is based on z-transformed correlations. CI = confidence interval. [AQ3]

Table 2. Correlations for the 10 Words and Phrases Most Associated With Miscategorizing Women as Men.[AQ4]

Word or Phrase	Percentage of Raters Who Rated Author as a Woman (r [95% CI])	Ground-Truth Femaleness (r_{pb})	Z (p)
State	−.111 [−.146, −.075]	−.078 [−.114, −.042]	−2.38 (.017)
Research	−.055 [−.091, −.019]	.005 [−.031, .041]	−4.30 (<.001)
Ebola	−.081 [−.117, −.045]	−.006 [−.042, .030]	−5.39 (<.001)
agaInst	−.178 [−.213, −.144]	−.117 [−.152, −.081]	−4.43 (<.001)
History	−.106 [−.141, −.070]	−.056 [−.092, −.020]	−3.60 (<.001)
Sports	−.154 [−.189, −.119]	−.086 [−.122, −.050]	−4.92 (<.001)
David	−.119 [−.154, −.083]	−.104 [−.139, −.068]	−1.08 (.280)
War	−.145 [−.180, −.110]	−.103 [−.138, −.067]	−3.04 (.002)
The	−.247 [−.280, −.213]	−.201 [−.235, −.167]	−3.40 (.001)
News	−.095 [−.131, −.059]	−.076 [−.112, −.040]	−1.37 (.171)

Note. Z is based on z-transformed correlations. CI = confidence interval.

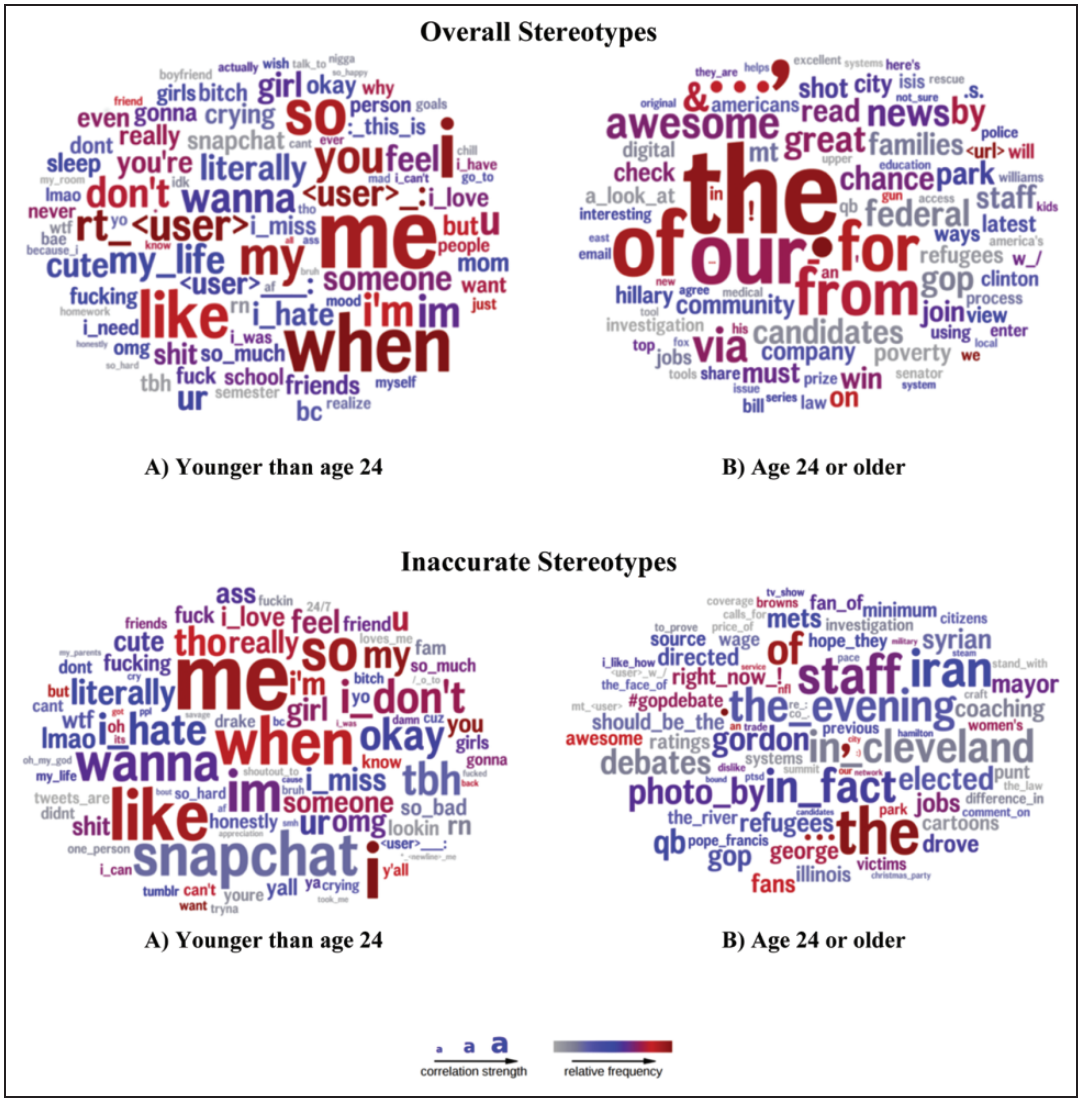

Figure 3. Words/phrases correlated with the ratio of total raters who categorized authors into each age category. "Overall stereotypes" indicate words/phrases categorized as (a) younger or (b) older, regardless of the ground truth. "Inaccurate stereotypes" indicate words (c) written by older users characterized as youth or (d) written by youth and characterized as older.

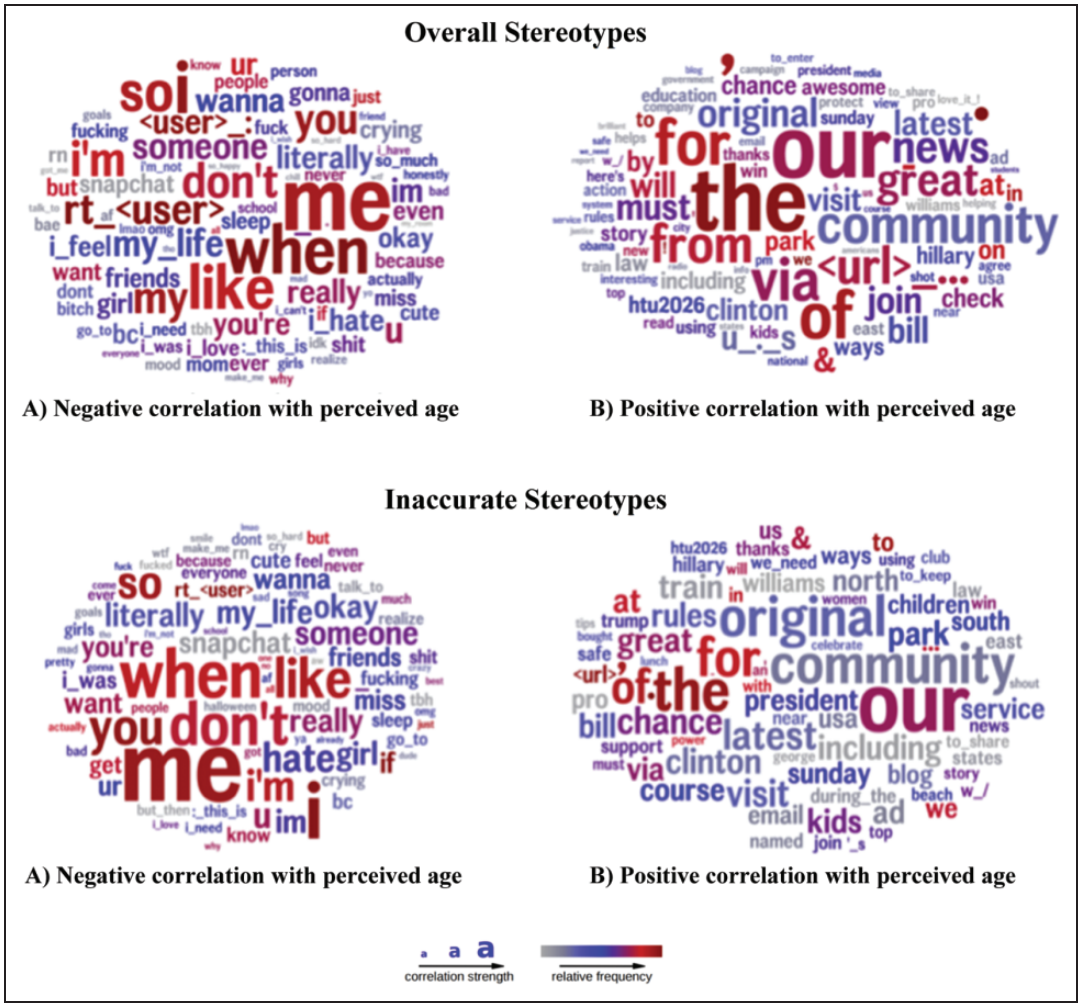

Figure 4. Words/phrases most strongly positively and negatively correlated with perceived age. Overall stereotypes" indicate words/phrases perceived to be (a) negatively correlated with perceived age or (b) positively correlated with perceived age, regardless of the ground truth. "Inaccurate stereotypes" indicate words (c) negatively correlated with perceived age or (d) positively correlated with perceived age, controlling for actual age.

older users were overly believed to mention business and politics.

As in Study 1, participants were generally accurate: 69.4% of categorizations were correct ($\chi^2 = 1{,}140.51$, $p < .001$). Younger authors were correctly identified 74% of the time and 65% were correct for older authors. Inaccurate stereotypes again were mostly exaggerated assessments of correct differences between age-groups (see Supplemental Tables S1 and S2).

To assess perceptions of age as a continuous variable, we first determined overall stereotypes by regressing words and phrases on participants' age guesses. Then, to isolate inaccurate stereotypes, we ran the same analysis controlling for authors' actual ages. The average absolute difference between real and predicted age was less than 10 years ($M = 6.80$, $SD = 7.28$), and 45% of participants' guesses were within 5 years of authors' actual ages. Real and predicted age were strongly correlated ($r = .63$). The language results were similar to those performed on age as a binary variable (Figure 4).

Study 3: Education

As illustrated in Figure 5, people without college degrees were overly assumed to use profanity and to be conversational (e.g., "lol," "wanna," "gonna"), while those with advanced degrees were exaggeratedly assumed to mention technology (e.g., "connect," tech, "web").

Raters again performed better than chance, with 45.5% of all categorizations accurate ($\chi^2 = 1{,}046.62$, $p < .001$). However, accuracy was unevenly distributed: 58.2% of ratings were correct for authors without college degrees, 55.1% were correct for authors with college degrees, and only 22.9% were correct for authors with advanced degrees. Raters had especially narrow and strict notions of the language of people with advanced degrees. As a result, inaccurate stereotypes were more likely to be the result of participants underestimating rather than overestimating education levels. (For specific inaccurate stereotypes within each group, see Figure S1 in the Supplemental Material.) Like in previous studies, many language cues were

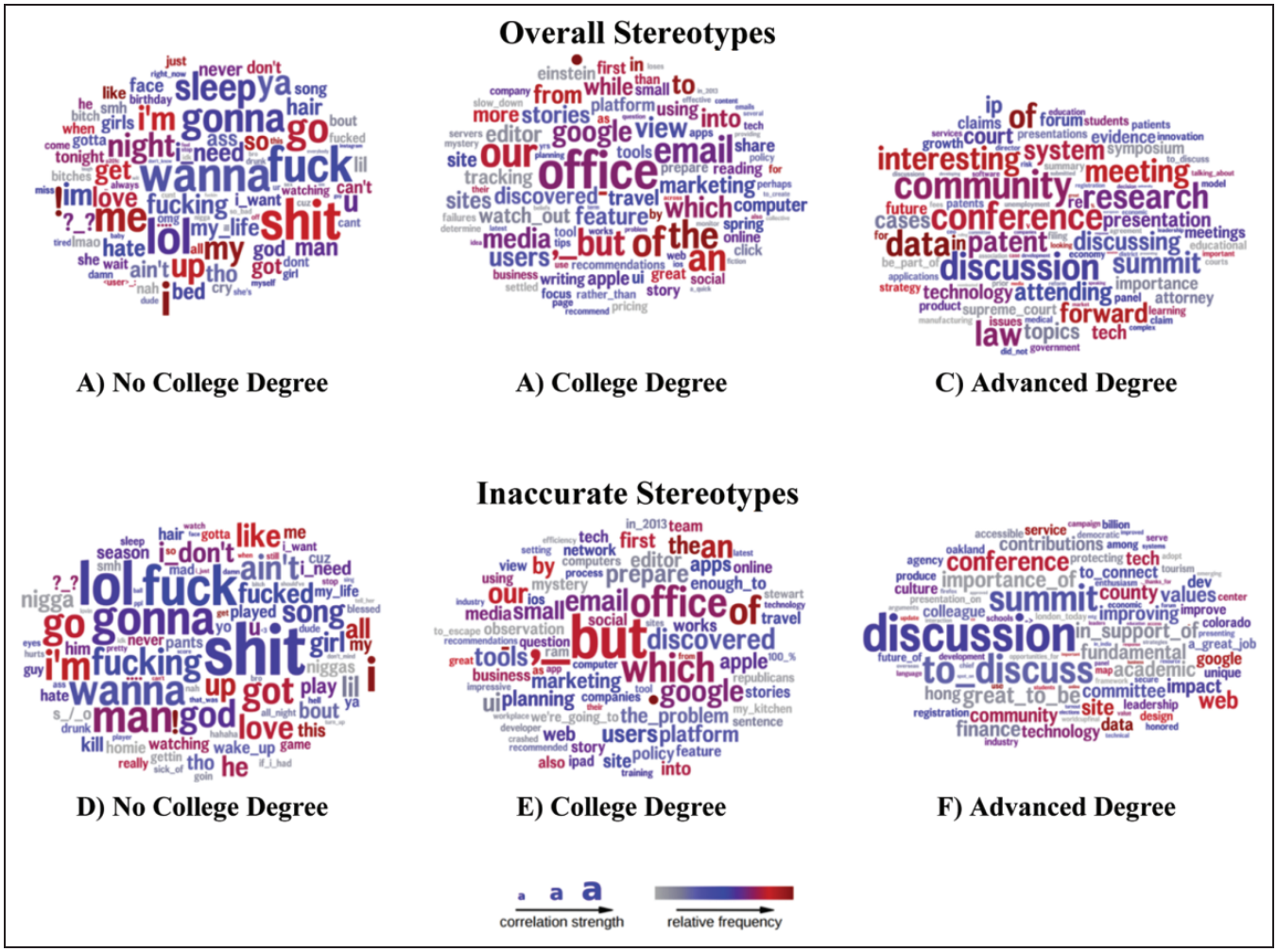

Figure 5. Words and phrases correlated with the ratio of total raters who categorized authors into each education category. "Overall stereotypes" indicate words/phrases categorized as (a) no college degree, (b) college degree, or (c) advanced degree, regardless of the ground truth. "Inaccurate stereotypes" indicate words (d) written by users with a degree but characterized as no degree, (e) written by people with no degree or advanced degrees but characterized as college degree, or (f) written by users without an advanced degree but characterized as advanced degree.

mostly exaggerated true differences (see Tables S3–S5 in the Supplemental Material for details).

One concern may be that stereotypes about education are inexorably bound up with stereotypes about age; it is possible that some authors were perceived as less educated because they were perceived as too young to have completed a college degree. Although we do not have age information for authors in Study 3, as a supplemental analysis, we estimated each authors' ages from the tweets themselves using previously validated language models (Sap et al., 2014). We then could correlate these predicted ages with both real education (i.e., the binary status of being in each of the three education levels) and perceived education (i.e., the proportion of raters who perceived authors to be in each of the three education levels).[5]

Predicted age had relatively weak correlations with the no college degree, $r = -.13$, $p < .001$; college degree, $r = .02$, $p = .59$; and advanced degree, $r = .12$, $p < .001$, categories. The relationship was stronger between predicted age and *perceived* education for perceived no degree, $r = -.38$, $p < .001$; perceived college degree, $r = .28$, $p < .001$; and perceived advanced degree, $r = .25$, $p < .00$. The relationship with perceived education was significantly higher than with actual education for all three categories (no degree: $Z = -7.48$, $p < .001$; college degree: $Z = -6.18$, $p < .001$; advanced degree: $Z = -.397$, $p < .001$). These results suggest that age-relevant stereotypes were indeed overly influential in raters' assessments of authors' levels of education.

Study 4: Politics

General stereotype information was highly political: Participants used explicit, obvious political cues when they could (Figure 6). Talking about sports was associated with mistakenly believing a liberal to be a conservative, while using conversational, feminine language was associated with mistakenly believing a conservative to be a liberal.

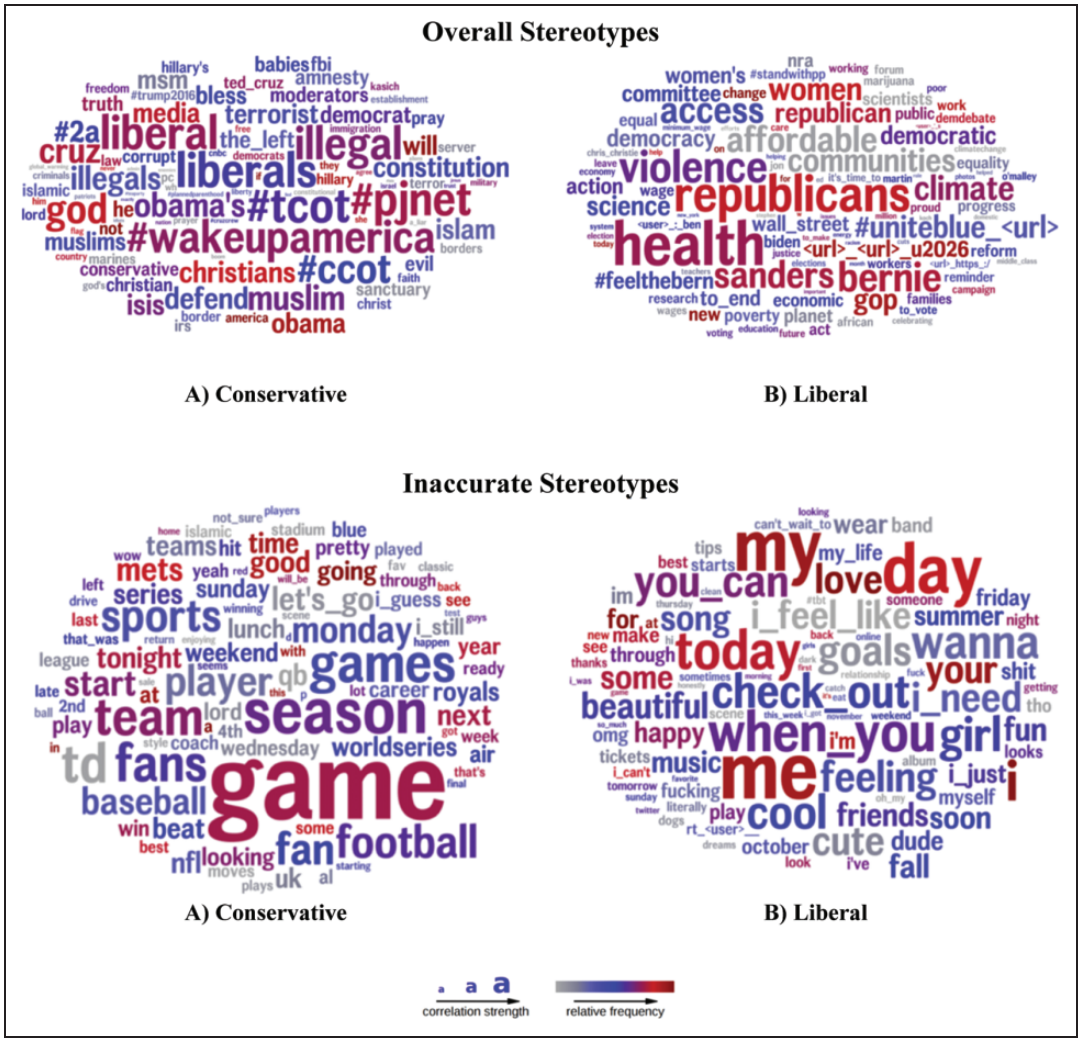

Figure 6. Words and phrases correlated with the ratio of total raters who categorized authors into each politics category. "Overall stereotypes" indicate words/phrases categorized as (a) conservative or (b) liberal, regardless of the ground truth. "Inaccurate stereotypes" indicate words (c) written by liberals characterized as conservative or (d) written by conservatives characterized as liberal.

Participants performed far better than chance, with 82% of categorizations correct ($\chi^2 = 9{,}021.19$, $p < .001$). Eighty-three percent of ratings were correct for liberal authors, and 80% of ratings were correct for conservative authors.

Unlike in Studies 1–3, participants did not simply exaggerate real-world language differences (see Supplemental Tables S6 and S7). Inaccurate stereotypes tended to be nonpolitical, but the specific effects differed by political group. Table 3 shows the words most strongly associated with falsely believing a liberal author is actually conservative. The word "game" was associated with inaccuracy for both conservative and for liberal authors. However, it was more often incorrectly believed to indicate that an author was conservative than liberal, which suggests an inaccurate association between that word and conservatism. In other words, when authors talked about nonpolitical topics, such as sports, participants were less accurate in identifying them across the board. However, words such as game, "season," and "team" were associated more strongly with thinking a liberal author was conservative than vice versa. A similar pattern occurred for incorrect stereotypes about liberals (Table 4).

Inaccurate stereotypes for liberals and conservatives appeared to be gendered in nature (compare "inaccurate stereotypes" in Figure 6 with overall stereotypes in Figure 2). We did not have gender information for Study 4 authors, but, similar to our technique in Study 3, we estimated the gender of each author directly from tweets (Sap et al., 2014). Predicted gender correlated with actual political orientation, such that authors predicted to be female were actually more liberal, $r_\varphi = .14$, $p < .001$. However, predicted gender had a stronger correlation with *perceived* political orientation, $r_\varphi = .21$, $p < .001$, a difference that was statistically significant, $Z = 6.10$, $p < .001$. This suggests that participants exaggerated the importance of gendered cues in determining the political orientation of authors.

Table 3. Correlations With Inaccurate Categorization for the 10 Words and Phrases Most Associated With Inaccurate Stereotypes of Conservatives in Study 4.

Word or Phrase	Inaccurate Belief That an Author Is Conservative (*r* [95% CI])	Inaccurate Belief That an Author Is Liberal (*r* [95% CI])	*Z* (*p*)
Game	.311 [.260, .361]	.126 [.071, .180]	4.87 (<.001)
Season	.213 [.161, .265]	.114 [.059, .168]	2.54 (.011)
Games	.194 [.141, .246]	.100 [.045, .154]	2.40 (.016)
TD	.192 [.139, .244]	.018 [−.037, .073]	4.40 (<.001)
Team	.190 [.137, .243]	.100 [.045, .154]	2.30 (.021)
Fans	.188 [.135, .241]	.098 [.043, .152]	2.30 (.021)
Sports	.174 [.121, .227]	.083 [.028, .138]	2.31 (.021)
Football	.172 [.118, .226]	.034 [−.021, .089]	3.49 (<.001)
Fan	.169 [.116, .224]	.100 [.045, .154]	1.76 (.078)
Player	.164 [.108, .216]	.109 [.054, .163]	1.40 (.162)

Note. *Z* is based on z-transformed correlations. CI = confidence interval.

Table 4. Correlations With Inaccurate Categorization for the 10 Words and Phrases Most Associated With Inaccurate Stereotypes of Liberals in Study 4.

Word or Phrase	Inaccurate Belief That an Author Is Liberal (*r* [95% CI])	Inaccurate Belief That an Author Is Conservative (*r* [95% CI])	*Z* (*p*)
Me	.278 [.219, .329]	.015 [−.040, .070]	6.75 (<.001)
My	.264 [.211, .314]	.073 [.018, .128]	4.93 (<.001)
Day	.252 [.200, .304]	.085 [.030, .140]	4.30 (<.001)
I	.227 [.174, .279]	.064 [.009, .119]	4.17 (<.001)
When you	.226 [.173, .278]	.028 [−.027, .083]	5.04 (<.001)
Today	.223 [.170, .275]	.030 [−.025, .085]	4.91 (<.001)
Wanna	.211 [.157, .263]	−.011 [−.066, .044]	5.62 (<.001)
Girl	.210 [.156, .262]	−.042 [−.097, .013]	6.37 (<.001)
Cool	.209 [.155, .261]	.048 (−.007, .103]	4.10 (<.001)
Check out	.208 [.154, .260]	.024 [.031, .079]	4.67 (<.001)

Note. *Z* is based on z-transformed correlations. CI = confidence interval.

General Discussion

Through four studies, we examined words and phrases that contribute to stereotypes. Consistent with previous literature that stereotypes are often accurate (e.g., Jussim et al., 2015), participants were generally skilled in guessing a person's group membership. Errors tended to be exaggerations rather being than completely wrong. The exception was politics, where nonpolitical language led to inaccuracy across the board; however, people exaggerated the association between women and liberalism.

The online, language-based context allowed for two important innovations. First, it lets us tease apart a very large number of interconnected, group-based associations and specifically identify those that led to inaccuracy. If we simply asked participants to describe overall group tendencies, it is likely that exaggerated stereotypes would appear to be accurate, since they match group-level differences in central tendency. However, these stereotypes cannot be considered accurate, since they were associated with objectively incorrect beliefs about a person's group membership.

Allport (1954) previously suggested that stereotypes are exaggerations of real group differences, but this has been controversial (e.g., McCauley, 1995). Our findings suggest that not all stereotypes are exaggerations; many stereotypes are correctly associated with a person's group membership. However, when a categorization is wrong, raters appear to have drawn on exaggerated aspects of a stereotype. For instance, in Study 1, writing about technology was primarily associated with women mistakenly being identified as men. Although men were more likely than women to post about technology, raters believed the difference was more indicative of maleness than it actually was, resulting in false positive categorizations of men. Novel techniques such as ours are needed to determine when stereotypic beliefs and associations are truly adaptive versus unadaptive.

These results imply potential targets for intervention. Not only does the stereotypical association between maleness and technology potentially result in negative societal consequences (e.g., Murphy, Steele, & Gross, 2007), it was associated with incorrect conclusions about men and women. Making people aware of their subtle, gendered associations with technology may reduce harmful biases that limit women's opportunities to advance in science, technology, engineering, and math fields (e.g. Shapiro & Williams, 2012; Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012) and might increase interpersonal accuracy.

Notably, there was similarity between the inaccurate stereotypes of politics and general stereotypes of gender. Our results suggest that people defaulted to gender stereotypes when attempting to guess people's political orientations, assuming that masculine people are conservative and feminine people are liberal. This pattern may emerge from stereotypes connecting both liberalism and women with warmth while connecting both conservatism and men with instrumentality (Huddy & Terkildsen, 1993). Female authors in our sample were more likely to be liberal but not as much as raters appeared to believe; this finding is thus another example of people exaggerating the diagnostic utility of an actual group association. A similar result was found in Study 3 between education level and age.

Our second innovation is that our method identifies the language that makes up stereotypes. In face-to-face interactions, people simultaneously use information from multiple channels to categorize others, which makes it ambiguous what cues were most important. Using social media language, lets us isolate a single channel within the context of everyday life, allowing us more certainty that the identified stereotypes are real.

Also, by not directly asking participants to explicitly list aspects of their stereotypes, our method avoids self-presentation concerns (e.g., Plant & Devine, 1998) and highlights information that was reliably associated with categorization, but which people may be unlikely to explicitly verbalize or even consciously notice. Our methods therefore

present a novel solution to the problem of how to identify very subtle and nuanced aspects of stereotypes.

Despite these novel findings, our methods had several limitations. The social media environment allowed us to isolate a single information channel, but behaviors on Twitter may not wholly generalize to other contexts. Our method highlighted the entire set of words most strongly related to miscategorizations, but we do not know if specific words or stylistic choices had the most directly causal impact on participants' incorrect guesses. Third, only a single characteristic was available in each study, but these characteristics most likely are correlated (e.g., age and education). In Studies 3 and 4, we estimated age and gender, providing some insights, but were limited by the data available. Finally, we treated all participants the same in their ratings. In the future, it will be useful to see if there are any individual differences associated with the ability to avoid the influence of misleading cues online.

Our studies indicate the power of big data methods to quantitatively compare actual and perceived behavioral tendencies across groups. Using social media text to unobtrusively measure both behaviors and perceptions of those behaviors can reveal surprising, important features of people's stereotypical beliefs and their levels of correctness.

Acknowledgments

The authors thank Johannes Eichstaedt, Laura Smith, H. Andrew Schwartz, Patrick Crutchley, and Eleanor Hanna for their feedback and suggestions for revisions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by a grant from the Templeton Religion Trust (ID #TRT0048).

Supplemental Material

The online [appendices/data supplements/etc] are available at http://spp.sagepub.com/supplemental.

Notes

1. For the purposes of this study, authors were considered either male or female. While this binary choice does not capture the full range of gender identity or perception, it matches automatic gender categorization that occurs in the real world (e.g., Quinn, Yahr, & Kuhn, 2002).
2. To minimize the possibility of raters assuming ongoing education, they were told that all authors were older than 22 years and also not currently in school.
3. Because age is a continuous variable, raters guessed the authors' ages in years. Both guessed age and real age were then applied to the two age categories, split at 23.
4. For quality control, we interspersed several authors who directly stated their group category (e.g., a male author saying "My beard is almost to the point where I can make other men jealous of my sweet beard"). If participants misidentified two of these unambiguous authors, they were unable to participate further and their data are not included in our results. In addition, raters had to spend at least 10 s on each task before being allowed to submit their guesses. Overall, 16, 8, 20, and 40 raters failed the attention checks in Studies 1, 2, 3, and 4, respectively.
5. For actual education, authors belonged to one age category, and a linear contrast (with weights of -1, 0, and 1 for *no college degree*, *college degree*, and *advanced degree*, respectively) could be used, regressing predicted age onto the linear contrast. As expected, there was a positive, linear relationship between predicted age and actual education level, $\beta = .15$, $p < .001$. For perceived education, the categories are nonindependent, with scores representing the proportion of ratings for that category. Although we could force each author into a single perceived education category using raters' majority vote, this would lose information, so we chose to retain all rating information and estimate the correlations for each category separately.

References

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

Argamon, S., Koppel, M., Pennebaker, J., & Schler, J. (2007). Mining the blogosphere: Age, gender, and the varieties of self-expression. *First Monday*, *12*. Retrieved from http://firstmonday.org/issues/issue12_9/argamon/index.html

Chung, C., & Pennebaker, J. W. (2007). The psychological functions of function words. In K. Fiedler (Ed.), *Social communication* (pp. 343–359). New York, NY: Psychology Press.

Diekman, A. B., Eagly, A. H., & Kulesa, P. (2002). Accuracy and bias in stereotypes about the social and political attitudes of women and men. *Journal of Experimental Social Psychology*, *38*, 268–282.

Dovidio, J. F., Brigham, J. C., Johnson, B. T., & Gaertner, S. L. (1996). Stereotyping, prejudice, and discrimination: Another look. *Stereotypes and Stereotyping*, *276*, 319.

Eagly, A. H. (1995). The science and politics of comparing women and men. *American Psychologist*, *50*, 145.

Eichstaedt, J. C., Tobolsky, V., Yaden, D. B., Schwartz, H. A., Kern, M. L., Park, G., . . . Seligman, M. E. P. (2016). From hypothesis-testing to hypothesis-generation: A review and quantitative comparison of open and closed-vocabulary approaches for text analysis. Manuscript in preparation. [AQ5]

Elias, P., & Birch, M. (2010). SOC2010: Revision of the standard occupational classification. *Economic and Labour Market Review*, *4*, 48–55.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4.

Haas, A. (1979). Male and female spoken language differences: Stereotypes and evidence. *Psychological Bulletin*, *86*, 616.

Huddy, L., & Terkildsen, N. (1993). Gender stereotypes and the perception of male and female candidates. *American Journal of Political Science*, *37*, 119–147.

Judd, C. M., & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review*, *100*, 109.

Jussim, L., Cain, T. R., Crawford, J. T., Harber, K., & Cohen, F. (2009). The unbearable accuracy of stereotypes. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 199–227). New York, NY: Taylor & Francis.

Jussim, L., Crawford, J. T., & Rubinstein, R. S. (2015). Stereotype (in) accuracy in perceptions of groups and individuals. *Current Directions in Psychological Science*, *24*, 490–497.

Jussim, L., & Zanna, M. P. (2005). Accuracy in social perception: Criticisms, controversies, criteria, components, and cognitive processes. *Advances in Experimental Social Psychology*, *37*, 1–93.

Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Park, G., Ungar, L. H., Stillwell, D. J., . . . Seligman, M. E. (2014). From "sooo excited!!!" to "so proud": Using language to study development. *Developmental Psychology*, *50*, 178–188.

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L., & Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological Methods*. http://dx.doi.org/10.1037/met0000091

Lee, I. A., & Preacher, K. J. (2013). *Calculation for the test of the difference between two dependent correlations with one variable in common* [Computer software]. Retrieved December, 2015, from http://quantpsy.org/corrtest/corrtest2.htm

Lui, M., & Baldwin, T. (2012, July). langid. py: An off-the-shelf language identification tool. *Proceedings of the ACL 2012 system demonstrations* (pp. 25–30). Association for Computational Linguistics, Jeju, Republic of Korea.

MacCallum, R., Zhang, S., Preacher, K., & Rucker, D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19.

McCauley, C. (1995). Are stereotypes exaggerated? A sampling of racial, gender, academic, occupational, and political stereotypes. In Y. Lee, L. Jussim, & C. R. McCauley (Eds.), *Stereotype accuracy: Toward appreciating group differences* (pp. 215–243). Washington, DC: American Psychological Association.

McCauley, C., & Stitt, C. L. (1978). An individual and quantitative measure of stereotypes. *Journal of Personality and Social Psychology*, *36*, 929.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*, 16474–16479.

Murphy, M. C., Steele, C. M., & Gross, J. J. (2007). Signaling threat how situational cues affect women in math, science, and engineering settings. *Psychological Science*, *18*, 879–885.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, *45*, 211–236.

Nguyen, D. P., Gravel, R., Trieschnigg, R. B., & Meder, T. (2013, July). *"How old do you think I am?" A study of language and age*. Proceedings of the seventh international AAAI conference on weblogs and social media, ICWSM, Cambridge, MA. [AQ6]

Nguyen, D. P., Trieschnigg, R. B., Doğruöz, A. S., Gravel, R., Theune, M., Meder, T., & de Jong, F. M. G. (2014, August). *Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment*. Berkeley, CA: Association for Computational Linguistics.

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

Pennacchiotti, M., & Popescu, A. M. (2011). A machine learning approach to Twitter user classification. *International Conference on Web and Social Media*, *11*, 281–288.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: LIWC 2001*. Mahwah, NJ: Lawrence Erlbaum Associates.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*, 811.

Preoţiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., & Aletras, N. (2015). Studying user income through language, behaviour and affect in social media. *PLoS ONE*, *10*, *e0138717*.

Prothro, E. T., & Melikian, L. H. (1955). Studies in stereotypes: V. Familiarity and the kernel of truth hypothesis. *The Journal of Social Psychology*, *41*, 3–10.

Quinn, P. C., Yahr, J., Kuhn, A., Slater, A. M., & Pascalis, O. (2002). Representation of the gender of human faces by infants: A preference for female. *Perception*, *31*, 1109–1121.

Rao, D., Paul, M. J., Fink, C., Yarowsky, D., Oates, T., & Coppersmith, G. (2011). Hierarchical Bayesian models for latent attribute detection in social media. *International Conference on Web and Social Media*, *11*, 598–601.

Ryan, C. (2003). Stereotype accuracy. *European Review of Social Psychology*, *13*, 75–109.

Sap, M., Park, G., Eichstaedt, J. C., Kern, M. L., Stillwell, D. J., Kosinski, M., . . . Schwartz, H. A. (2014). Developing age and gender predictive lexica over social media. *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1146–1151.

Schwartz, H. A., Eichstaedt, J. C., Dziurzynski, L., Kern, M. L., Blanco, E., Kosinski, M., . . . Ungar, L. H. (2013). *Toward personality insights from language exploration in social media.* Proceedings of the AAAI spring symposium series: Analyzing Microtext, Stanford, CA.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., . . . Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, *8*, e73791.

Shapiro, J. R., & Williams, A. M. (2012). The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields. *Sex Roles*, *66*, 175–183.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, *73*, 751–754.

Stangor, C. (1995). Content and application inaccuracy in social stereotyping. In Y. Lee, L. Jussim, & C. R. McCauley (Eds.), *Stereotype accuracy: Toward appreciating group differences* (pp. 275–292). Washington, DC: American Psychological Association.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245.

Sylwester, K., & Purver, M. (2015). Twitter language use reflects psychological differences between Democrats and Republicans. *PloS ONE*, *10*, *e0137422*.

Author Biographies

Jordan Carpenter is a postdoctoral researcher at the University of Pennsylvania. He received his PhD in social psychology from the University of North Carolina in 2013.

Daniel Preoţiuc-Pietro is a postdoctoral researcher in natural language processing working for the World Well-Being Project in the Positive Psychology Center of the University of Pennsylvania. His current research leverages large-scale social media footprints to aid with psychology and health problems.

Lucie Flekova is a PhD Student in natural language processing at the Department of Computer Science, Technische Universität Darmstadt. She focuses on stylistic and semantic analysis of text with applications in author profiling.

Salvatore Giorgi is a research programmer at the World Well-Being Project at the University of Pennsylvania.

Courtney Hagan is a research assistant with the World Well-Being Project.

Margaret L. Kern is a senior lecturer at the Centre for Positive Psychology at the University of Melbourne's Graduate School of Education. Her research examines the question of who flourishes in life (physically, mentally, and socially), why, and what enhances or hinders healthy life trajectories.

Anneke E. K. Buffone currently is the lead research scientist and a postdoctoral fellow at the University of Pennsylvania's World Well-Being Project. Buffone's research specializes in other-focused motivations, emotions, and cognitions and its effects on social interactions.

Lyle Ungar is a professor of computer and information science at the University of Pennsylvania, where he also holds appointments in other departments in the schools of Engineering, Arts and Sciences, Medicine, and Business. His current research interests include machine learning, text mining, statistical natural language processing, and psychology.

Martin E. P. Seligman is the Zellerbach Family Professor of psychology and director of the Positive Psychology Center at the University of Pennsylvania, where he focuses on positive psychology, learned helplessness, depression, and optimism.