Held Stereotypes' Effects on Scalar Inference: a Norming Study

Dean Manko with Brandon Waldon, Judith Degen, and Zion Mengesha

LINGUIST 245B — Spring 2020
Stanford University

Listeners have been shown to incorporate world knowledge and context in their interpretation of sentences and scalar inferences, but how do prior beliefs like racial ideologies play a role in this interpretation? In the first step of a larger study analyzing how stereotypical beliefs and racialized judgments affect scalar inferences, we focused on quantifying how participants view the likelihood of strong meanings associated with agents of different races. Participants (n=125) were presented with sentences that both evoked and repressed the stereotype that Black Americans are associated with criminality, and asked to judge how likely the sentences were to be true. They saw ten critical sentences including both Black-normed names and White-normed names. The data presents some weak evidence for the elicitation of these stereotypical beliefs, with the average response for Black-normed names seeming to be both more likely to evoke and less likely to repress the stereotype, but the findings were ultimately not statistically significant.

I. Introduction

The way we interpret and understand sentences largely has to do with the context in which they appear. To this point, the concept of context affecting the way we extract meaning — such as scalar inferences — has been demonstrated in a range of literature. This norming study is the first step of a larger project hoping to expand on current literature linking scalar inference to context-affected parsing. Context's effect on parsing has been demonstrated regularly, including listeners adjusting parsing expectations with context to not be surprised to hear a story about a peanut being in love in Nieuwland & Van Berkum's N400 studies (2006). But within the literature of processing, scalar inference presents a window into language interpretation beyond the level of parsing sentences. Scalar inference looks at the phenomenon whereby the use of a less informative term (e.g., *some of*) is inferred to mean the negation of a more informative term (e.g., to mean *not all of*) (Politzer-Ahles & Fiorentino, 2013). Work such as Geurts (2010) looked at the affect of *a priori* probabilities on scalar inferences, and found that scale inferences arose even when prior probably of a stronger meaning was high. Additionally Degen & Tanenhaus (2015) found that "prior world knowledge" of an object (e.g. its physical properties of floating, sinking, or sticking to a wall) affected the scalar inferences drawn by comprehenders. This study aims to take this work in a different direction by looking at internalized attitudes as a different form of context, departing from world knowledge to look at listener-specific prior beliefs.

In the context of critical race theory as presented by Moya and Markus (2011), race and ethnicity are not necessarily essential to people, but are social processes. Differences emerge between races by the different social processes people of different races engage with, by both doing and having different actions done to them. This is only further ingrained in attitudes by popular media and the controlling images it creates, which play on stereotypical ideas to portray marginalized groups in media in ways that are gendered, classed, and racialized (Collins 2004).

In a small step towards looking how prior beliefs — especially ones that have the clear dangerous and material implications that stereotypes do— manifest in parsing, we are analyzing how stereotypical beliefs and racialized judgments affect scalar inferences. To do so involved looking at how the racial ideologies of listening subjects[1] play a role in interpreting weak and strong alternatives on pragmatic scales. This norming study took a first step in this direction by trying to measure subjects' racial ideologies through investigating whether stereotype-evoking sentences and stereotype-repressing sentences are judged to be more or less likely to be true given who the participant thinks they are about. The goal was to understand the *a priori* holdings of stereotypes (and, as such, the prior beliefs) by looking at how participants view the likelihood of strong meanings associated with agents of different races. This study focused on the stereotypical association between Black Americans and criminality to elicit stereotypical beliefs.

[1] "Listening subjects" reflects the idea of observers or non-group members determine the attitudes around a given group, such as the stereotypes held and propagated by White American about Black Americans (Inoue, 2006; Flores & Rosa, 2019).

In measuring these judgments, we make the linking assumption that the mean judgment of likelihood is able to shed light on stereotypical beliefs. Higher mean responses for Black-normed names than White-normed names on stimuli that evoke the stereotype suggest that this stereotype is a belief that is held by the participant. Similarly, higher mean responses for White-normed names than Black-normed names on stimuli that *repress* the stereotype suggest that this stereotype is again a belief that is held by suggesting that subjects believe Black-normed names are less likely to repress the stereotype.

As such, it is hypothesized that: 1) For sentences that evoke the criminality stereotype: higher likelihood ratings when the sentence features a stereotypically black name (than when the name is stereotypically white); 2) For sentences that repress the criminality stereotype: higher likelihood ratings when the sentence features a stereotypically white name (than when the name is stereotypically black).

All code, data, and analysis are on GitHub: https://github.com/dmanko99/Manko245B_Project
The experiment's OSF preregistration can be found here: https://osf.io/rn2jw

II. Norming Study

Methods

An example of the example stimuli for the experiment is shown below. In (1a), the stimulus is evoking the criminality stereotype, whereas (1b) is repressing the stereotype:

1. (a) All of Deandre's cousins have spent time in prison.

   (b) All of Deandre's cousins work at a prison.

This subject of the statement was either a Black-normed name, as shown in (1), or a White-normed name, as we see in (2):

2. (a) Tanner has threatened a judge and a police officer.

   (b) All of Tanner's friends respect the police.

Presenting participants with these critical trials, the measurement of concern was participants' judgments of how likely the sentences were to be true.

PARTICIPANTS

For this study, we recruited 140 participants on Amazon Mechanical Turk (mTurk). We wished to collect an average of 20 responses per critical sentence per name category (this amounts to an average of 40 responses / sentence), multiplied by 3 because we have 3 experimental lists (120). We include an additional 20 participants on the assumption that some will fail our attention check items. After exclusion categories, we were left excluding 15 participants (n = 125). Of the 125 participants, all of them were based in the United States and between the ages of 19 and 66. They had all completed a minimum of 1000 hits on mTurk prior to the experiment, and were paid $0.70 for its completion which took an average of 4 minutes to complete.

MATERIALS

To norm the stereotype-based stimuli, we focused on criminality as a racial stereotype commonly held about Black Americans; this was a decision informed by both sociological and computational data. Golash-Boza recognizes the "Black thug stereotype" as a prominently held attitude in the United States, and shows how Black Americans are disproportionately mediatized as thugs, gang members, or criminals (2016). Additionally, we sourced the criminality stereotype from Sap et al.'s Social Bias Inference Corpus (SBIC), a collection of annotations of potentially offensive social media posts including more than 18,000 posts where annotators identify posts targeting different identities, including the Black community (2019).

To create sentences for the presented stimuli, we used five pragmatic scales for which scalar implicatures can arise — <*or*, *and*>, <*some*, *all*>, <*looks like*, *is*>, <*possible*, *certain*>, and <*n*, *n*+*m*>. However, because we were focused on the strong meanings of these scales, we eliminated the weak alternatives for stimuli generation (leaving us with *and*, *all*, *is*, *certain*[2], and *n*+*m*).

We combined these items from the pragmatic scales with words whose (pre-trained word2vec) vector representations are close (cosine distance) to the word *criminal*. Then, to source the names, we looked at Black- and White-normed names from Stelter and Degner (2018), selecting fifteen of each name.[3] These result in the critical stimuli containing named names, and stories relating to the criminality stereotype. In addition to the critical stimuli, we presented participants with filler items that are assumed not to evoke racialized stereotypes, and simple math problems that served as an attention check and exclusion criteria for participants who might not have been completing the measures correctly — answering any attention check incorrectly (e.g. saying 1+1=2 is closer to being *not at all likely* to be true than it is to being *extremely likely*) will lead to exclusion of all data from the participant.

[2] The lexical item *certain* is excluded from the certain/possible stimuli because we determined that it was awkward to have participants rate the likelihood of sentences of the form ‘it is certain that p’. We assume that participants’ ratings of bare ‘p’ can serve as proxies for their hypothetical ratings of ‘it is certain that p’.

[3]Black-normed names included: Trevon, Tyree, Deion, Marquis, Jermaine, Lamont, Tyrone, Deandre, Tremayne, Lamar, Kareem, Hakeem, Jamal, Rasheed, Deshawn.
White-normed names included: Peter, Brad, Ethan, Ian, Cody, Brett, Paul, Connor, Jack, Logan, Roger, Dylan, Hunter, Dustin, Ryan

PROCEDURE

This is a within-subjects design. Participants see 10 critical sentences of interest, two from each pragmatic scale, balanced for evocation/repression of the criminality stereotype. In addition, participants see ten filler items — five of which (coded 'highbias') we believe will elicit high ratings overall, and five of which (coded 'lowbias'). Furthermore, each participant sees four attention checks over the course of the study. When presented with a sentence, the participant is asked how likely they think that the statement is true, and provides a value from "not at all likely" to "extremely likely" using a slider scale.

Results

GENERAL FINDINGS

Data was collected using mechanical Turk, and analyzed to look at the mean response for the category of the normed name (i.e. Black- or White-normed names) for sentences both evoking and repression the criminality stereotype. The general pattern of results for critical stimuli suggested that there was a small, but not statistically significant effect. As seen in Figure 1, responses to White-named stimuli were higher on average in the repress condition than the

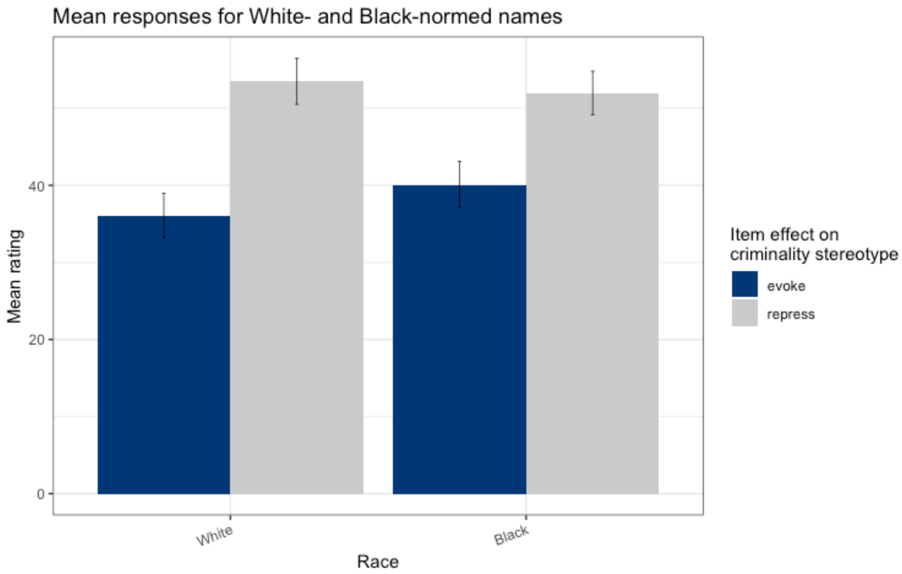


FIGURE 1: On average, White-normed names elicited higher mean ratings than Black-normed names in the repress condition, and lower mean ratings in the evoke condition.

responses to Black-normed names (53.49 vs. 51.97), and the responses to Black-normed names were higher on average in the evoke condition (39.98 vs. 36.05 for White-normed names).

However, this difference was not shown to be significant. To analyze the results we ran a Bayesian linear mixed effects regression predicting slider bar response from fixed effects of name category (Black-/White-normed names), relation of sentence to criminality stereotype (evoke / suppress), and the interaction of name category and stereotype. The model included by-item, by-name, and by-participant random intercepts; a by-item random slope for name category; a by-name random slope for stereotype; and by-participant slopes for name category, stereotype, and their interaction. This is the maximal random effects structure justified by the design (participants see items of both name categories and both stereotype features; items vary by name category but not by stereotype features). This was conducted using the default priors over parameter values provided by the brms package in R. The model's 95% confidence intervals included 0 (Table 1), which does not suggest a strong positive interaction.

TABLE 1: Summary of Bayesian linear mixed effects regression model

	β	L-95% CI	U-95%CI	r̂
response ~ stimulus race * stereotype evocation + (1+stimulus race|item) + (1+stereotype evocation|stimulus name) + (1+stimulus race * stereotype evocation|participant)	4.72	-0.31	9.72	1.00

TABLE 1: The lower confidence interval of the model suggests a lack of a strong, significant effect between the response and the interaction between the stimulus race and stereotype evocation.

Further testing showed evidence for a weak positive interaction, using non-linear hypothesis testing of fixed effects parameters. A posterior probability under the hypothesis against its alternative of 0.97 suggests that there *might* be weak evidence for a small effect.

FIGURE 2

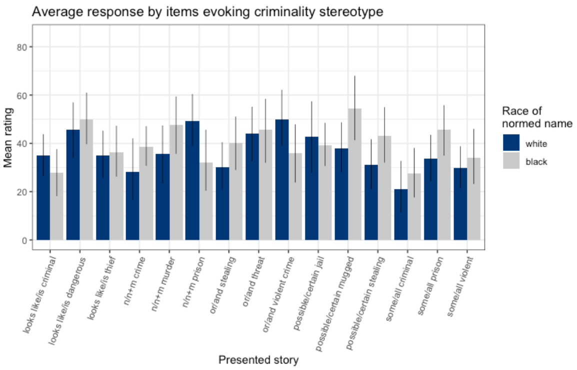
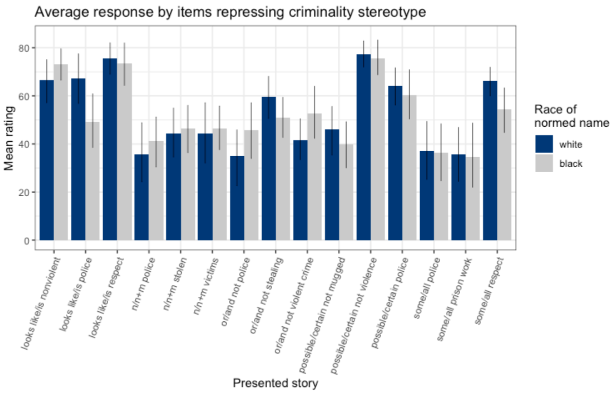


FIGURE 3



FIGURES 2 & 3: The average mean ratings for items showed high by-item variability and no significant patterns.

Inter-item and inter-participant variability showed that, for most cases, the evoke stimuli elicited lower average responses than stimuli that repressed the stereotype. However, as seen in Figures 2 & 3, the responses themselves depended on the story presented in the stimulus, with some stories eliciting higher responses on average than others (but the level of variability within each story prevented any definitive conclusions).

SUBSETTING SUBJECTS: RACE & POLITICAL AFFILIATION

With the issue of stereotypes, personal identities often play a role in perception of societal — and, by extension, racial — groups and the beliefs we hold about them (Moya & Markus, 2011). As such, looking at data by the participants racial and social identities allows us to more closely examine the idea of the "listening subject" (Inoue, 2006; Flores & Rosa, 2019). When breaking down the subjects by race, using categories laid out by Davenport (2018), we are left with very few non-white participants and no real observable difference in pattern (Figure 4).

In a measure of salient social identity, we also broke the responses down by the subjects' political affiliation. Subjects self-reported political identity on a scale adapted from the Pew Research Center, and these categories were grouped into larger categories with no effect on the shown patterns. Figure 5 shows no different observable pattern from the already-observed findings in the main three categories (Democrat, Independent, and Republican).

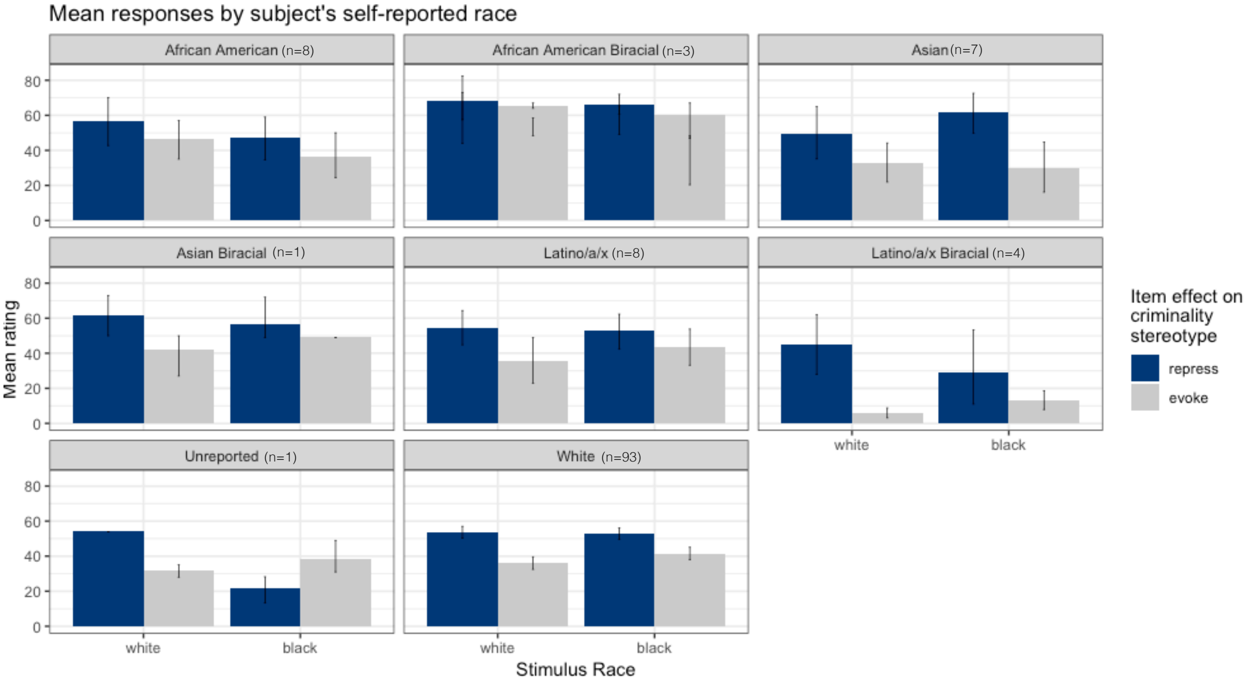


FIGURE 4: Grouping the subjects by their self-reported race presents similar patterns to Figure 1; most of the non-conforming results are groups of non-white speakers that only contain a handful of participants.

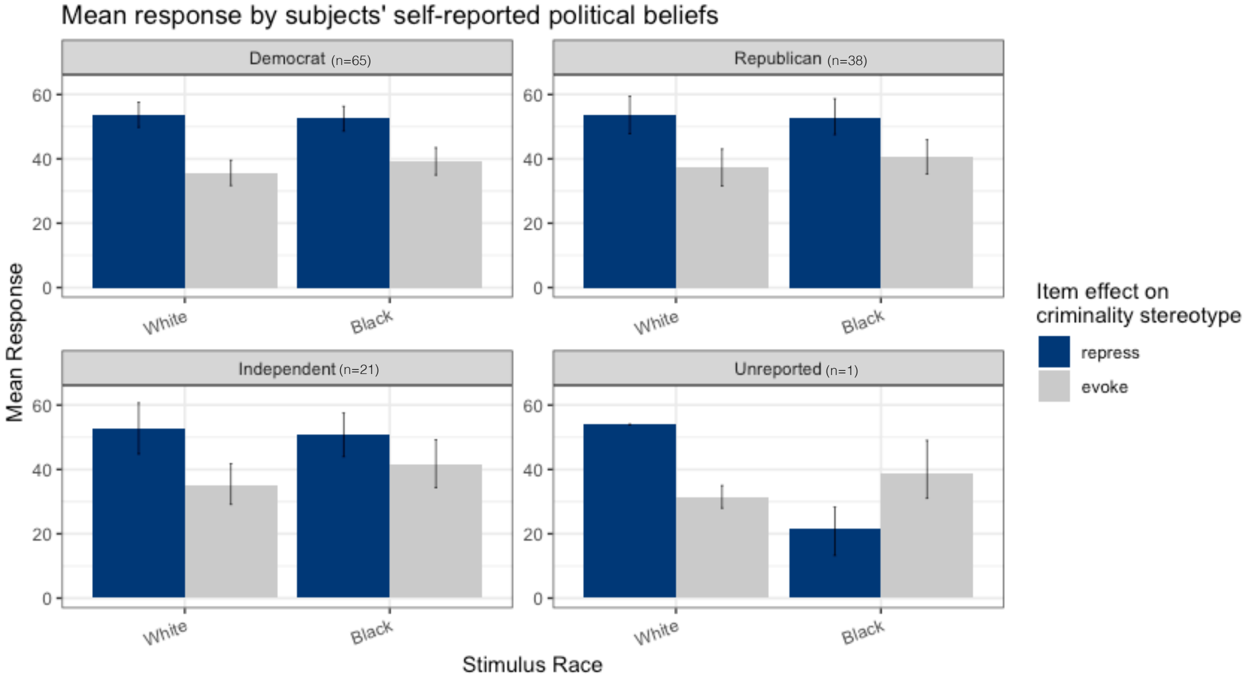


FIGURE 5: Grouping the subjects by their self-reported political beliefs largely similar patterns to the larger overview in Figure 1, with the exception of the (very small) group of participants who chose not to report.

III. Discussion

In this early step of measuring how world views and stereotypical beliefs affect scalar implicatures and, by proxy, language processing, judgments of sentences evoking and repressing the stereotype that Black Americans are associated with criminality were collected. The general pattern of findings does not provide evidence of a significant affect of the name category on the judgment made. The Bayesian linear mixed effects regression model provided evidence that there was no significant correlation between Black-normed names and higher judgments for stimuli evoking the criminality stereotype.

Although the hypotheses were not borne out, the data still leaves us with valuable information and a variety of possible next steps. The posterior probability of the Bayesian model suggests that there might be an effect, but encourages taking different approaches to isolate this effect. Going forward, in the attempt to norm these stimuli and elicit stereotypical beliefs, one possible step is to control for a specific identity of participant. While no effects jumped out for participant subsets, a higher-powered look at groups of participants (e.g. selecting for a political belief) might show stronger effects. Additionally, looking at the grouped participant data in Figure 5 suggests that the format of the trials might be too direct and putting participants on the spot. The "Unreported" category, although underpowered, is the only category in which the hypotheses were borne out significantly. Changing the dependent measure to a less intense probe — e.g. *what does your neighbor/the average American believe?* — has been shown to elicit stereotypical judgments in psychological literature (e.g. Chang & Demyan, 2007). This has the potential to elicit more honest or revelatory judgments from participants. Another direction for future research includes looking at names of both gender, avoiding the often-made equation between blackness and maleness seen in literature.

This norming study provides an encouraging start to a larger project, although not providing the anticipated results. Further building on both the experimental paradigm established here and the data allows for a more informed and directed approach in both eliciting and measuring stereotypical beliefs and expectations with regards to the strong alternatives of a pragmatic scale.

IV. References

Chang, D. F., & Demyan, A. L. (2007). Teachers' stereotypes of Asian, Black, and White students. *School Psychology Quarterly*, *22*(2), 91.

Collins, P. H. (2004). *Black sexual politics: African Americans, gender, and the new racism*. Routledge.

Davenport, L. D. (2018). *Politics Beyond Black and White: Biracial Identity and Attitudes in America*. Cambridge University Press.

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint−based approach. *Cognitive science*, *39*(4), 667-710.

Flores, N., & Rosa, J. (2019). Bringing race into second language acquisition. *The Modern Language Journal*, *103*, 145-151.

Geurts, B. (2010). *Quantity implicatures*. Cambridge University Press.

Golash-Boza, T. (2016). A critical and comprehensive sociological theory of race and racism. *Sociology of Race and Ethnicity*, *2*(2), 129-141.

Inoue, M. (2006). *Vicarious language: Gender and linguistic modernity in Japan* (Vol. 11). Univ of California Press.

Moya, P., & Markus, H. R. (2011). Doing race: A conceptual overview. *Doing Race: 21 Essays for the 21st Century*, 1-102.

Nieuwland, M. S., & Van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of cognitive neuroscience*, *18*(7), 1098-1111.

Politzer-Ahles, S., & Fiorentino, R. (2013). The realization of scalar inferences: Context sensitivity without processing cost. *PloS one*, *8*(5).

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., & Choi, Y. (2019). Social Bias Frames: Reasoning about Social and Power Implications of Language. *arXiv preprint arXiv:1911.03891*.

Stelter, M., & Degner, J. (2018). Recognizing Emily and Latisha: Inconsistent Effects of Name Stereotypicality on the Other-Race Effect. *Frontiers in psychology*, *9*, 486.