Pre-Analysis Plan for: What Should We Be Allowed to Post? Citizens' Preferences for Online Hate Speech Regulation

Simon Munzert (Hertie School of Governance)
Richard Traunmüller (University of Mannheim & Goethe University Frankfurt)
Andrew Guess (Princeton University)
Pablo Barberá (University of Southern California)
JungHwan Yang (University of Illinois at Urbana-Champaign)

July 26, 2019

Contents

1 Summary

This document describes a pre-analysis plan for a combined framing, priming, and vignette experiment embedded in an two-country online survey that examines citizen preferences towards hate speech regulation under various conditions as well as downstream consequences of hate speech exposure. We construct vignettes in forms of social media posts, mimicking actual cases of hate speech to bolster external validity, that vary along various dimensions of hate speech regulation, such as sender as well as target characteristics, act of speech, and targets' reaction. Respondents are asked to judge the posts with regards to actions that should be taken by the platform providers and other consequences the sender of hate speech should face. The tasks are embedded in different frames to test whether respondents can be framed to take a rather firm or soft stance towards hateful content. Finally, downstream consequences of exposure to hate speech content are examined. To that end, the entire vignette setup itself is considered a treatment. We then test how it affects peoples attitudes towards the need for hate speech regulation on the one hand and their willingness to express potentially controversial preferences on the other. We plan to publish the results in two papers.

2 Hypotheses

In a world that is becoming increasingly culturally diverse and digitally connected, "hate speech" has grown into a central concern across the globe. Next to polluting the quality of public discourse, "hate speech" is linked to detrimental effects on mental health and even violent inter-group conflict. Yet, whether and how to restrict speech that is considered offensive or promotes hate toward particular groups is highly contentious. Next to struggles over definitions of what constitutes "hate speech" in the first place, considerable disagreement concerns the adequate regulatory response to discriminatory speech: should "hate speech" be discouraged by social pressures alone or prohibited by law? Answering this question is difficult because positions for and against the restriction of hate speech are rooted in conflicting principles of freedom and equality. In this study, we plan to shed light on citizens' preferences for online hate speech sensitivity and regulation using evidence from vignette experiments.

2.1 Hate speech sensitivity and preferences for regulation conditional on content and context

A priori, we believe that sensitivity for controversial content as well as preferences towards hate speech regulation are not flat but conditional. We hypothesize that a variety of content- and context-specific factors influence citizens' perceptions of the offensiveness and hatefulness of online content, and also shape preferences for action that should be taken with regards to hate speech content and by whom. The following list summarizes our conjectures:

1. **Issues matter.** Respondents will not be agnostic towards the issue addressed by the controversial content. We expect to observe differences in both the sensitivity for controversial content as well as preferences towards hate speech regulation across four issues involving Muslim immigrants, women, the ideological Left, and the ideological Right. In particular, we expect that controversial content is considered relatively more hateful and offensive when the issue involves treatment of minorities, such as Muslim immigrants, and relatively less hateful and offensive when it involves treatment of broad groups, such as political movements.

2. **Target/sender identities matter.** Respondents will be more prone to judge the controversial content as hateful and offensive and to penalize it harder when the target is not anonymous or when the sender is anonymous.

3. **Addressing scope matters:** Respondents will be most prone to judge the controversial content as hateful and offensive and to penalize it harder when the target is directly addressed ("You..."). Proneness is expected to decrease with a narrowing of the scope: ("You..." > "All..." > "Most..." > "Extreme...").

4. **Message context matters:** Respondents will be most prone to judge the controversial content as hateful and offensive and to penalize it harder when the immediate context, that is the target's original message to which the sender replies, expresses an identification with ("I'm a...) rather than a mere support of ("I support") the target group.

5. **Content severity matters:** Respondents will be most prone to judge the controversial content as hateful and offensive and to penalize it harder when the target group is threatened with violence. Proneness is expected to decrease along the following content categories: violence > insult > vilification > discrimination. Moreover, the extreme

version of each category will evoke a more sensitive and punishing reaction than the moderate one.

6. **Target reactions matter:** Respondents will be most prone to judge the controversial content as hateful and offensive and to penalize it harder when the target does not react to the sender's message, thereby not anticipating any other action. Proneness is expected to decrease along the following content categories: none > appealing to norms > platform action > counter-aggression.

2.2 Hate speech sensitivity and preferences for regulation conditional on respondent characteristics

Moreover, we believe that hate speech sensitivity and preferences for regulation are conditional on respondent characteristics. In particular, we hypothesize that respondents will be biased in favor of their own group identity. In the context of our experiment, this implies:

1. If respondents share the gender, religious, and/or ideological identity with the target, they will be more prone to judge the controversial content as hateful and offensive and to penalize it harder.

2. If the respondents share the gender, religious, and/or ideological identity with the sender, they will be less prone to judge the controversial content as hateful and offensive and to penalize it less hard.

3. If respondents do not share the gender, religious, and/or ideological identity with the target, they will be less prone to judge the controversial content as hateful and offensive and to penalize it less hard.

4. If respondents do not share the gender, religious, and/or ideological identity with the sender, they will be more prone to judge the controversial content as hateful and offensive and to penalize it harder.

In addition to these considerations, we will test for heterogeneous effects in the above hypotheses in various subgroups, which are motivated by our reading of the literature. A global survey has recently found strong variation to the extent citizens in 64 countries support free expression (Wike and Simmons 2015). On an index from 0 to 8 (least to most supportive of free expression), US citizens ranked highest with a mean of 5.73, Germans much lower with 4.34. Preferences for free speech also differ across socio-demographics and social identity (Lalonde, Doan and Patterson 2000; Gross and Kinder 1998; Chong 2006). Men rate freedom of speech more important than women (Downs and Cowan 2012). Older people are less willing to censor hate speech than younger people (Lambe 2004). Blacks do not differ from Whites in their preferences for hate speech regulation (Gross and Kinder 1998; Chong 2006).

Attitudes toward free speech and its regulation are structured in large parts along political ideology (Lalonde et al. 2000; Gross and Kinder 1998; Chong 2006) as well as more fundamental psychological value orientations (Lalonde et al. 2000). Individualism is related to higher support for free speech (Downs and Cowan 2012). Social dominance orientation is positively related to the acceptance of hate speech (Bilewicz, Soral, Marchlewska and Winiewski 2017), whereas right-wing authoritarianism is positively related to hate-speech restriction (Bilewicz et al. 2017) and negatively related to the importance of free speech (Downs and Cowan 2012). Concern for political correctness is associated with more liberal beliefs and ideologies and with less right-wing authoritarianism (Strauts and Blanton 2015). Harell (2010) shows that exposure to racial and ethnic diversity in ones social networks decreases political tolerance of racist speech while simultaneously having a positive effect on political tolerance of other types of objectionable speech.

One particular challenge to the study of citizens preferences for free speech and its regulation is the inherent complexity and conditionality of the norms of free speech. A classic result in research on political tolerance is that people "express strong endorsement of the general principles of free expression and great reluctance to sustain these principles when asked to apply them to noxious groups" (Marcus, Sullivan, Theiss-Morse and Wood 1995). In particular, there seems to be an interaction between political ideology and content of speech acts. Suedfeld, Steel and Schmidt (1994) suggest that liberals are more likely to support the censorship of racist, sexist, and homophobic messages whereas political conservatives are more likely to support the censorship of pornography and offensive content regarding religious faith and conservative values. While Fisher, Lilie, Evans, Hollon, Sands, Depaul, Brady, Lindbom, Judd, Miller et al. (1999) find that support for censorship is generally higher for the political right than the left—regardless of content—, they also document left support for politically correct censorship, especially on university campus. Apart from the content, preferences for free speech also depend on who speaks. Grant and Rudolph (2003) show that people give greater weight to free speech when they consider the speech of their most-liked group, and they give less weight to free speech when they consider the speech of their least-liked group. Lindner and Nosek (2009) found that the manipulation of the speakers ethnicity (Black American, White American, or Arab Muslim) did not alter speech protection. However, there is an interaction between the speaker and the content of the speech act. Respondents protected a White speaker more strongly than an Arab Muslim speaker when he criticized Americans. Conversely, respondents protected an Arab Muslim speaker more strongly than a White speaker when he criticized Arabs.

Based on our reading of the literature as well as our own considerations, we will test for heterogeneity of effects along the following respondent characteristics and traits:

1. Hate speech experience and preferences
2. Feeling towards discussing politics with others
3. Political interest
4. Political ideology
5. Political issue preferences
6. Free speech regulation preferences
7. Party preferences
8. Partisanship
9. Social media usage
10. Internet usage
11. Racial resentment
12. Gender
13. Age
14. Education
15. Religion

2.3 The role of governmental and civil action for hate speech preferences

We hypothesize that respondents' perceptions of controversial content as well as their preferences towards hate speech regulation are not completely robust to external input. They may depend on the societal and political context. Depending on this context, respondents can be framed to take a rather firm or soft stance towards hateful content. To test this claim, we embed the vignette task in a framing experiment. In the introductory text to the vignettes, respondents are not only introduced to the task but the task is motivated in two different ways. There is also a control group that does not receive any motivational information. This setup generates four major expectations:

1. If the task is motivated by **looming government regulation to censor offensive or hateful social media content protecting potential victims of hate speech**, respondents, sensitized for the interests of potential victims, will be more prone to support tougher actions by both platform providers and other actors than non-primed respondents.

2. If the task is motivated by **looming government regulation to censor offensive or hateful social media content protecting potential victims of hate speech**, respondents, sensitized for the interests of potential victims, will perceive the content shown as more offensive and hateful than non-primed respondents.

3. If the task is motivated by **civil rights groups advocating for the right for free speech and against censorship online**, respondents, sensitized for adverse effects of potential censorship, will be less prone to support tougher actions by both platform providers and other actors than non-primed respondents.

4. If the task is motivated by **civil rights groups advocating for the right for free speech and against censorship online**, respondents, sensitized for adverse effects of potential censorship, will perceive the content shown as less offensive and hateful than non-primed respondents.

2.4 Downstream consequences of hate speech exposure

Finally, we consider downstream consequences of exposure to hate speech content. Being exposed to eight more or less offensive or hateful messages and being asked to engage with this content, that is, the entire vignette setup itself, can be considered a treatment in and of itself that both affects people's attitudes towards the need for hate speech regulation on the one hand and their willingness to express potentially controversial preferences. In particular, we expect the following:

1. Exposure to hate speech content via the vignette task reduces people's propensity to support the idea that people should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people.

2. Exposure to hate speech content via the vignette task reduces people's propensity to support the idea that people can use the Internet without government censorship.

3. Exposure to hate speech content via the vignette task reduces people's propensity to support potentially controversial opinions.

Furthermore, we expect these hypothesized effects to be more pronounced among those who received the government regulation prime than those who received the free speech advocacy prime (see above).

3 Design

3.1 Logistics

We received a grant from the Faculty Activity Fund of the Hertie School of Governance, Berlin, to run this experiment as part of a broader study, funded by the the Volkswagen Foundation, on how media exposure affects public opinion. The experiment is embedded in two panel surveys fielded on initially about 1,500 respondents recruited for the YouGov U.S. Pulse panel and about 1,500 respondents recruited for the YouGov German Pulse panel, which enables tracking of people's web usage on desktop and mobile devices. The Pulse panel is a subset of YouGov's traditional survey panels, where respondents opt in to install tracking software on their devices. The wave in which the vignette experiment is embedded was launched in early 2019 in both Germany and the United States. **Data on the experiments were not made accessible to the researchers before the publication of this pre-analysis plan.**

In both surveys, panelists that installed the web tracking software RealityMine on their computers and cell phones agreed to participate in a "Politics and Media" study with multiple survey waves. Their participation was rewarded using YouGov's proprietary point system and included a bonus if the respondent completed all waves in order to disincentivize attrition. Participation was voluntary and respondents were able to opt-out from the web tracking part of the study at any point in time. Respondents were sampled using age, gender, party identification, and education quotas and then re-weighted in order to obtain a sample that is representative of the U.S. population on these characteristics.

3.2 Experimental setup

We implement our research design using three components:

1. An **experimental manipulation priming hate speech regulation or free speech advocacy** (versus a control group) (preceded by an attention check)
2. A set of **vignettes in combination with several outcome measures**
3. An **experimental randomization of the positioning of a question on preferences towards hate speech regulation and other sensitive issues** (before or after the vignette experiment)
4. Multiple **items for subgroup and effect heterogeneity analyses** (already part of the core survey)

3.2.1 Treatment primes

The vignettes are preceded by an introduction that describes the judgment task related to the vignettes. Respondents will be randomly assigned to one of three versions which frame the task differently. (see Figures 1 and 2):

1. a neutral version (version 1),
2. a government regulation prime (version 2), and
3. a free speech advocacy prime (version 3).

A content warning was added to each prime, pointing out that the contents of some of the messages may be unpleasant or repugnant. We inform respondents about the option to skip each of the vignettes.

To check whether respondents actually take the time to carefully read the fictitious hate speech law, we run an attention check just before the experimental manipulation (Berinsky,

Figure 1: Treatment primes, US version

US VERSION 1 (NEUTRAL)
In the following, you will see a couple of messages posted online by social media users. Some of these messages, marked with a red arrow, are potentially problematic. We want you to take a close look at these messages and then answer a few questions.
We also want to point out that the contents of some of these messages may be unpleasant or repugnant to you. If you do not want to see any more such messages, you can skip these questions without answering.

US VERSION 2 (GOVERNMENT REGULATION PRIME)
As you may have heard, the government is serious about tackling online hate speech. Potential victims of online hate speech should be protected. This means that a large number of social media messages containing offensive or hateful content will be deleted and prosecuted.
In the following, you will see a couple of messages posted online by social media users. Some of these messages, marked with a red arrow, are potentially problematic. We want you to take a close look at these messages and then answer a few questions.
We also want to point out that the contents of some of these messages may be unpleasant or repugnant to you. If you do not want to see any more such messages, you can skip these questions without answering.

US VERSION 3 (FREE SPEECH ADVOCACY PRIME)
As you may have heard, civil society organizations are struggling to counter censorship of content on the Net. The right to freedom of expression should be protected. This means that social media messages containing offensive or hateful content should not be deleted or prosecuted.
In the following, you will see a couple of messages posted online by social media users. Some of these messages, marked with a red arrow, are potentially problematic. We want you to take a close look at these messages and then answer a few questions.
We also want to point out that the contents of some of these messages may be unpleasant or repugnant to you. If you do not want to see any more such messages, you can skip these questions without answering.

Huber and Lenz 2012). In this check, respondents are asked to ignore the initial question (about smartphone ownership) and to just type 'read' into the open text field (see Figure 3). We will contrast the results between the full sample and the reduced sample of respondents who passed the attention check. The attention check is presented to all respondents irrespective of treatment status.

3.2.2 Vignette design and attributes

The vignettes are constructed in a way that mimics posts on a popular social media platform (here: Facebook). Irrelevant features of the message, such as time stamp or features to interact with it, are dropped. Only features that represent relevant attributes of the vignettes

Figure 2: Treatment primes, German version

GERMAN VERSION 1 (NEUTRAL)

Im Folgenden sehen Sie Nachrichten, die von Social-Media-Nutzern gepostet wurden. Einige dieser Nachrichten, markiert mit einem roten Pfeil, sind potentiell problematisch. Wir möchten, dass Sie sich diese Nachrichten genau ansehen und anschließend einige Fragen dazu beantworten.

Wir möchten Sie außerdem darauf hinweisen, dass die Inhalte einiger dieser Nachrichten möglicherweise unangenehm oder abstoßend auf Sie wirken könnten. Wenn Sie deshalb keine weiteren solchen Nachrichten sehen möchten, können Sie diese Fragen ohne zu antworten überspringen.

GERMAN VERSION 2 (GOVERNMENT REGULATION PRIME)

Wie Sie vielleicht gehört haben, bemüht sich die Regierung sehr ernsthaft Online-Hassrede zu bekämpfen. Potentielle Opfer von Online-Hassrede sollen so geschützt werden. Das bedeutet, dass eine große Anzahl an Social-Media-Nachrichten mit beleidigenden oder hasserfüllten Inhalten gelöscht und strafrechtlich verfolgt werden.

Im Folgenden sehen Sie Nachrichten, die von Social-Media-Nutzern gepostet wurden. Einige dieser Nachrichten, markiert mit einem roten Pfeil, sind potentiell problematisch. Wir möchten, dass Sie sich diese Nachrichten genau ansehen und anschließend einige Fragen dazu beantworten.

Wir möchten Sie außerdem darauf hinweisen, dass die Inhalte einiger dieser Nachrichten möglicherweise unangenehm oder abstoßend auf Sie wirken könnten. Wenn Sie deshalb keine weiteren solchen Nachrichten sehen möchten, können Sie diese Fragen ohne zu antworten überspringen.

GERMAN VERSION 3 (FREE SPEECH ADVOCACY PRIME)

Wie Sie vielleicht gehört haben, bemühen sich Bürgerrechtsorganisationen darum, der Zensur von Inhalten im Netz entgegenzutreten. Das Recht auf freie Meinungsäußerung soll so geschützt werden. Das bedeutet, dass Social-Media-Nachrichten mit beleidigenden oder hasserfüllten Inhalten nicht gelöscht oder strafrechtlich verfolgt werden sollten.

Im Folgenden sehen Sie Nachrichten, die von Social-Media-Nutzern gepostet wurden. Einige dieser Nachrichten, markiert mit einem roten Pfeil, sind potentiell problematisch. Wir möchten, dass Sie sich diese Nachrichten genau ansehen und anschließend einige Fragen dazu beantworten.

Wir möchten Sie außerdem darauf hinweisen, dass die Inhalte einiger dieser Nachrichten möglicherweise unangenehm oder abstoßend auf Sie wirken könnten. Wenn Sie deshalb keine weiteren solchen Nachrichten sehen möchten, können Sie diese Fragen ohne zu antworten überspringen.

Figure 3: Design of attention check before prime

ATTENTION CHECK BEFORE PRIME, US SURVEY

Many people own smart phones nowadays. How about you: Do you own one, and if yes, what type of smartphone? Specifically, we want to know whether you actually take your time to read the questions and follow our instructions. To demonstrate that you read this far, skip this question and just type read in the text field below.

○ Apple iPhone
○ Samsung Galaxy
○ Huawei Mate
○ Google Pixel
○ LG V40
○ Sony XPeria
○ Other: _______________
○ I do not own a smartphone

ATTENTION CHECK BEFORE PRIME, GERMAN SURVEY

Viele Leute besitzen heutzutage ein Smartphone. Wie ist das mit Ihnen? Besitzen Sie ein Smartphone, und wenn ja, welches? Genauer gesagt möchten wir von Ihnen wissen, ob Sie sich eigentlich die Zeit nehmen die Fragen zu lesen und den Anweisungen zu folgen. Um zu zeigen, dass Sie bis hierhin gelesen haben, tragen Sie bitte “gelesen” in das Feld “Anderes, und zwar” unten ein.

○ Apple iPhone
○ Samsung Galaxy
○ Huawei Mate
○ Google Pixel
○ LG V40
○ Sony XPeria
○ Anderes, und zwar: _______________
○ Ich besitze kein Smartphone.

are kept. These attributes cover issues, sender as well as target characteristics, and sender message’s and target message’s characteristics. Table 1 provides an overview of the attributes and attribute levels.

Tables 2 to 5 provide the detailed components of the messages across the four different issues/target groups (Muslim immigrants, women, ideological Left, ideological Right) for the US survey. Tables 6 to 9 provide the same information for the German survey. Tables 10 and 11 provide information on sender and target characteristics. The prenames and surnames were chosen based on lists of popular female and male, Muslim and non-Muslim names in the United States and Germany, respectively. The thumbnail images were taken from licence-free stock photo platforms. For the non-Muslim group, only whites were used. For the Muslim group, mostly people with dark complexion and, in part, headgear, were used. It is important to note that we use the visual and name characteristics as deliberately suggestive cues to implicitly signal in-group or out-group membership for the groups “Muslim”, “non-Muslim”, “woman”, and “no woman”. Also, we use two senders/targets who are anonymous (no gender or religion cues) but provide liberal and conservative cues. The liberal anonymous account is “Team Global”, featured with the rainbow flag. The conservative anonymous account is “Team USA” featured with the US flag (“Team Deutschland” with German flag in the German survey).

Figures 4 and 5 provide a random sample of vignettes used (the label IDs were not shown to the respondents).

Table 1: Vignette attributes and attribute levels

Attribute	**Attribute Level**
Issue ∈	• Muslim immigrants • Women • Ideological Left • Ideological Right
Target's identity ∈	• male/female • Muslim/non-Muslim • liberal anonymous/conservative anonymous
Target's initial message ∈	• identification with target group • support of target group
Sender's identity ∈	• male/female • Muslim/non-Muslim • liberal anonymous/conservative anonymous
Sender message's target ∈	• "You..." • "All..." • "Most..." • "Extreme..."
Sender's message ∈	• violence (moderate/extreme) • insult (moderate/extreme) • discrimination (moderate/extreme) • vilification (moderate/extreme)
Target's reaction ∈	• appealing to norms • counter-aggression • platform action (blocking, reporting) • none

Table 2: Muslim immigrant vignettes: message components, US survey

Target message	**Sender message's target**	**Sender message**	**Target reply**
I'm a proud Muslim immigrant! [identification]	Most Muslim immigrants	should be banned from coming into our country. [discrimination, moderate]	Please stop attacking others. [appealing to norms 1]
I support Muslim immigrants. [support]	All Muslim immigrants	should be deported. [discrimination, extreme]	Please stay civil. We all have far more in common with each other than things that divide us. [appealing to norms 2]
	Extremist Muslim immigrants	should be punched in the face. [violence, moderate]	You are so dumb! [counter-aggression 1]
	You	should be killed. [violence, extreme]	Shut up asshole! [counter-aggression 2]
		are dumb. [insult, moderate]	[target] has reported [sender]'s post. [platform action 1]
		are stupid animals. [insult, extreme]	[target] has blocked [sender]'s account. [platform action 2]
		are just different from us. [vilification, moderate]	none
		are genetically inferior to us. [vilification, extreme]	

Table 3: Women vignettes: message components, US survey

Target message	Sender message's target	Sender message	Target reply
I'm a proud feminist! [identification]	Most women	should not be allowed to serve in the army. [discrimination, moderate]	Please stop attacking others. [appealing to norms 1]
I support feminism. [support]	All women	should be caring mothers and not pursue a selfish career. [discrimination, extreme]	Please stay civil. We all have far more in common with each other than things that divide us. [appealing to norms 2]
	Extremist feminist women	should be punched in the face. [violence, moderate]	You are so dumb! [counter-aggression 1]
	You	should be killed. [violence, extreme]	Shut up asshole! [counter-aggression 2]
		are dumb. [insult, moderate]	[target] has reported [sender]'s post. [platform action 1]
		are stupid animals. [insult, extreme]	[target] has blocked [sender]'s account. [platform action 2]
		are just different from men. [vilification, moderate]	none
		are genetically inferior to men. [vilification, extreme]	

Table 4: Ideological Left vignettes: message components, US survey

Target message	Sender message's target	Sender message	Target reply
I'm a proud Democrat! [identification]	Most Democrats	should be stopped from spreading falsehoods. [discrimination, moderate]	Please stop attacking others. [appealing to norms 1]
I support Democrats. [support]	All Democrats	should be not allowed to vote. [discrimination, extreme]	Please stay civil. We all have far more in common with each other than things that divide us. [appealing to norms 2]
	Socialist Democrats	should be punched in the face. [violence, moderate]	You are so dumb! [counter-aggression 1]
	You	should be killed. [violence, extreme]	Shut up asshole! [counter-aggression 2]
		are dumb. [insult, moderate]	[target] has reported [sender]'s post. [platform action 1]
		are stupid animals. [insult, extreme]	[target] has blocked [sender]'s account. [platform action 2]
		are just not as clever as Republicans. [vilification, moderate]	none
		are genetically inferior to Republicans. [vilification, extreme]	

Table 5: Ideological Right vignettes: message components, US survey

Target message	Sender message's target	Sender message	Target reply
I'm a proud Republican! [identification]	Most Republicans	should be stopped from spreading falsehoods. [discrimination, moderate]	Please stop attacking others. [appealing to norms 1]
I support Republican. [support]	All Republicans	should be not allowed to vote. [discrimination, extreme]	Please stay civil. We all have far more in common with each other than things that divide us. [appealing to norms 2]
	Trumpist Republicans	should be punched in the face. [violence, moderate]	You are so dumb! [counter-aggression 1]
	You	should be killed. [violence, extreme]	Shut up asshole! [counter-aggression 2]
		are dumb. [insult, moderate]	[target] has reported [sender]'s post. [platform action 1]
		are stupid animals. [insult, extreme]	[target] has blocked [sender]'s account. [platform action 2]
		are just not as clever as Democrats. [vilification, moderate]	none
		are genetically inferior to Democrats. [vilification, extreme]	

Table 6: Muslim immigrant vignettes: message components, German survey

Target message	**Sender message's target**	**Sender message**	**Target reply**
Ich bin ein stolzer muslimischer Einwanderer! [identification]	Die meisten muslimischen Einwanderer	sollten daran gehindert werden, in unser Land zu kommen. [discrimination, moderate]	Bitte höre auf, andere anzugreifen. [appealing to norms 1]
Ich unterstütze muslimische Einwanderer. [support]	Alle muslimischen Einwanderer	sollten abgeschoben werden. [discrimination, extreme]	Bitte bleibe höflich. Wir haben alle mehr gemeinsam, als uns voneinander trennt. [appealing to norms 2]
	Extremistische muslimische Einwanderer	sollten eine reingehauen bekommen. [violence, moderate]	Du bist so blöd! [counter-aggression 1]
	Du	sollten abgeknallt werden. [violence, extreme]	Halt's Maul Arschloch! [counter-aggression 2]
		sind Idioten. [insult, moderate]	[target] hat [sender]s Nachricht angezeigt. [platform action 1]
		sind dumme Kreaturen. [insult, extreme]	[target] hat [sender]s Account blockiert. [platform action 2]
		sind einfach anders als wir. [vilification, moderate]	none
		sind uns genetisch unterlegen. [vilification, extreme]	

Table 7: Women vignettes: message components, German survey

Target message	Sender message's target	Sender message	Target reply
Ich bin ein stolzer Feminist! [identification]	Die meisten Frauen	sollten fürsorgliche Mütter sein und keine egoistische Karriere verfolgen. [discrimination, moderate]	Bitte höre auf, andere anzugreifen. [appealing to norms 1]
Ich unterstütze Feminismus. [support]	Alle Frauen	sollten nicht in der Bundeswehr dienen dürfen. [discrimination, extreme]	Bitte bleibe höflich. Wir haben alle mehr gemeinsam, als uns voneinander trennt. [appealing to norms 2]
	Extrem feministische Frauen	sollten eine reingehauen bekommen. [violence, moderate]	Du bist so blöd! [counter-aggression 1]
	Du	sollten abgeknallt werden. [violence, extreme]	Halt's Maul Arschloch! [counter-aggression 2]
		sind Idioten. [insult, moderate]	[target] hat [sender]s Nachricht angezeigt. [platform action 1]
		sind dumme Kreaturen. [insult, extreme]	[target] hat [sender]s Account blockiert. [platform action 2]
		sind einfach anders als Männer. [vilification, moderate]	none
		sind genetisch Männern unterlegen. [vilification, extreme]	

Table 8: Ideological Left vignettes: message components, German survey

Target message	**Sender message's target**	**Sender message**	**Target reply**
Ich bin ein stolzes Mitglied der Grünen! [identification]	Die meisten Grünen	sollten davon abgehalten werden, Lügen zu verbreiten. [discrimination, moderate]	Bitte höre auf, andere anzugreifen. [appealing to norms 1]
Ich unterstütze die Grüne. [support]	Alle Grünen	sollten nicht wählen dürfen. [discrimination, extreme]	Bitte bleibe höflich. Wir haben alle mehr gemeinsam, als uns voneinander trennt. [appealing to norms 2]
	Linksextreme Grüne	sollten eine reingehauen bekommen. [violence, moderate]	Du bist so blöd! [counter-aggression 1]
	Du	sollten abgeknallt werden. [violence, extreme]	Halt's Maul Arschloch! [counter-aggression 2]
		sind Idioten. [insult, moderate]	[target] hat [sender]s Nachricht angezeigt. [platform action 1]
		sind dumme Kreaturen. [insult, extreme]	[target] hat [sender]s Account blockiert. [platform action 2]
		sind einfach nicht so schlau wie AfDler. [vilification, moderate]	none
		sind AfDlern genetisch unterlegen. [vilification, extreme]	

Table 9: Ideological Right vignettes: message components, German survey

Target message	**Sender message's target**	**Sender message**	**Target reply**
Ich bin ein stolzes Mitglied der AfD! [identification]	Die meisten AfDler	sollten davon abgehalten werden, Lügen zu verbreiten. [discrimination, moderate]	Bitte höre auf, andere anzugreifen. [appealing to norms 1]
Ich unterstütze die AfD. [support]	Alle AfDler	sollten nicht wählen dürfen. [discrimination, extreme]	Bitte bleibe höflich. Wir haben alle mehr gemeinsam, als uns voneinander trennt. [appealing to norms 2]
	Rechtsextreme AfDler	sollten eine reingehauen bekommen. [violence, moderate]	Du bist so blöd! [counter-aggression 1]
	Du	sollten abgeknallt werden. [violence, extreme]	Halt's Maul Arschloch! [counter-aggression 2]
		sind Idioten. [insult, moderate]	[target] hat [sender]s Nachricht angezeigt. [platform action 1]
		sind dumme Kreaturen. [insult, extreme]	[target] hat [sender]s Account blockiert. [platform action 2]
		sind einfach nicht so schlau wie Grüne. [vilification, moderate]	none
		sind Grünen genetisch unterlegen. [vilification, extreme]	

3.2.3 Construction of vignette universe

To construct the vignette universe, which is later used to sample from to generate the vignette decks, we create a data frame of all combinations of all attribute levels and later exclude observations that are illogical or implausible. The artificial variation across sender and message characteristics is reduced: For instance, in the case of two vignettes that are exactly equal but differ only on the sender's name (e.g., a female Muslim named Fatima Abad vs. a female Muslim named Nazia Karimi), one of the vignettes is randomly discarded. Furthermore, the following rules are implemented:

- Target and sender must be different persons.
- A sender message addressing the out-group target (e.g., a male target in hate speech addressing women or a non-Muslim target in hate speech addressing Muslims) directly ("You...") can only be a message of class *violence* or *insult*, not *discrimination* or *vilification*.
- If the message is addressing the target directly ("You..."), some messages have to be grammatically adapted (e.g., replace "You are stupid animals" with "You are a stupid animal").
- The sender has to be of type out-group. That is, in hate speech targeting Muslims, the sender cannot be Muslim, in hate speech targeting women, the sender cannot be female, in hate speech targeting the ideological Left, the sender cannot be ideologically left, and in hate speech targeting the ideological Right, the sender cannot be ideologically right.

Keeping only vignettes that comply with these rules gives us a set of 40,960 unique vignettes.

3.2.4 Construction of vignette decks

In the next step, we construct the vignette decks, which consist of eight individual vignettes each. The goal is to achieve approximate balance of all attribute levels in the sample of vignettes used in the surveys, to avoid repeated use of persons in individual decks as often as possible, and to maximize variation of attribute levels in individual decks. To that end, we define the a priori distribution of features within single decks ("stratification"). The a priori distribution is:

- topic: 2 Muslim immigrant, 2 woman, 2 ideological Left, 2 ideological Right
- gender, sender: 3 male, 5 female
- religion, sender: 3 non-Muslim, 5 Muslim
- ideology, sender: 6 unknown, 1 conservative, 1 liberal

Table 10: Sender and target characteristics, US survey

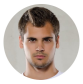
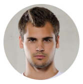
Ideology	Religion cue	Gender	Prename	Surname	Thumbnail
unknown	Muslim	female	Fatima	Abad	
			Nazia	Karimi	
			Saba	Malek	
			Zainab	Omer	
		male	Amir	Rahman	
			Muhammad	Nazir	
			Nadeem	Shakir	
			Rashid	Farra	
	non-Muslim	female	Anna	Krueger	
			Lisa	Mueller	
			Laura	Harris	
			Carolyn	Clark	
		male	Paul	Miller	
			Mark	Schmitt	
			Lucas	Baker	
			Florian	Smith	
liberal	unknown	unknown	Team	Global	
conservative			Team	USA	

Table 11: Sender and target characteristics, German survey

Ideology	Religion cue	Gender	Prename	Surname	Thumbnail
unknown	Muslim	female	Fatima	Abad	
			Nazia	Karimi	
			Saba	Malek	
			Zeynep	Omer	
		male	Amir	Rahman	
			Muhammad	Nazir	
			Nadeem	Shakir	
			Rashid	Farra	
	non-Muslim	female	Anna	Schneider	
			Lisa	Meier	
			Laura	Fischer	
			Carolin	Weber	
		male	Paul	Wagner	
			Mark	Schmidt	
			Lukas	Becker	
			Florian	Schulz	
liberal	unknown	unknown	Team	Global	
conservative			Team	Deutschland	

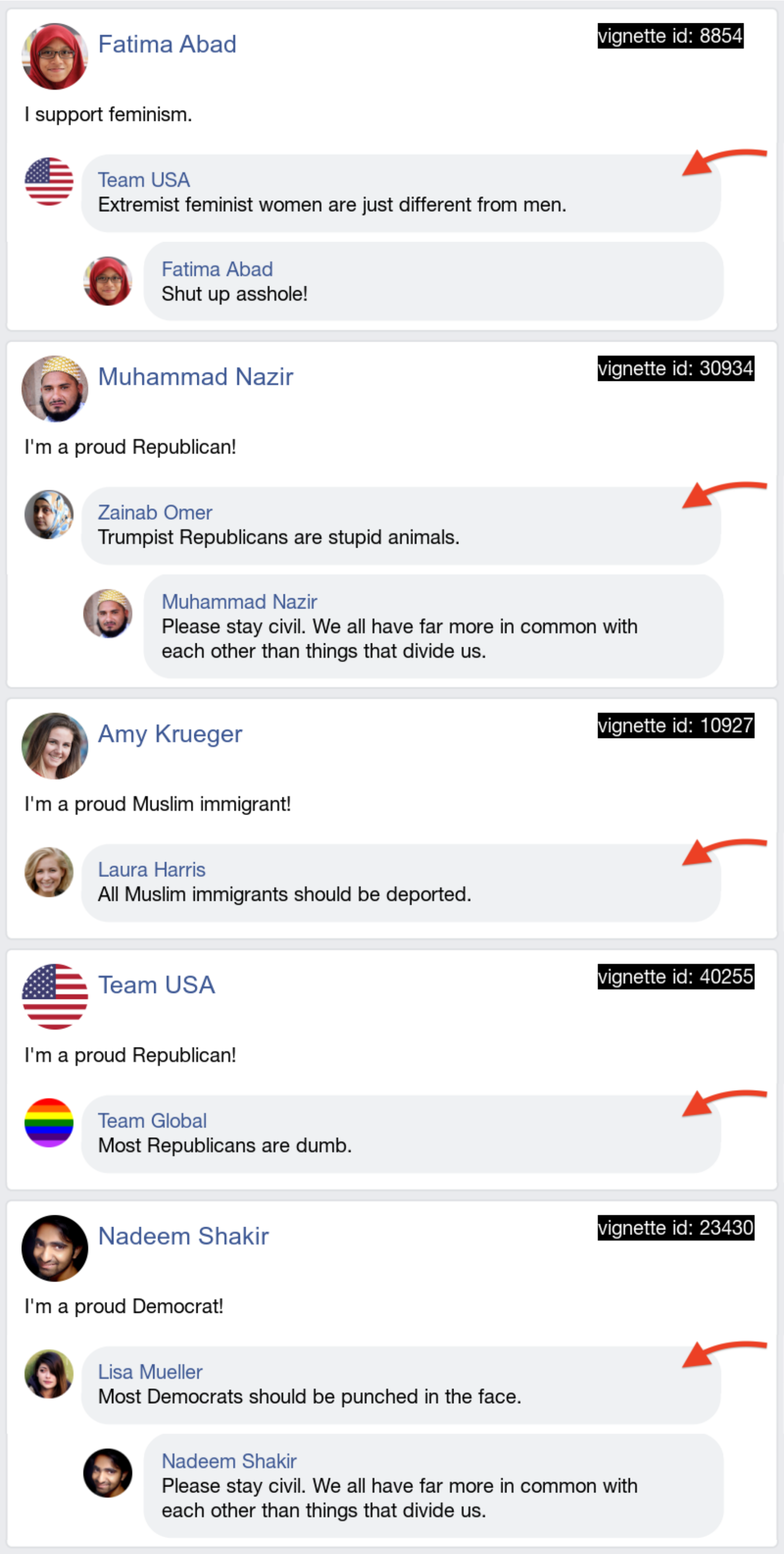
- target group category: 2 most, 2 all, 2 extreme, 2, you
- target message category: 4 proud, 4 support
- sender message category: 2 discrimination, 2 insult, 2 vilification, 2 violence
- target reply category: 2 appealing to norms, 2 counter-aggression, 2 platform action, 2 none

Note that, a priori, we over-sampled female and Muslim senders because they are later replaced with male or non-Muslim senders in women and Muslim immigrant vignettes (see rule above: sender has to be of type out-group).

Based on this distribution of features, we sample actual patterns of vignette sets. In the next step, we randomly draw from all those vignettes from the vignette universe that match the sampled patterns. These rules do not guarantee yet that the occurrence of sender/target character duplicates within individual decks is minimized. Therefore, we generate a gross sample of vignette decks and keep only those with 13 or more unique characters among senders and targets.

Overall, the described procedure generates a largely balanced representation of the attribute levels in the sample of vignette decks (see Figures 6 to 11).

Figure 4: Sample of vignettes



Fatima Abad

vignette id: 8854

I support feminism.

Team USA
Extremist feminist women are just different from men.

Fatima Abad
Shut up asshole!

Muhammad Nazir

vignette id: 30934

I'm a proud Republican!

Zainab Omer
Trumpist Republicans are stupid animals.

Muhammad Nazir
Please stay civil. We all have far more in common with each other than things that divide us.

Amy Krueger

vignette id: 10927

I'm a proud Muslim immigrant!

Laura Harris
All Muslim immigrants should be deported.

Team USA

vignette id: 40255

I'm a proud Republican!

Team Global
Most Republicans are dumb.

Nadeem Shakir

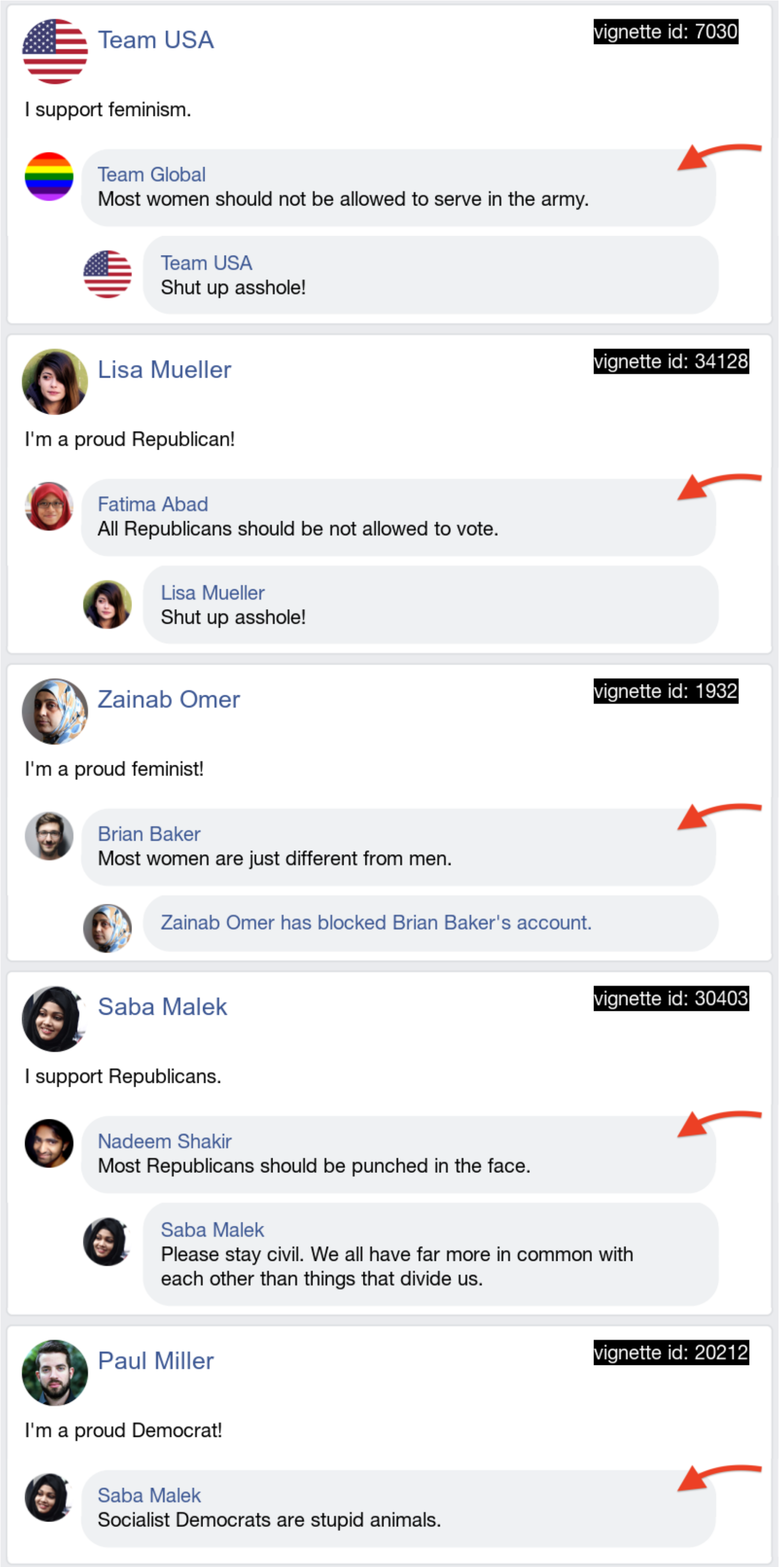
vignette id: 23430

I'm a proud Democrat!

Lisa Mueller
Most Democrats should be punched in the face.

Nadeem Shakir
Please stay civil. We all have far more in common with each other than things that divide us.

Figure 5: Sample of vignettes, *continued*



Team USA

vignette id: 7030

I support feminism.

Team Global
Most women should not be allowed to serve in the army.

Team USA
Shut up asshole!

Lisa Mueller

vignette id: 34128

I'm a proud Republican!

Fatima Abad
All Republicans should be not allowed to vote.

Lisa Mueller
Shut up asshole!

Zainab Omer

vignette id: 1932

I'm a proud feminist!

Brian Baker
Most women are just different from men.

Zainab Omer has blocked Brian Baker's account.

Saba Malek

vignette id: 30403

I support Republicans.

Nadeem Shakir
Most Republicans should be punched in the face.

Saba Malek
Please stay civil. We all have far more in common with each other than things that divide us.

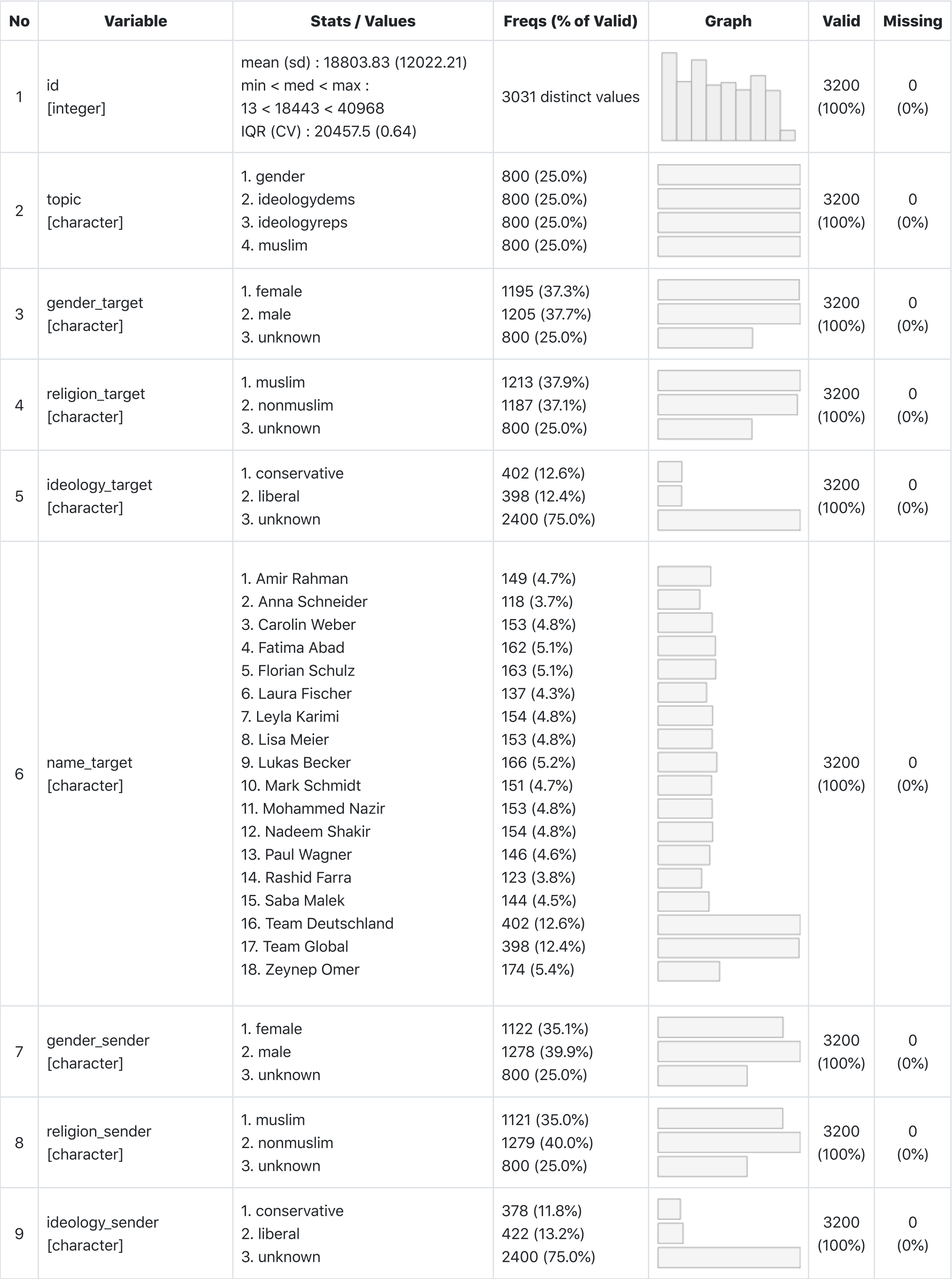
Paul Miller

vignette id: 20212

I'm a proud Democrat!

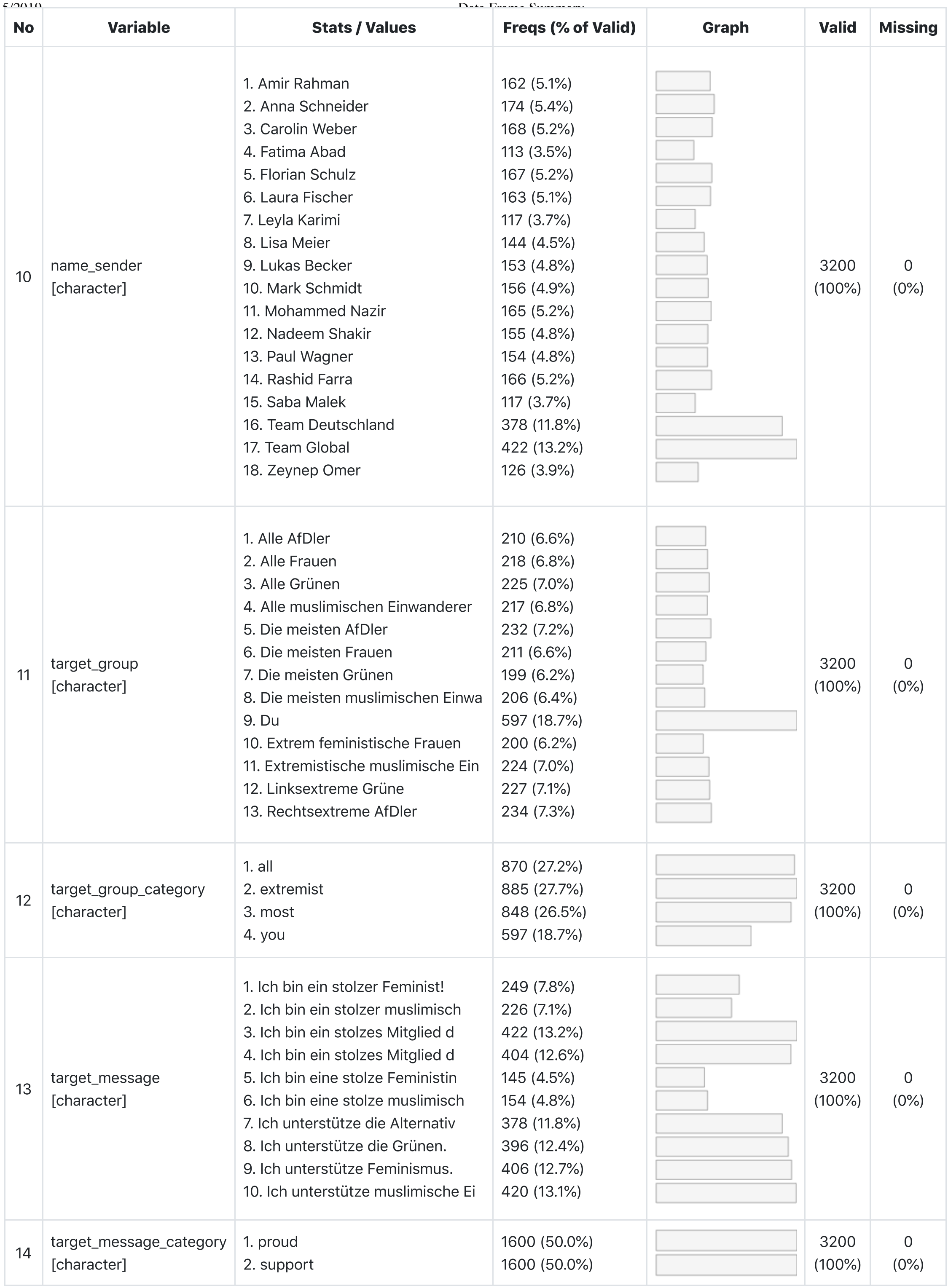
Saba Malek
Socialist Democrats are stupid animals.

Figure 6: Balance of vignette attributes in gross sample (US survey)



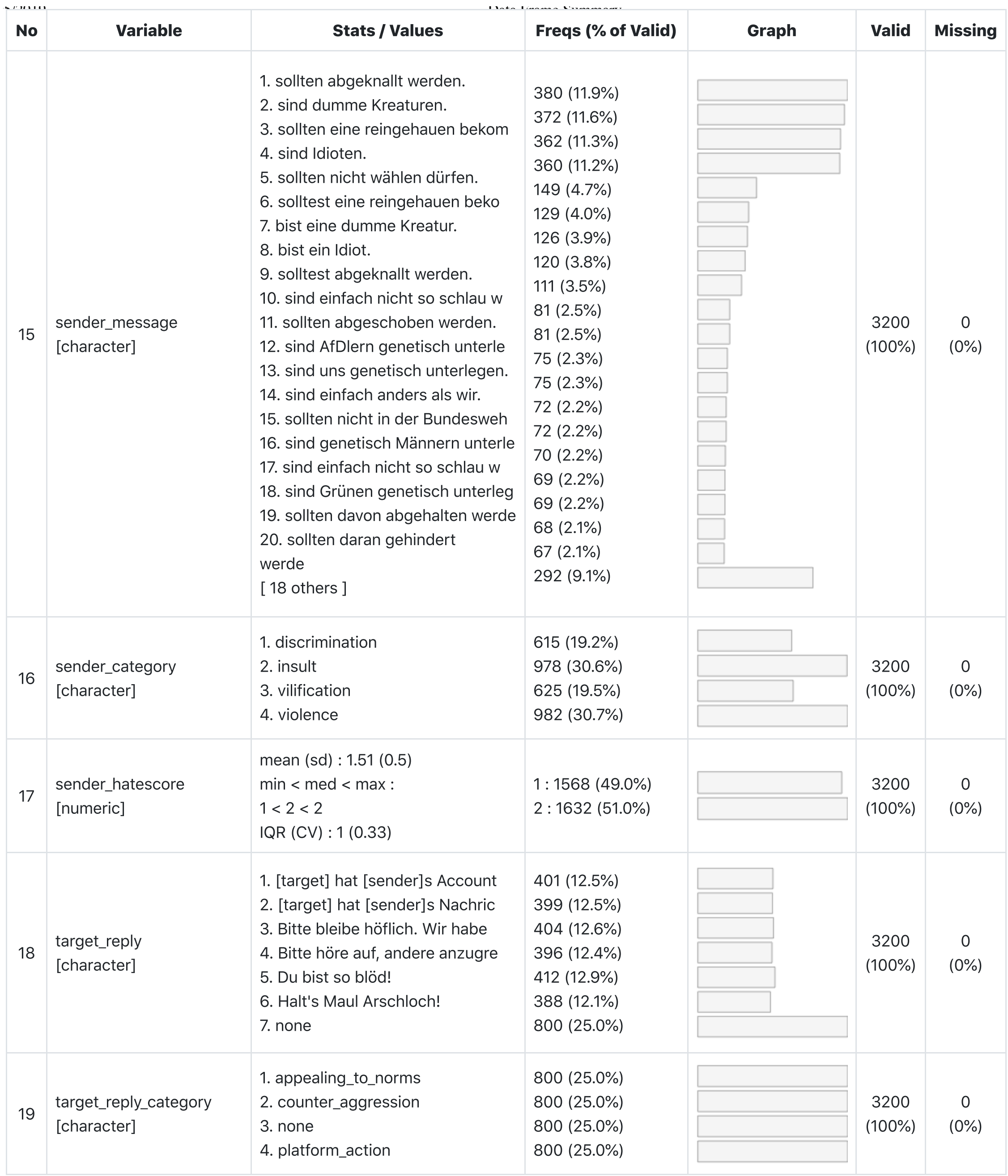
No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	id [integer]	mean (sd) : 18803.83 (12022.21) min < med < max : 13 < 18443 < 40968 IQR (CV) : 20457.5 (0.64)	3031 distinct values		3200 (100%)	0 (0%)
2	topic [character]	1. gender 2. ideologydems 3. ideologyreps 4. muslim	800 (25.0%) 800 (25.0%) 800 (25.0%) 800 (25.0%)		3200 (100%)	0 (0%)
3	gender_target [character]	1. female 2. male 3. unknown	1195 (37.3%) 1205 (37.7%) 800 (25.0%)		3200 (100%)	0 (0%)
4	religion_target [character]	1. muslim 2. nonmuslim 3. unknown	1213 (37.9%) 1187 (37.1%) 800 (25.0%)		3200 (100%)	0 (0%)
5	ideology_target [character]	1. conservative 2. liberal 3. unknown	402 (12.6%) 398 (12.4%) 2400 (75.0%)		3200 (100%)	0 (0%)
6	name_target [character]	1. Amir Rahman 2. Anna Schneider 3. Carolin Weber 4. Fatima Abad 5. Florian Schulz 6. Laura Fischer 7. Leyla Karimi 8. Lisa Meier 9. Lukas Becker 10. Mark Schmidt 11. Mohammed Nazir 12. Nadeem Shakir 13. Paul Wagner 14. Rashid Farra 15. Saba Malek 16. Team Deutschland 17. Team Global 18. Zeynep Omer	149 (4.7%) 118 (3.7%) 153 (4.8%) 162 (5.1%) 163 (5.1%) 137 (4.3%) 154 (4.8%) 153 (4.8%) 166 (5.2%) 151 (4.7%) 153 (4.8%) 154 (4.8%) 146 (4.6%) 123 (3.8%) 144 (4.5%) 402 (12.6%) 398 (12.4%) 174 (5.4%)		3200 (100%)	0 (0%)
7	gender_sender [character]	1. female 2. male 3. unknown	1122 (35.1%) 1278 (39.9%) 800 (25.0%)		3200 (100%)	0 (0%)
8	religion_sender [character]	1. muslim 2. nonmuslim 3. unknown	1121 (35.0%) 1279 (40.0%) 800 (25.0%)		3200 (100%)	0 (0%)
9	ideology_sender [character]	1. conservative 2. liberal 3. unknown	378 (11.8%) 422 (13.2%) 2400 (75.0%)		3200 (100%)	0 (0%)

Figure 7: Balance of vignette attributes in gross sample (US survey), *continued*



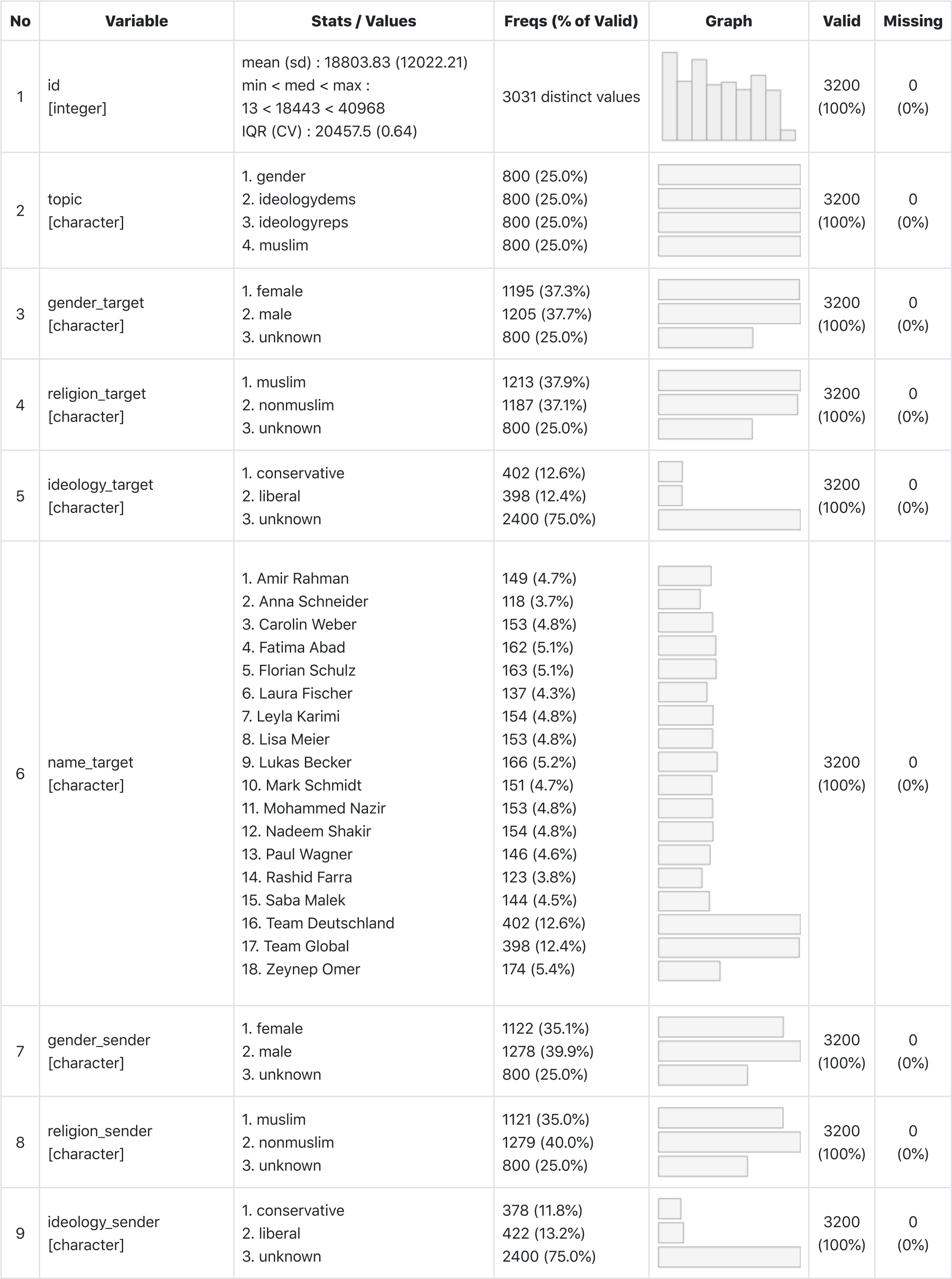
No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
10	name_sender [character]	1. Amir Rahman 2. Anna Schneider 3. Carolin Weber 4. Fatima Abad 5. Florian Schulz 6. Laura Fischer 7. Leyla Karimi 8. Lisa Meier 9. Lukas Becker 10. Mark Schmidt 11. Mohammed Nazir 12. Nadeem Shakir 13. Paul Wagner 14. Rashid Farra 15. Saba Malek 16. Team Deutschland 17. Team Global 18. Zeynep Omer	162 (5.1%) 174 (5.4%) 168 (5.2%) 113 (3.5%) 167 (5.2%) 163 (5.1%) 117 (3.7%) 144 (4.5%) 153 (4.8%) 156 (4.9%) 165 (5.2%) 155 (4.8%) 154 (4.8%) 166 (5.2%) 117 (3.7%) 378 (11.8%) 422 (13.2%) 126 (3.9%)		3200 (100%)	0 (0%)
11	target_group [character]	1. Alle AfDler 2. Alle Frauen 3. Alle Grünen 4. Alle muslimischen Einwanderer 5. Die meisten AfDler 6. Die meisten Frauen 7. Die meisten Grünen 8. Die meisten muslimischen Einwa 9. Du 10. Extrem feministische Frauen 11. Extremistische muslimische Ein 12. Linksextreme Grüne 13. Rechtsextreme AfDler	210 (6.6%) 218 (6.8%) 225 (7.0%) 217 (6.8%) 232 (7.2%) 211 (6.6%) 199 (6.2%) 206 (6.4%) 597 (18.7%) 200 (6.2%) 224 (7.0%) 227 (7.1%) 234 (7.3%)		3200 (100%)	0 (0%)
12	target_group_category [character]	1. all 2. extremist 3. most 4. you	870 (27.2%) 885 (27.7%) 848 (26.5%) 597 (18.7%)		3200 (100%)	0 (0%)
13	target_message [character]	1. Ich bin ein stolzer Feminist! 2. Ich bin ein stolzer muslimisch 3. Ich bin ein stolzes Mitglied d 4. Ich bin ein stolzes Mitglied d 5. Ich bin eine stolze Feministin 6. Ich bin eine stolze muslimisch 7. Ich unterstütze die Alternativ 8. Ich unterstütze die Grünen. 9. Ich unterstütze Feminismus. 10. Ich unterstütze muslimische Ei	249 (7.8%) 226 (7.1%) 422 (13.2%) 404 (12.6%) 145 (4.5%) 154 (4.8%) 378 (11.8%) 396 (12.4%) 406 (12.7%) 420 (13.1%)		3200 (100%)	0 (0%)
14	target_message_category [character]	1. proud 2. support	1600 (50.0%) 1600 (50.0%)		3200 (100%)	0 (0%)

Figure 8: Balance of vignette attributes in gross sample (US survey), *continued*



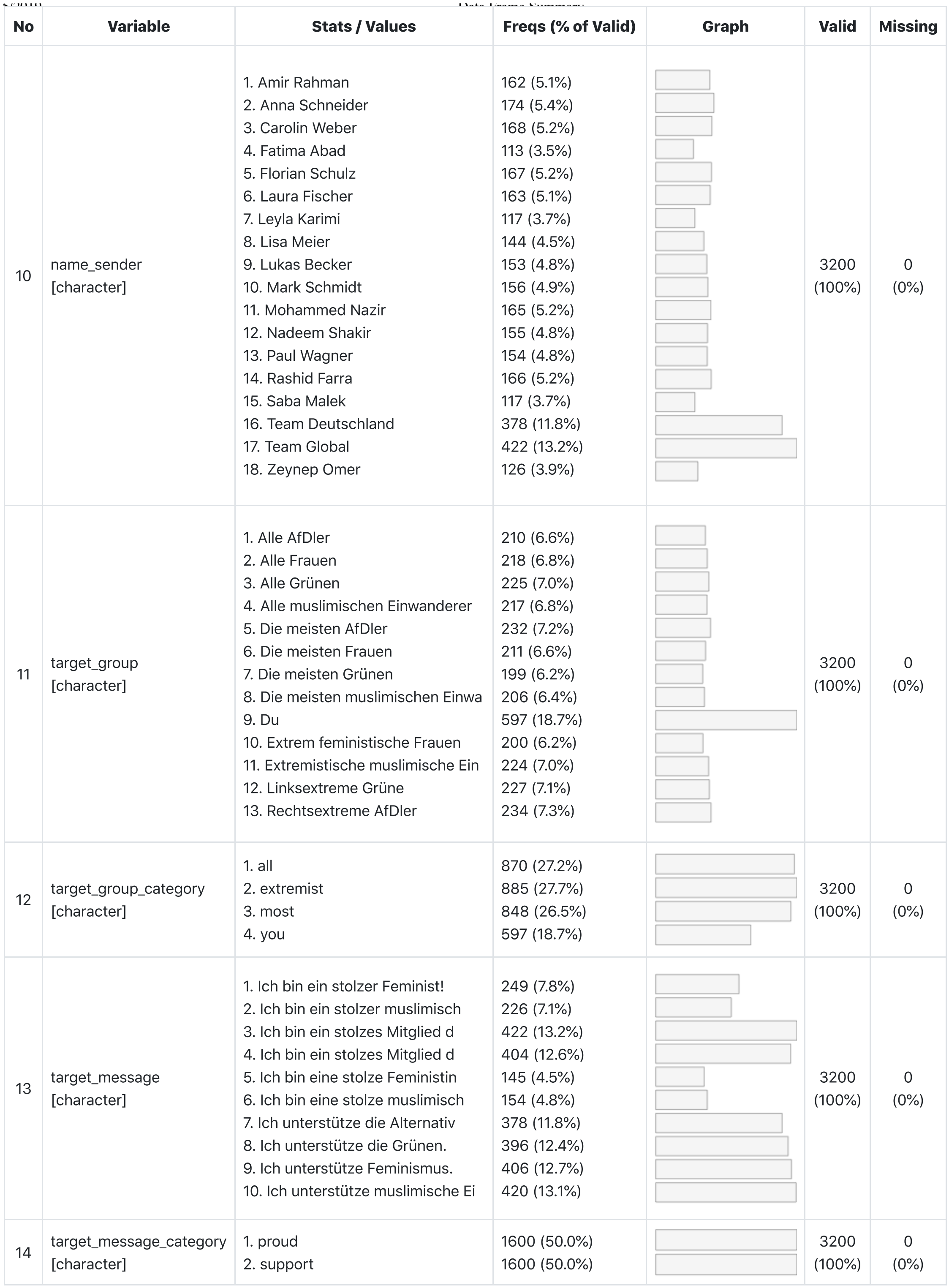
No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
15	sender_message [character]	1. sollten abgeknallt werden. 2. sind dumme Kreaturen. 3. sollten eine reingehauen bekom 4. sind Idioten. 5. sollten nicht wählen dürfen. 6. solltest eine reingehauen beko 7. bist eine dumme Kreatur. 8. bist ein Idiot. 9. solltest abgeknallt werden. 10. sind einfach nicht so schlau w 11. sollten abgeschoben werden. 12. sind AfDlern genetisch unterle 13. sind uns genetisch unterlegen. 14. sind einfach anders als wir. 15. sollten nicht in der Bundesweh 16. sind genetisch Männern unterle 17. sind einfach nicht so schlau w 18. sind Grünen genetisch unterleg 19. sollten davon abgehalten werde 20. sollten daran gehindert werde [18 others]	380 (11.9%) 372 (11.6%) 362 (11.3%) 360 (11.2%) 149 (4.7%) 129 (4.0%) 126 (3.9%) 120 (3.8%) 111 (3.5%) 81 (2.5%) 81 (2.5%) 75 (2.3%) 75 (2.3%) 72 (2.2%) 72 (2.2%) 70 (2.2%) 69 (2.2%) 69 (2.2%) 68 (2.1%) 67 (2.1%) 292 (9.1%)		3200 (100%)	0 (0%)
16	sender_category [character]	1. discrimination 2. insult 3. vilification 4. violence	615 (19.2%) 978 (30.6%) 625 (19.5%) 982 (30.7%)		3200 (100%)	0 (0%)
17	sender_hatescore [numeric]	mean (sd) : 1.51 (0.5) min < med < max : 1 < 2 < 2 IQR (CV) : 1 (0.33)	1 : 1568 (49.0%) 2 : 1632 (51.0%)		3200 (100%)	0 (0%)
18	target_reply [character]	1. [target] hat [sender]s Account 2. [target] hat [sender]s Nachric 3. Bitte bleibe höflich. Wir habe 4. Bitte höre auf, andere anzugre 5. Du bist so blöd! 6. Halt's Maul Arschloch! 7. none	401 (12.5%) 399 (12.5%) 404 (12.6%) 396 (12.4%) 412 (12.9%) 388 (12.1%) 800 (25.0%)		3200 (100%)	0 (0%)
19	target_reply_category [character]	1. appealing_to_norms 2. counter_aggression 3. none 4. platform_action	800 (25.0%) 800 (25.0%) 800 (25.0%) 800 (25.0%)		3200 (100%)	0 (0%)

Figure 9: Balance of vignette attributes in gross sample (German survey)



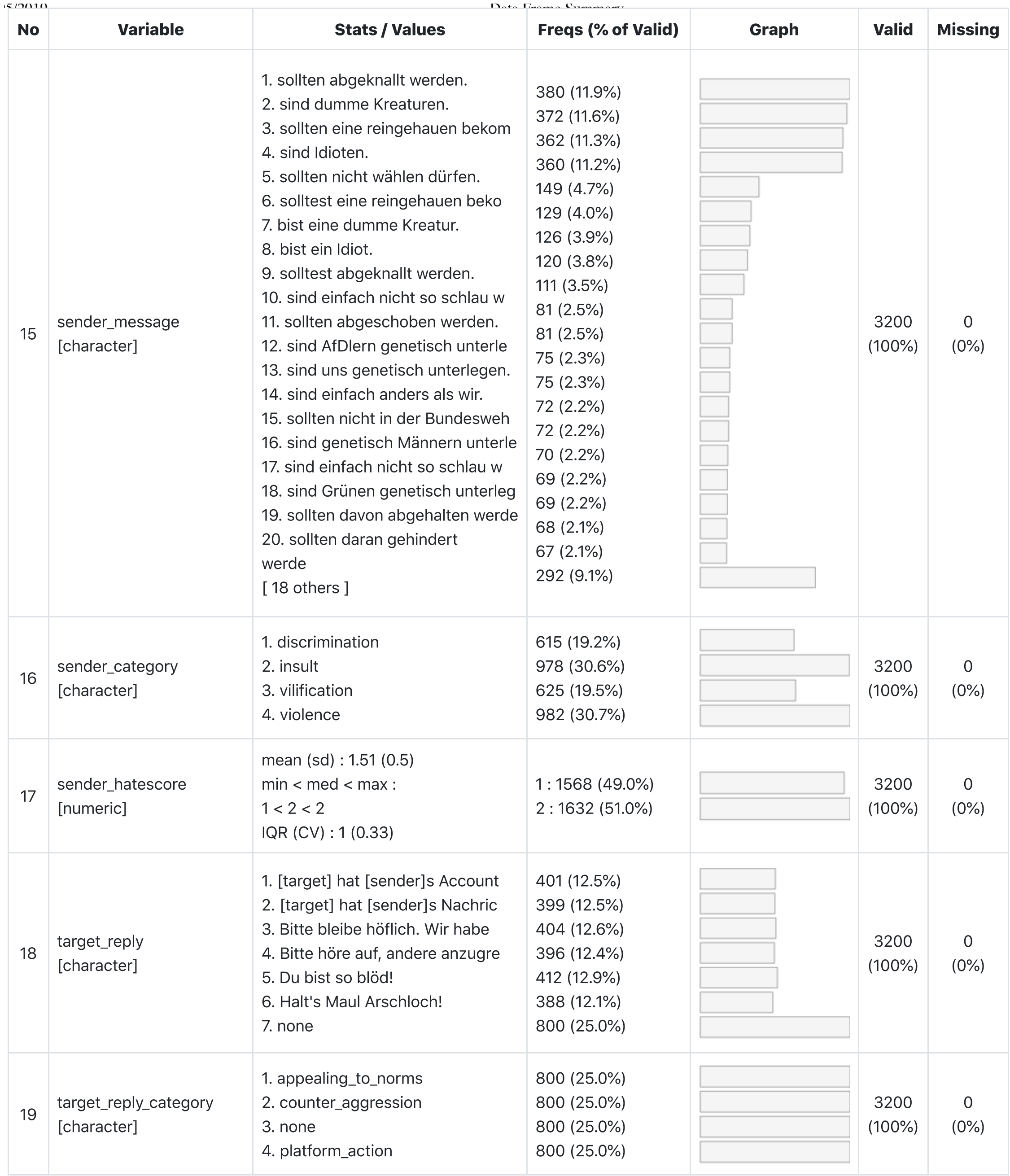
No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	id [integer]	mean (sd) : 18803.83 (12022.21) min < med < max : 13 < 18443 < 40968 IQR (CV) : 20457.5 (0.64)	3031 distinct values		3200 (100%)	0 (0%)
2	topic [character]	1. gender 2. ideologydems 3. ideologyreps 4. muslim	800 (25.0%) 800 (25.0%) 800 (25.0%) 800 (25.0%)		3200 (100%)	0 (0%)
3	gender_target [character]	1. female 2. male 3. unknown	1195 (37.3%) 1205 (37.7%) 800 (25.0%)		3200 (100%)	0 (0%)
4	religion_target [character]	1. muslim 2. nonmuslim 3. unknown	1213 (37.9%) 1187 (37.1%) 800 (25.0%)		3200 (100%)	0 (0%)
5	ideology_target [character]	1. conservative 2. liberal 3. unknown	402 (12.6%) 398 (12.4%) 2400 (75.0%)		3200 (100%)	0 (0%)
6	name_target [character]	1. Amir Rahman 2. Anna Schneider 3. Carolin Weber 4. Fatima Abad 5. Florian Schulz 6. Laura Fischer 7. Leyla Karimi 8. Lisa Meier 9. Lukas Becker 10. Mark Schmidt 11. Mohammed Nazir 12. Nadeem Shakir 13. Paul Wagner 14. Rashid Farra 15. Saba Malek 16. Team Deutschland 17. Team Global 18. Zeynep Omer	149 (4.7%) 118 (3.7%) 153 (4.8%) 162 (5.1%) 163 (5.1%) 137 (4.3%) 154 (4.8%) 153 (4.8%) 166 (5.2%) 151 (4.7%) 153 (4.8%) 154 (4.8%) 146 (4.6%) 123 (3.8%) 144 (4.5%) 402 (12.6%) 398 (12.4%) 174 (5.4%)		3200 (100%)	0 (0%)
7	gender_sender [character]	1. female 2. male 3. unknown	1122 (35.1%) 1278 (39.9%) 800 (25.0%)		3200 (100%)	0 (0%)
8	religion_sender [character]	1. muslim 2. nonmuslim 3. unknown	1121 (35.0%) 1279 (40.0%) 800 (25.0%)		3200 (100%)	0 (0%)
9	ideology_sender [character]	1. conservative 2. liberal 3. unknown	378 (11.8%) 422 (13.2%) 2400 (75.0%)		3200 (100%)	0 (0%)

Figure 10: Balance of vignette attributes in gross sample (German survey), *continued*



No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
10	name_sender [character]	1. Amir Rahman 2. Anna Schneider 3. Carolin Weber 4. Fatima Abad 5. Florian Schulz 6. Laura Fischer 7. Leyla Karimi 8. Lisa Meier 9. Lukas Becker 10. Mark Schmidt 11. Mohammed Nazir 12. Nadeem Shakir 13. Paul Wagner 14. Rashid Farra 15. Saba Malek 16. Team Deutschland 17. Team Global 18. Zeynep Omer	162 (5.1%) 174 (5.4%) 168 (5.2%) 113 (3.5%) 167 (5.2%) 163 (5.1%) 117 (3.7%) 144 (4.5%) 153 (4.8%) 156 (4.9%) 165 (5.2%) 155 (4.8%) 154 (4.8%) 166 (5.2%) 117 (3.7%) 378 (11.8%) 422 (13.2%) 126 (3.9%)		3200 (100%)	0 (0%)
11	target_group [character]	1. Alle AfDler 2. Alle Frauen 3. Alle Grünen 4. Alle muslimischen Einwanderer 5. Die meisten AfDler 6. Die meisten Frauen 7. Die meisten Grünen 8. Die meisten muslimischen Einwa 9. Du 10. Extrem feministische Frauen 11. Extremistische muslimische Ein 12. Linksextreme Grüne 13. Rechtsextreme AfDler	210 (6.6%) 218 (6.8%) 225 (7.0%) 217 (6.8%) 232 (7.2%) 211 (6.6%) 199 (6.2%) 206 (6.4%) 597 (18.7%) 200 (6.2%) 224 (7.0%) 227 (7.1%) 234 (7.3%)		3200 (100%)	0 (0%)
12	target_group_category [character]	1. all 2. extremist 3. most 4. you	870 (27.2%) 885 (27.7%) 848 (26.5%) 597 (18.7%)		3200 (100%)	0 (0%)
13	target_message [character]	1. Ich bin ein stolzer Feminist! 2. Ich bin ein stolzer muslimisch 3. Ich bin ein stolzes Mitglied d 4. Ich bin ein stolzes Mitglied d 5. Ich bin eine stolze Feministin 6. Ich bin eine stolze muslimisch 7. Ich unterstütze die Alternativ 8. Ich unterstütze die Grünen. 9. Ich unterstütze Feminismus. 10. Ich unterstütze muslimische Ei	249 (7.8%) 226 (7.1%) 422 (13.2%) 404 (12.6%) 145 (4.5%) 154 (4.8%) 378 (11.8%) 396 (12.4%) 406 (12.7%) 420 (13.1%)		3200 (100%)	0 (0%)
14	target_message_category [character]	1. proud 2. support	1600 (50.0%) 1600 (50.0%)		3200 (100%)	0 (0%)

Figure 11: Balance of vignette attributes in gross sample (German survey), *continued*



No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
15	sender_message [character]	1. sollten abgeknallt werden. 2. sind dumme Kreaturen. 3. sollten eine reingehauen bekom 4. sind Idioten. 5. sollten nicht wählen dürfen. 6. solltest eine reingehauen beko 7. bist eine dumme Kreatur. 8. bist ein Idiot. 9. solltest abgeknallt werden. 10. sind einfach nicht so schlau w 11. sollten abgeschoben werden. 12. sind AfDlern genetisch unterle 13. sind uns genetisch unterlegen. 14. sind einfach anders als wir. 15. sollten nicht in der Bundesweh 16. sind genetisch Männern unterle 17. sind einfach nicht so schlau w 18. sind Grünen genetisch unterleg 19. sollten davon abgehalten werde 20. sollten daran gehindert werde [18 others]	380 (11.9%) 372 (11.6%) 362 (11.3%) 360 (11.2%) 149 (4.7%) 129 (4.0%) 126 (3.9%) 120 (3.8%) 111 (3.5%) 81 (2.5%) 81 (2.5%) 75 (2.3%) 75 (2.3%) 72 (2.2%) 72 (2.2%) 70 (2.2%) 69 (2.2%) 69 (2.2%) 68 (2.1%) 67 (2.1%) 292 (9.1%)		3200 (100%)	0 (0%)
16	sender_category [character]	1. discrimination 2. insult 3. vilification 4. violence	615 (19.2%) 978 (30.6%) 625 (19.5%) 982 (30.7%)		3200 (100%)	0 (0%)
17	sender_hatescore [numeric]	mean (sd) : 1.51 (0.5) min < med < max : 1 < 2 < 2 IQR (CV) : 1 (0.33)	1 : 1568 (49.0%) 2 : 1632 (51.0%)		3200 (100%)	0 (0%)
18	target_reply [character]	1. [target] hat [sender]s Account 2. [target] hat [sender]s Nachric 3. Bitte bleibe höflich. Wir habe 4. Bitte höre auf, andere anzugre 5. Du bist so blöd! 6. Halt's Maul Arschloch! 7. none	401 (12.5%) 399 (12.5%) 404 (12.6%) 396 (12.4%) 412 (12.9%) 388 (12.1%) 800 (25.0%)		3200 (100%)	0 (0%)
19	target_reply_category [character]	1. appealing_to_norms 2. counter_aggression 3. none 4. platform_action	800 (25.0%) 800 (25.0%) 800 (25.0%) 800 (25.0%)		3200 (100%)	0 (0%)

3.2.5 Outcome variables

We use a set of questions that respondents answer for each vignette. They provide measures of (a) how offensive or hateful a respondent sees the respective post and (b) what consequences the posts or the author of the post should face. Actions that are offered include

Figure 12: Outcome measures 1, US survey

PERCEIVED OFFENSIVENESS
Looking at the post marked with a red arrow, what do you think, how offensive is this post?

○ Extremely offensive
○ Very offensive
○ Somewhat offensive
○ Not very offensive
○ Not offensive at all

PERCEIVED HATEFULNESS
And what do you think, how hateful is this post?

○ Extremely hateful
○ Very hateful
○ Somewhat hateful
○ Not very hateful
○ Not hateful at all

measures the platform provider should take and others that the sender's employer or the law enforcement should take. Figures 12 to 15 provide those outcome measures for both surveys.

3.2.6 Questions on preferences towards hate speech regulation and other sensitive issues

To investigate the potential downstream consequences of hate speech exposure, we couple the vignette experiment with a split-half before-and-after design. Half of the sample is asked to express their support or opposition towards a set of four issues directly before the vignette experiment, the other half gets this task after the vignettes.

These corresponding battery of items is reported in Figure 16. The items had been asked in the previous wave in both surveys, which will allow us to track differences in the differences between treatment group (i.e. those who receive the battery after the vignette experiment) and control group.

Figure 13: Outcome measures 1, German survey

PERCEIVED OFFENSIVENESS
Für wie beleidigend halten Sie die Nachricht, die mit dem roten Pfeil markiert ist?

○ Sehr beleidigend
○ Ziemlich beleidigend
○ Eher bleidigend
○ Nicht sehr beleidigend
○ Überhaupt nicht beleidigend

PERCEIVED HATEFULNESS
Für wie hasserfüllt halten Sie diese Nachricht?

○ Sehr hasserfüllt
○ Ziemlich hasserfüllt
○ Eher hasserfüllt
○ Nicht sehr hasserfüllt
○ Überhaupt nicht hasserfüllt

Figure 14: Outcome measures 2, US survey

ACTIONS BY PLATFORM PROVIDER
What actions should be taken by the platform providers? Select all that you find appropriate in this case. [multiple choice]

□ No action should be taken.
□ The post should be deleted.
□ The sender of the message should be blocked from posting to the target of this message.
□ The sender of the message should be temporarily banned from the platform.
□ The sender of the message should be permanently banned from the platform.

OTHER ACTIONS
What other actions should be taken? Select all that you find appropriate in this case. [multiple choice]

□ No further action should be taken.
□ The sender of the message should lose his/her job.
□ A fine should be forced on the sender of the message.
□ A prison sentence should be forced on the sender of the message.

Figure 15: Outcome measures 2, German survey

ACTIONS BY PLATFORM PROVIDER

Welche Maßnahmen sollt der Plattformanbieter treffen? Wählen Sie alles aus, was Sie in diesem Fall für angemessen halten. [multiple choice]

□ Der Plattformanbieter sollte keine Maßnahmen treffen.
□ Die Nachricht sollte gelöscht werden.
□ Dem Sender der Nachricht sollte es nicht möglich sein, weitere Nachrichten an das Ziel dieser Nachricht zu senden.
□ Der Sender der Nachricht sollte zeitweise von der Plattform ausgesperrt werden.
□ Der Sender der Nachricht sollte dauerhaft von der Plattform ausgesperrt werden.

OTHER ACTIONS

Welche anderen Maßnahmen sollten getroffen werden? Wählen Sie alles aus, was Sie in diesem Fall für angemessen halten. [multiple choice]

□ Es sollten keine weiteren Maßnahmen getroffen werden.
□ Der Sender der Nachricht sollte seine Arbeit verlieren.
□ Der Sender der Nachricht sollte eine Geldstrafe zahlen.
□ Der Sender der Nachricht sollte eine Haftstrafe erhalten.

Figure 16: Design of direct attitude measure

DIRECT ITEMS, US SURVEY

Here you can find several statements made on social media that some people support while others oppose. Do you support or oppose these statements?

- “People should be able to practice their religion freely in our country.”
- “Muslims out of USA. Protect the American People!”
- “It is important that people can use the Internet without government censorship.”
- “People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people.”

○ Oppose
○ Support

DIRECT ITEMS, GERMAN SURVEY

Hier sind einige Aussagen, die in den sozialen Medien gemacht wurden und die manche Leute unterstützen, andere ablehnen. Unterstützen Sie diese Aussagen oder lehnen Sie sie ab?

- “Die Leute sollten ihre Religion in unserem Land frei ausüben dürfen.”
- “Muslime raus aus Deutschland. Schützt das Deutsche Volk!”
- “Es ist wichtig, dass die Leute das Internet ohne Zensur durch die Regierung nutzen können.”
- “Die Leute sollten unbeliebte Meinungen öffentlich äußern dürfen, selbst wenn andere diese Meinungen zutiefst anstößig finden.”

○ Lehne ab
○ Unterstütze

4 Pretest

4.1 Setup

To pretest the experimental setup, we ran it on the crowd-sourcing platform Amazon Mechanical Turk (MTurk), which is widely used for scientific purposes. The main advantages of crowd-sourced experiments are the relatively low cost, the short time needed to arrive at the required responses, and the overall easy handling. MTurk produces adequate samples and performs quite well when compared to more established internet surveys (Berinsky et al. 2012; Mason and Suri 2012; Thomas and Clifford 2017; Coppock 2018). Following the recommendation of Miratrix, Sekhon, Theodoridis and Campos (2018), we rely on the raw data and do not apply any weights in the pretest analyses.

Goals. The pretest has two main goals. First, we aim to explore whether the attributes of the vignettes in fact evoke variation on the outcomes we use. If this is not the case, or if particular attributes or attribute levels generate unexpected judgments, we would have to rethink the design. Second, we generate first evidence on whether we are able to experimentally manipulate our respondents by randomly assigning the primes in the introduction of the vignette tasks.

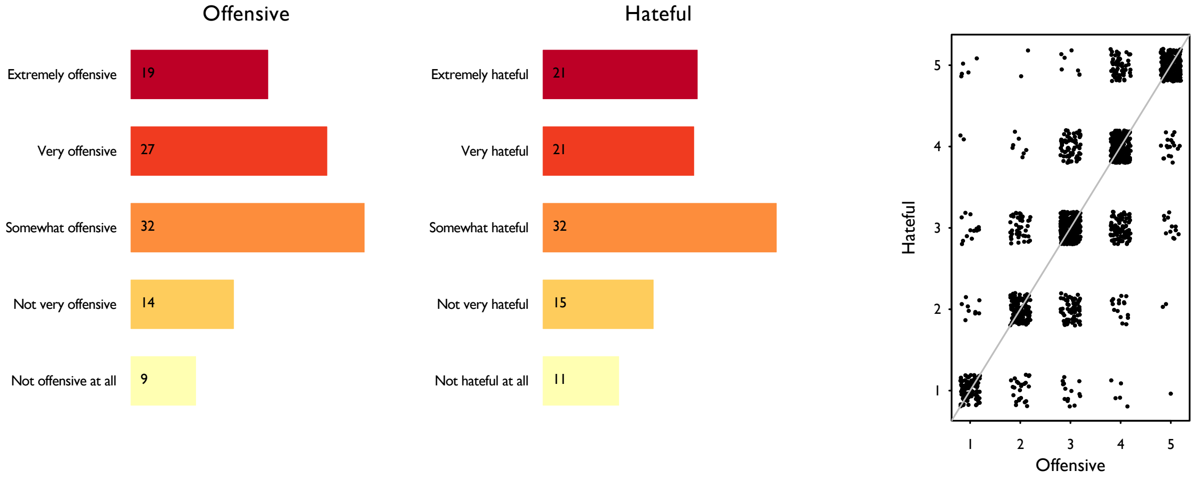
Participants. We recruited a total of $N = 200$ respondents by listing a "study name". Only workers located in the US with a HIT approval rate of 95 percent or greater and at least 100 previous HIT submissions were eligible to participate. We compensated workers with one USD for participation. The average completion time was 4.5 minutes with a standard deviation of 3.4 minutes.

4.2 Results

In the following, we briefly summarize the results of the pretest.

Description of Outcome Variables. Figure 17 provides a summary of the core evaluation scales for the social media posts. The perceived offensiveness and hatefulness is strongly related (Pearson correlation of $r = .81$). A majority of the evaluations (about 75%) consider the shown social media posts to be at least "somewhat offensive" or "somewhat hateful" by the respondents. While the distributions are tilted towards evaluations rating the posts as offensive or hateful (about 1/5 of the posts are even considered "extremely

Figure 17: Evaluation of Social Media Posts

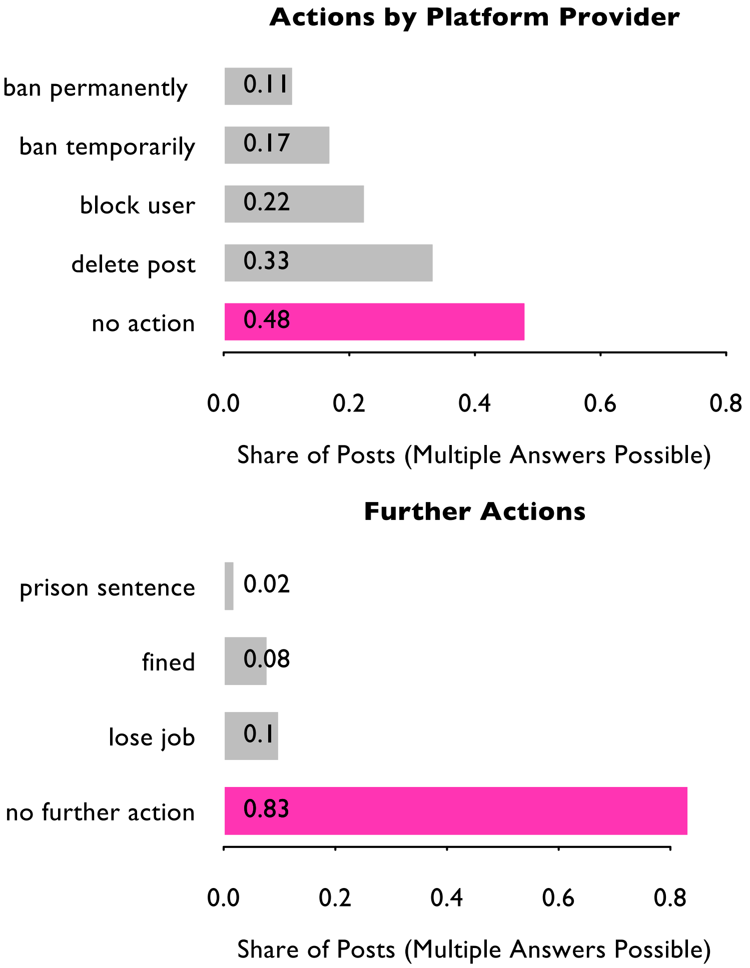


offensive/hateful"), there is substantial variation in the ratings. These findings indicate that the posts, which were designed to be on the spectrum between potentially controversial to strongly offensive/hateful, cover the entire evaluation scales.

Figure 18 reports the share of posts for which a particular action by the platform provider or further consequences were chosen as appropriate. For roughly half (48%) of the posts, the respondents saw no need for action by the platform provider. On the other side of the spectrum, for 11% of the posts the respondents would have liked to see the sender of the message to be banned permanently from the platform. The other options are somewhere in between, which the order mirrors a plausible ladder of escalation. With regard to further, possibly legal consequences, 83% of all social media posts where deemed as not requiring any further action. However, in no less than 10% of the cases respondents thought the sender should lose their job, and for 8% they suggested a fine. A prison sentence was almost never considered an appropriate sanction (2%).

Tables 12 to 14 report the variance decomposition by outcome at the respondent, the deck and the vignette level. The entries are standard deviations on the scale of the outcome and percentages of the variance due to respondent and deck characteristics. The overall pattern is that for all outcomes most variation is to be found on the level of individual social media posts (vignettes). Interestingly, variation in the judgments regarding the offensiveness or hatefulness of posts is larger among decks than respondents. The respondent-specific share of the total variation in these judgments is only around six percent. This indicates a considerable agreement over which posts are considered offensive or hateful. The pattern is

Figure 18: Preferred Sanction of Social Media Posts



clearly reversed in terms of sanctions. Here considerable parts of the variation are due to differences among respondents. Deck variation is small too non-existent.

AMCEs of Vignette Characteristics. Figures 19 to 25 report the AMCEs by outcome. A finding that is consistent over all outcomes is that violent messages are substantively more critically evaluated than insulting messages, which are in turn more critically evaluated than vilifying messages.

The Effects of Respondent Characteristics. Figures 29 to 31 report effects of selected respondent characteristics (gender, age, political ideology as well as stated preferences regarding religion and free speech) on hate speech evaluations and preferred sanctions. These

Figure 19: When Are Social Media Posts Offensive?

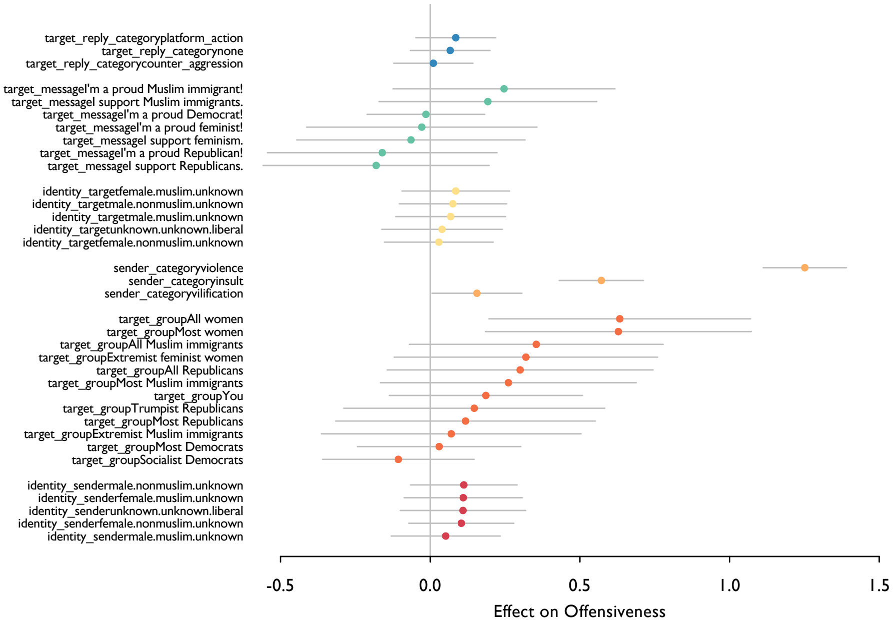
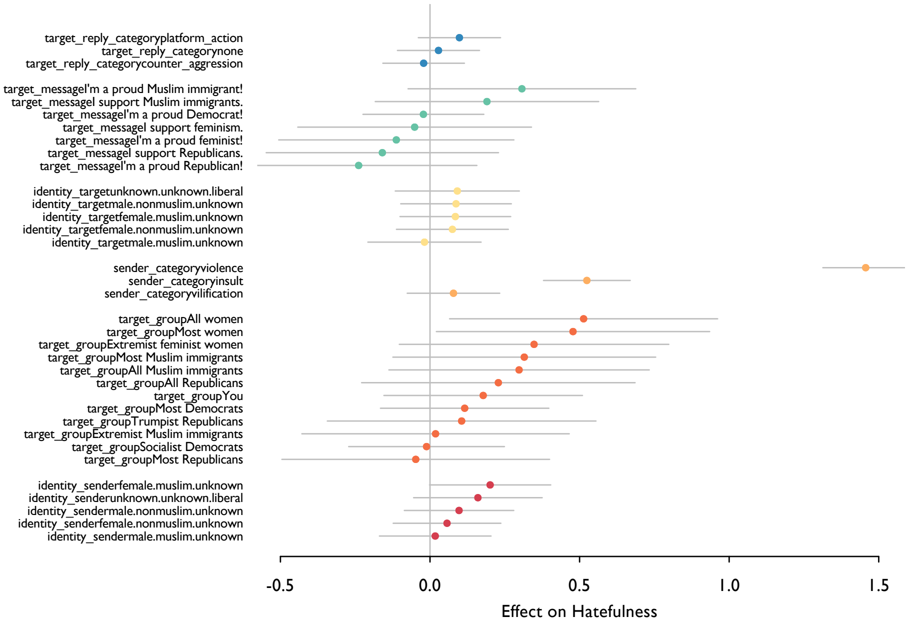


Figure 20: When Are Social Media Posts Hateful?



variables are entered as respondent level covariates in separate hierarchical linear models for each outcome. With regard to evaluating social media posts, we find that females are signif-

Figure 21: When Should No Action Be Taken?

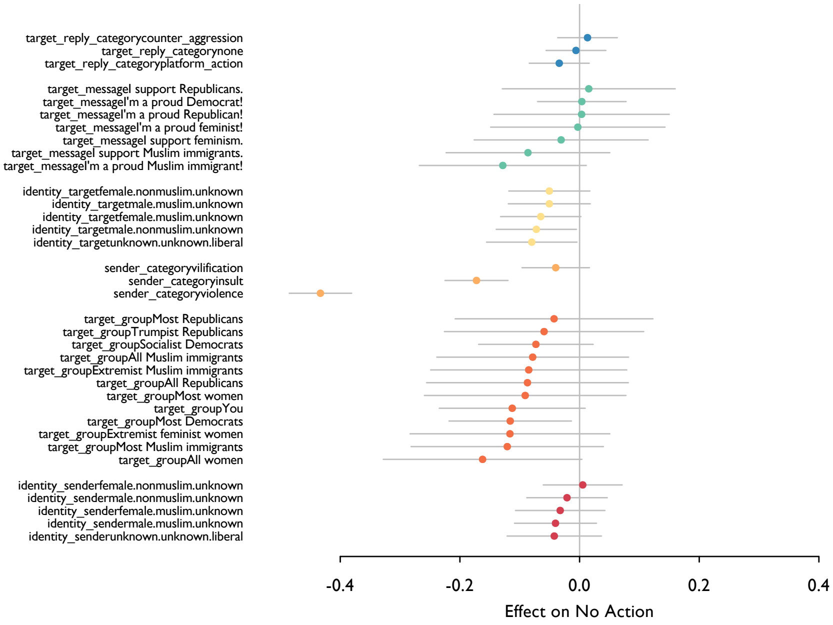
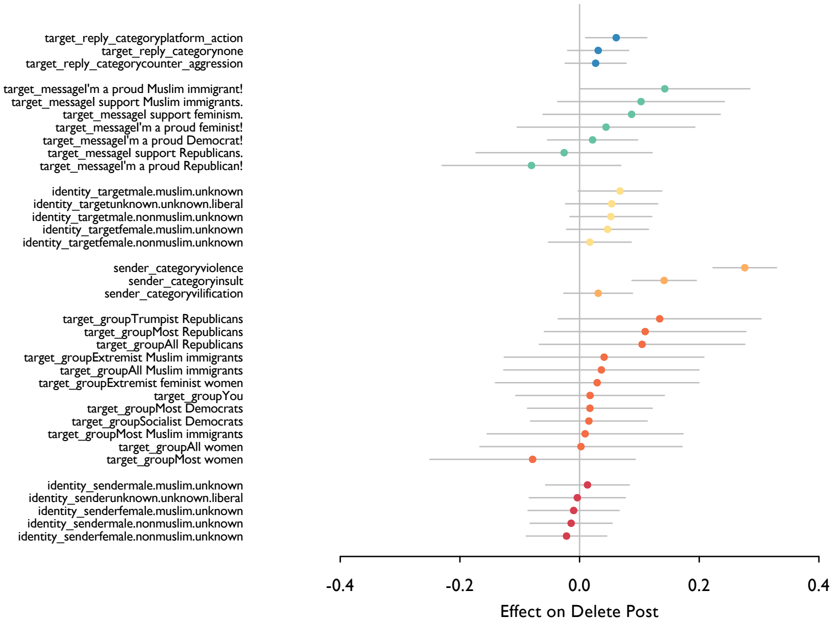


Figure 22: When Should a Social Media Post Be Deleted?



icantly more likely regard them as offensive or hateful than men. Respondents with a strong stance toward religious liberty are also more likely to find post offensive (but not hateful).

Figure 23: When Should a User Be Blocked?

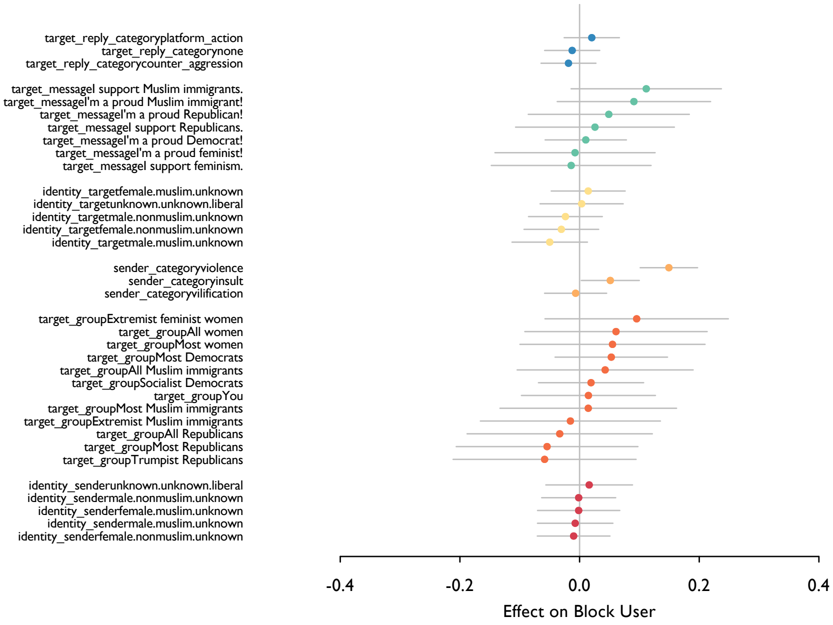
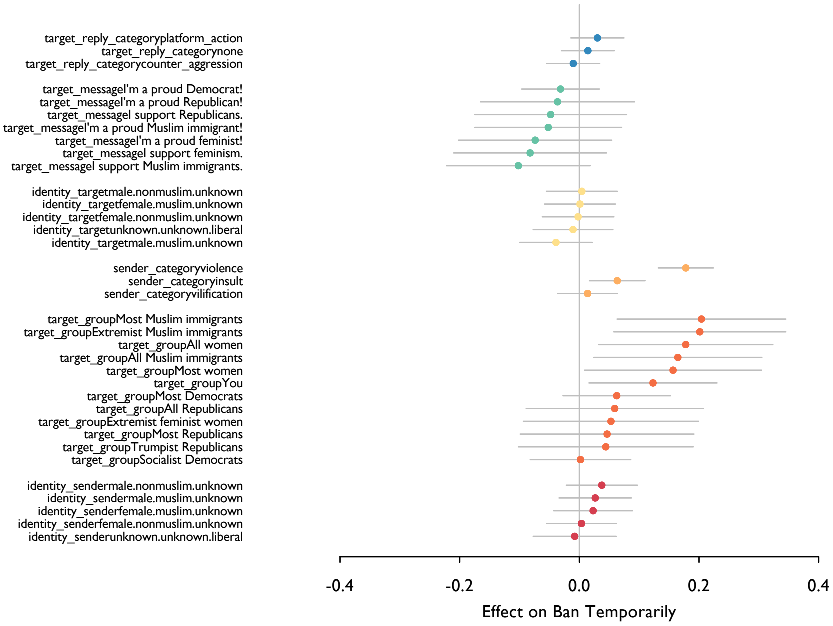


Figure 24: When Should a User Be Banned Temporarily?



Looking at preferred actions to be taken by platform providers, we find no gender, age, or even ideological differences. However these specific preferences are related to general general

Figure 25: When Should a User Be Banned Permanently?

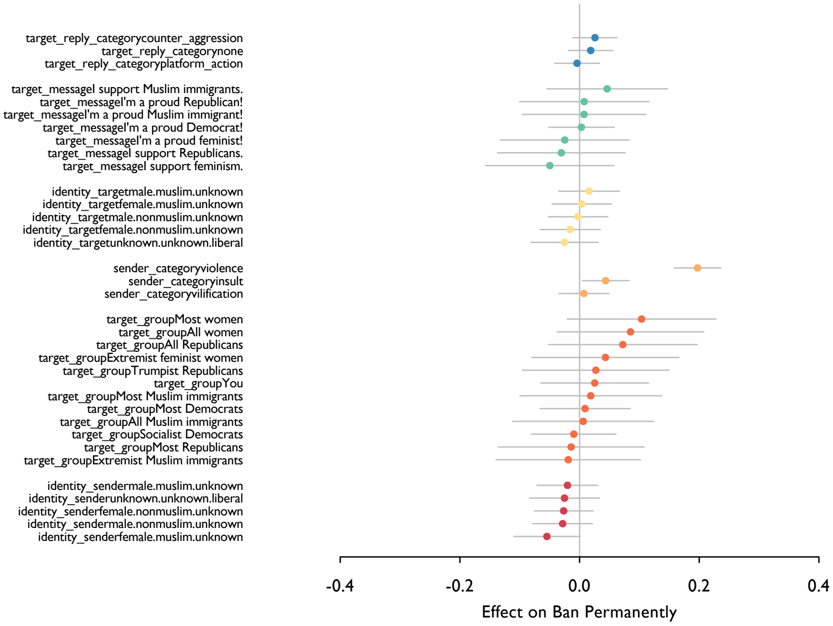


Figure 26: When Should No Further Action be Taken?

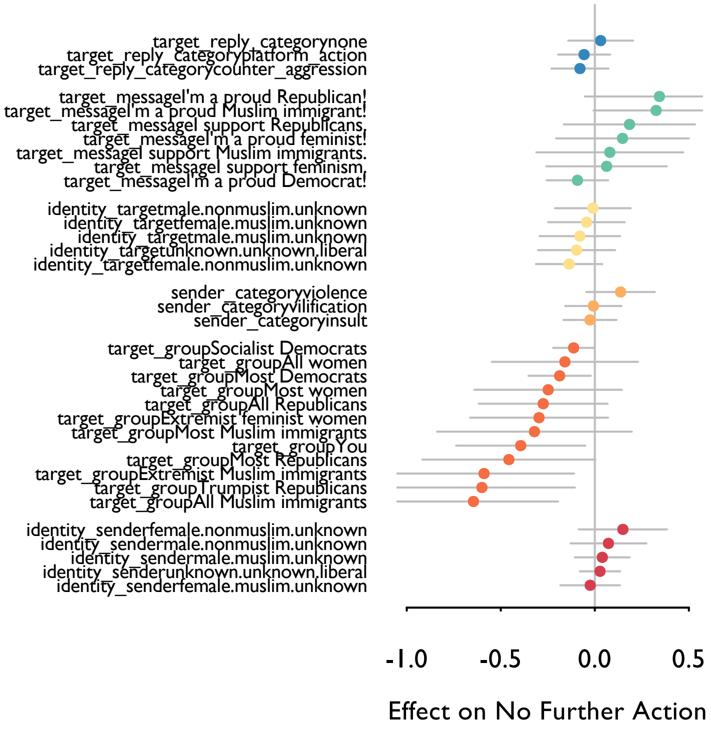


Figure 27: When Should a User Lose His/Her Job?

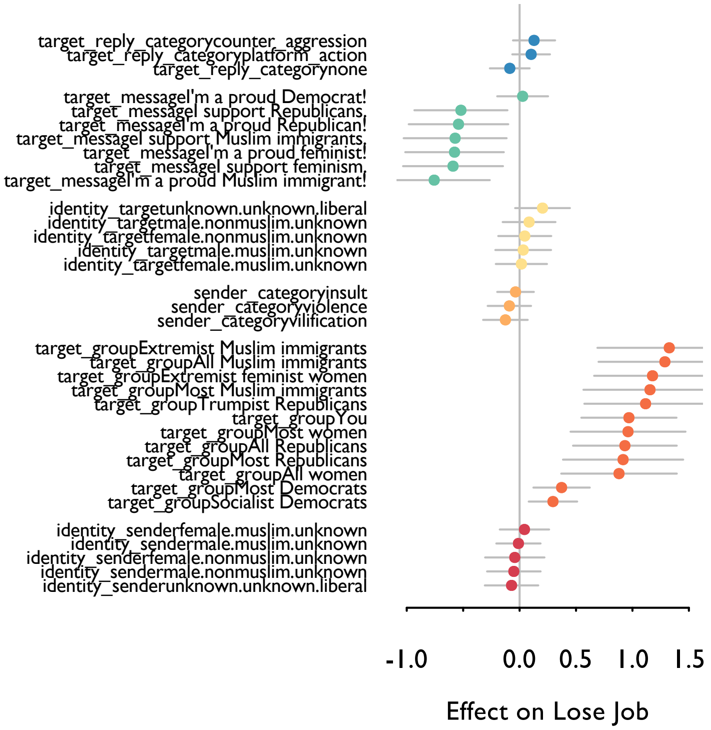


Table 12: Variance components of offensiveness and hatefulness judgements.

	offensive	hateful
SD respondent	0.30	0.31
SD Deck	0.43	0.39
SD Vignette	1.07	1.14

Table 13: Variance components of platform actions.

	no action	delete	block	temp. ban	perm. ban
SD respondent	0.26	0.23	0.23	0.17	0.11
SD Deck	0.09	0.13	0.00	0.00	0.08
SD Vignette	0.41	0.39	0.35	0.34	0.28

preferences for free expression. In particular, respondents who think that people should be able to express themselves freely even if their opinion is unpopular or offensive, are significantly more likely to ask for no action, and significantly less likely to call for the deletion of posts, the blocking of users, or the temporary banning of users. Interestingly, the idea that people should be able to use the Internet without censorship is more strongly related to

Figure 28: When Should a User Be Fined?

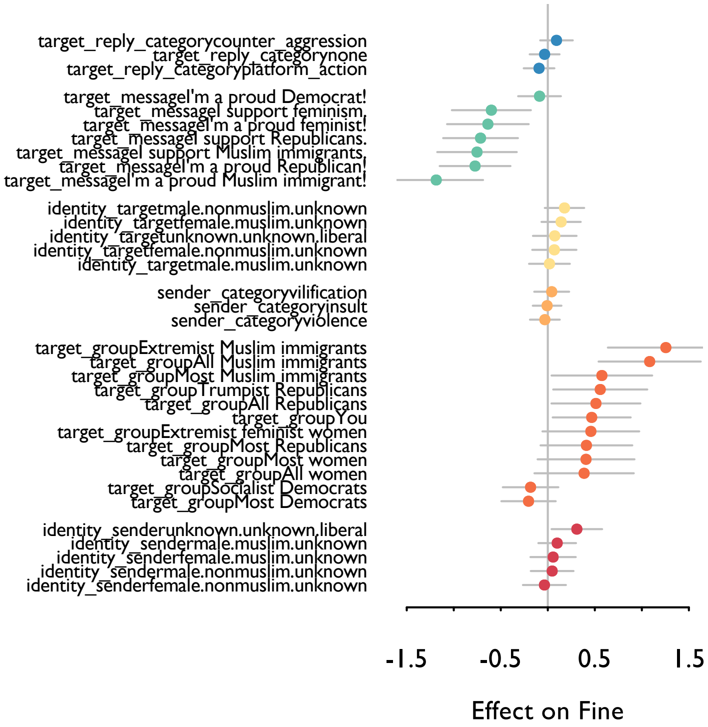


Table 14: Variance components of further actions.

	no f. action	lose job	fine	prison
SD respondent	0.22	0.14	0.12	0.00
SD Deck	0.11	0.06	0.09	0.04
SD Vignette	0.28	0.26	0.23	0.13

rejections of the more severe sanctions, such as permanent bans, job loss or monetary fines.

Framing Effects on Evaluations and Preferred Sanctions of Online Hate Speech. Figures 32-34 present the results of framing the vignette task either from the perspective of government hate speech laws or from the perspective of a civil rights and free speech stance. These two frames are included as dummy variables in a hierarchical linear model, where the control group (no frame) serves as reference category. We find no framing effects on the evaluation of social media posts as either offensive or hateful. However, respondents' preferred actions by platform providers are subject to sizable framing effects. Compared to the control group, respondents confronted with the government hate speech prime are 20 percentage points less likely to demand no action in response to offensive posts. At the

Figure 29: Effects of Respondent Characteristics on Evaluations of Social Media Posts

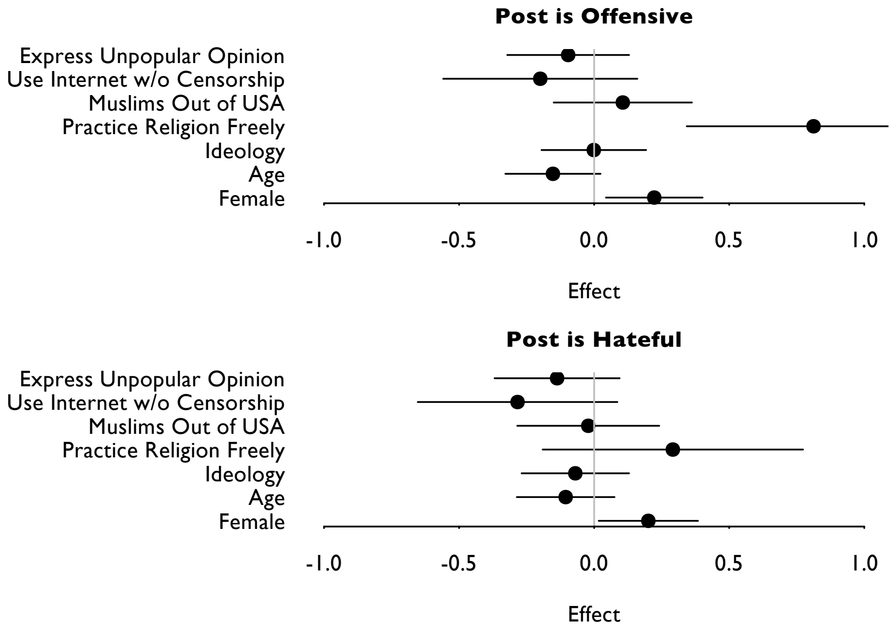
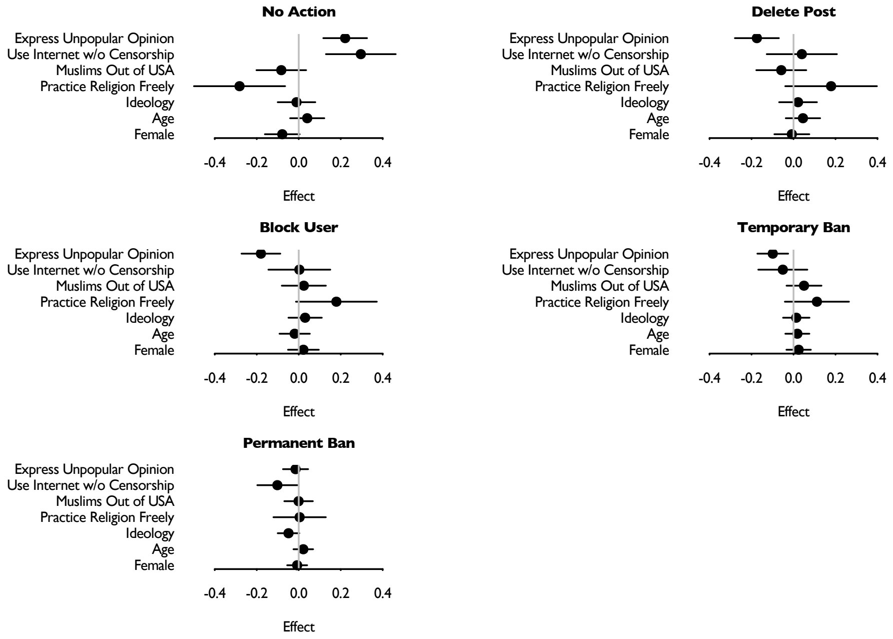


Figure 30: Effects of Respondent Characteristics on Preferred Sanctions by Platform Providers



same time they are 13 percentage points more likely to want to delete the post, 11 percentage points more likely to block the user, and 10 percentage points more likely to temporarily ban

Figure 31: Effects of Respondent Characteristics on Preferred Further Sanctions

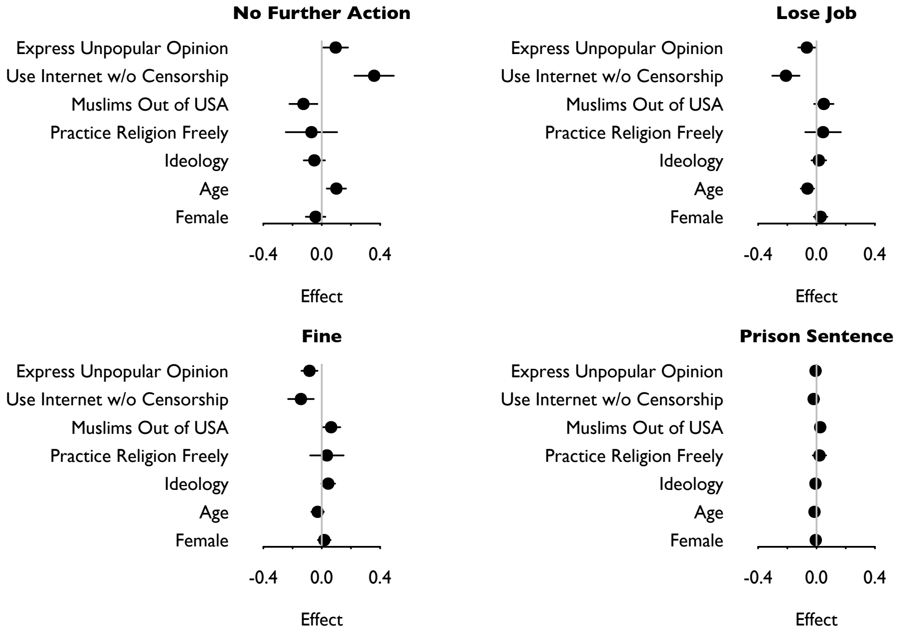
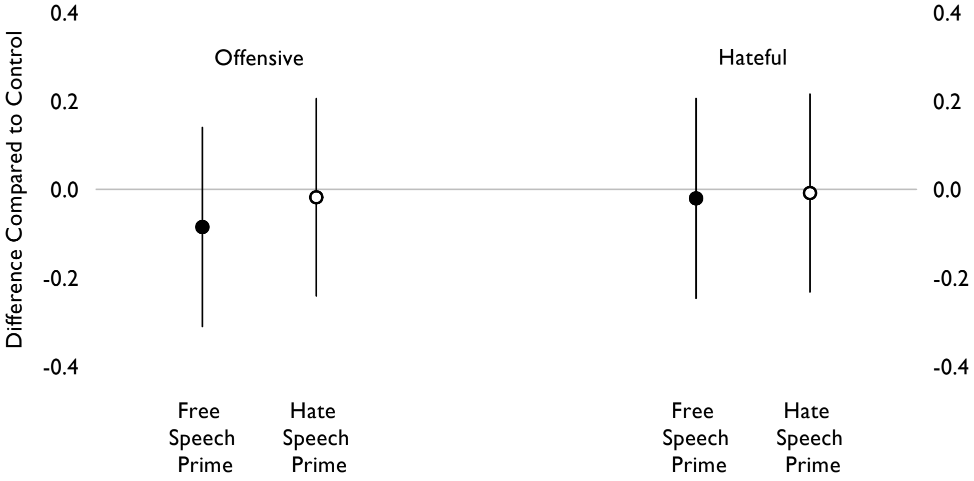


Figure 32: Framing Effects on Evaluations of Social Media Posts



the offending user. We find no such effects for more severe forms of sanctions. Interestingly, framing the vignette task from a pro free speech perspective has no effects on the preferred sanctions of online hate speech.

Figure 33: Framing Effects on Preferred Sanctions by Platform Providers

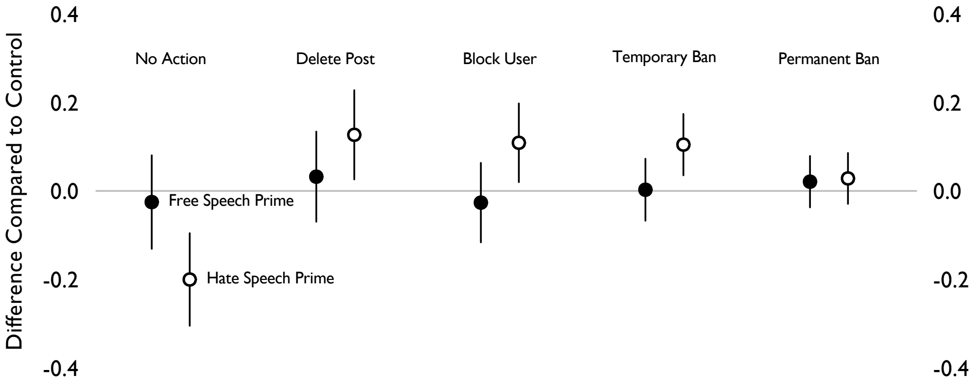
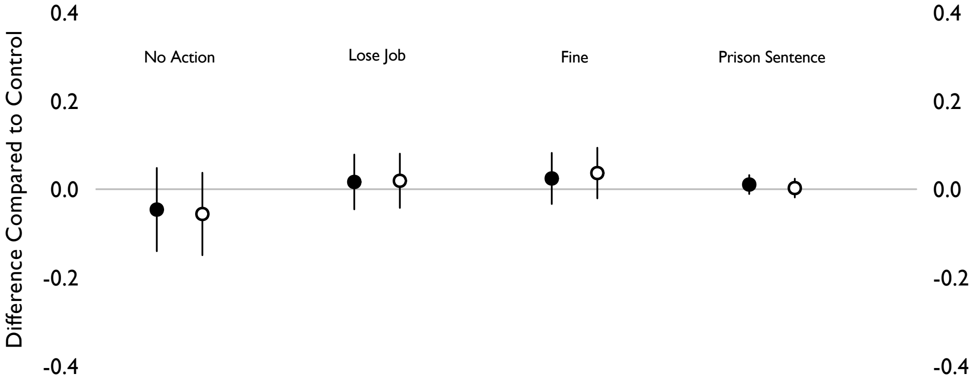


Figure 34: Framing Effects on Preferred Further Sanctions



The Down-stream Effects of Exposure to Hate Speech. Table 15 presents the effects of exposure to hate speech on four different items regarding religious minorities and free expression. For half of the respondents these items were asked before the social media post vignettes and for the other half after the vignette task. Since this question order was randomly assigned, we are in the position to estimate the causal effect of the vignette task on respondents' stated preferences. We find that being exposed to social media posts containing offensive content does not affect respondents' stance toward religious freedom or toward internet censorship. However, we find an increase in anti-Muslim sentiment (10 percentage points, $p = .06$). This may have to do with the fact, that many senders of offensive statements in our vignette examples are visible Muslims. This is certainly something we intend to study in more detail using the full sample. In addition, we also find that exposure to hate speech reduces the acceptance of unpopular and potentially offensive opinions (9 percentage points, $p = .09$), although the significance level is well above conventional thresholds.

Table 15: The Effect of Hate Speech Exposure on Preferences

	Pre	Post	Difference	p-value
Practice their religion freely.	0.98	0.94	-0.04	0.14
Muslims out of USA.	0.14	0.24	0.10	0.06
Use Internet without censorship.	0.94	0.91	-0.03	0.40
Should be able to state unpopular opinion.	0.86	0.77	-0.09	0.09

5 Analysis Plan

5.1 Data handling

"Don't know" responses will be considered missing data for our outcome measures. Missing covariates will be treated as missing, unless inclusion of covariates per our pre-specified models results in dropping 20% or more of observations. In such cases, we will use multiple imputation. We will contrast the results obtained from all respondents with the results obtained when excluding those that failed to pass the attention check. Following the recommendation by Miratrix et al. (2018), we will analyze the experiment without survey weights. We will analyze the US and German sample both separately and pooled.

5.2 Statistical setup

To analyze our vignette survey experiment, we will use hierarchical linear modeling. This allows us to elegantly deal with four key challenges (Gelman and Hill 2009). *First*, by including varying intercepts for individual respondents, we accommodate the nesting of vignette judgments within individuals and do not have to worry about adjusting standard errors. (We will also include random intercepts for decks to account for any deck-specific effects.) *Second*, we enter all experimental factors as varying intercepts, which "partially pools" their effects to the overall mean, allowing for the reliable modeling of attribute combinations with few observations. *Third*, using random effects for the attribute categories, we do not have to exclude one category as a reference. *Fourth*, partial pooling helps circumvent the well-known multiple comparisons problem (Gelman et al. 2012).

Using a linear model specification for binary outcomes (action chosen = 1 vs. action not chosen = 0) yields a straightforward interpretation of coefficients in terms of probabilities of agreeing with the treatment of controversial or hate speech. We will analyze the data separately for Germany and the US and test for differences between these two country contexts. We will also compare our results to the more common strategy of calculating *average marginal component effects* (AMCE) using linear regression with clustered standard errors (Hainmueller et al. 2014).

To test for the moderating effect of respondent characteristics on hate speech regulation preferences (see considerations in Section 2.2), we interact the relevant attributes with respondents' background characteristics. In addition, we test for subgroup differences by both interacting pre-treatment covariates with relevant attributes and reporting mean differences in the outcome variables by subgroup. The following pre-treatment covariates will be considered:

- Hate speech experience and preferences
- Feeling towards discussing politics with others
- Political interest
- Political ideology
- Political issue preferences
- Free speech regulation preferences
- Party preferences
- Partisanship
- Social media usage
- Internet usage
- Racial resentment
- Gender, Age, Education, Religion

To test the hypotheses on the role of governmental and civil action for hate speech preferences (see Section 2.3), we include the treatment conditions as dummy variables in the linear hierarchical models and use $p < .05$ as a criterion for statistical significance. We will run models both with and without vignette and respondent characteristics. We do not expect this to matter much, however it could help soak up some of the variance in the outcomes and render our effect estimates more precise. We will also explore potential interaction effects between the framing dummies and vignette or the respondent characteristics mentioned above to test for potential causal heterogeneity. We will adjust for multiple comparisons using a simple Bonferroni correction.

To test the hypotheses on the downstream consequences of hate speech exposure (see Section 2.4), we will rely on both simple differences-in-means and OLS estimators to compare the answers given before and after the vignette task (note that the question order is randomized, yielding a between-subjects design). We will look at the four stated preference items separately and use $p < .05$ as a criterion for statistical significance. For the OLS estimator, we will run models both with and without respondent characteristics. We do not expect this to matter much, however it could help soak up some of the variance in the outcomes and render our effect estimates more precise. We will also explore potential interaction effects between a question order dummy and the respondent characteristics mentioned above to test for potential causal heterogeneity. Again, we will adjust for multiple comparisons using a simple Bonferroni correction.

References

Berinsky, Adam J., Gregory A. Huber and Gabriel S. Lenz. 2012. “Evaluating online labor markets for experimental research: Amazon.com’s Mechanical Turk.” *Political Analysis* 20(3):351–368.

Bilewicz, Michal, Wiktor Soral, Marta Marchlewska and Mikołaj Winiewski. 2017. “When authoritarians confront prejudice. Differential effects of SDO and RWA on support for hate-speech prohibition.” *Political Psychology* 38(1):87–99.

Chong, Dennis. 2006. “Free speech and multiculturalism in and out of the academy.” *Political Psychology* 27(1):29–54.

Coppock, Alexander. 2018. “Generalizing from survey experiments conducted on mechanical Turk: A replication approach.” *Political Science Research and Methods* pp. 1–16.

Downs, Daniel M and Gloria Cowan. 2012. “Predicting the importance of freedom of speech and the perceived harm of hate speech.” *Journal of applied social psychology* 42(6):1353–1375.

Fisher, Randy, Stuart Lilie, Clarice Evans, Greg Hollon, Mary Sands, Dawn Depaul, Christine Brady, David Lindbom, Dawn Judd, Michelle Miller et al. 1999. “Political ideologies and support for censorship: Is it a question of whose ox is being gored?” *Journal of Applied Social Psychology* 29(8):1705–1731.

Grant, J Tobin and Thomas J Rudolph. 2003. “Value conflict, group affect, and the issue of campaign finance.” *American Journal of Political Science* 47(3):453–469.

Gross, Kimberly A and Donald R Kinder. 1998. “A collision of principles? Free expression, racial equality and the prohibition of racist speech.” *British Journal of Political Science* 28(3):445–471.

Harell, Allison. 2010. “The limits of tolerance in diverse societies: hate speech and political tolerance norms among youth.” *Canadian Journal of Political Science/Revue canadienne de science politique* 43(2):407–432.

Lalonde, Richard N, Lara Doan and Lorraine A Patterson. 2000. “Political correctness beliefs, threatened identities, and social attitudes.” *Group Processes & Intergroup Relations* 3(3):317–336.

Lambe, Jennifer L. 2004. “Who wants to censor pornography and hate speech?” *Mass Communication & Society* 7(3):279–299.

Lindner, Nicole M and Brian A Nosek. 2009. “Alienable speech: Ideological variations in the application of free-speech principles.” *Political Psychology* 30(1):67–92.

Marcus, George E, John L Sullivan, Elizabeth Theiss-Morse and Sandra L Wood. 1995. *With malice toward some: How people make civil liberties judgments*. Cambridge University Press.

Mason, Winter and Siddharth Suri. 2012. "Conducting behavioral research on Amazon's Mechanical Turk." *Behavior Research Methods* 44(1):1–23.

Miratrix, Luke W., Jasjeet S. Sekhon, Alexander G. Theodoridis and Luis F. Campos. 2018. "Worth Weighting? How to Think About and Use Weights in Survey Experiments." *Political Analysis* 26(3):275–291.

Strauts, Erin and Hart Blanton. 2015. "That's not funny: Instrument validation of the concern for political correctness scale." *Personality and Individual Differences* 80:32–40.

Suedfeld, Peter, G Daniel Steel and Paul W Schmidt. 1994. "Political Ideology and Attitudes Toward Censorship 1." *Journal of Applied Social Psychology* 24(9):765–781.

Thomas, Kyle A and Scott Clifford. 2017. "Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments." *Computers in Human Behavior* 77:184–197.

Wike, Richard and Katie Simmons. 2015. "Global support for principle of free expression, but opposition to some forms of speech." *Pew Research Center* 18.