Pre-Analysis Plan for: The Effect of Hate Speech Regulation on Preference Falsification

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

January 9, 2019

Contents

1 Summary

This document describes a pre-analysis plan for a double list experiment embedded in an two-country online survey that examines the effect of hate speech regulation on public discourse and online civility. Our core treatment is a prime based on fictional hate speech law that mimics existing legislation. After reporting their level of support towards the content of the legislation, respondents are asked to disclose preferences on a set of issues related to free speech, religious freedom, and treatment of religious groups. The reported opinions are then compared with revealed positions on the same items that were elicited pre-treatment using a set of double-list experiments.

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.

On a more pragmatic level, the debate over hate speech legislation also rests on a series of untested claims about the expected effectiveness and likely consequences of regulatory intervention. We propose to inform the debate on hate speech regulation by testing some of these claims against experimental evidence. While the most obvious but also most doubtful proposition is that hate speech laws induce a true change in discriminatory attitudes, we are interested in testing hypotheses on potential unintended consequences of hate speech legislation.

The first hypothesis holds that far from reducing obnoxious ideas, **restricting "hate speech" merely drives them underground and thus results in an increase in preference falsification (H1)**. Faced with sanctions for hate speech, people will show a greater difference between what they truly believe and what the say they believe with regards to a social group protected under hate speech legislation. This behavioral effect is

highly problematic because it exacerbates the sanctioning of hateful speech: when people who hold problematic views do not express them openly, we do not know who and how many they are and their ideas cannot be challenged. As substantive application, we will look at an offensive statement toward Muslim immigrants, which reads: "Muslims out of USA. Protect the American People!"

The second hypothesis states that, regardless of its effect on true hate speech, hate speech regulation results in a **chilling effect** on public discourse, **where citizens are reluctant to openly debate controversial, yet crucial political issues, choosing to self-censor instead (H2)**. Thus faced with sanctions for hate speech, people will be less willing to reveal their policy preference. This effect is particularly harmful for democracy, which rests on the free debate of policy options. As substantive application, we will focus on respondents' preference towards freedom of speech using the statement "People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people."

Third, we expect **country differences** in the baseline levels of preferences as revealed in the list experiments. The US and Germany not only differ in their cultural tradition concerning free expression (First Amendment vs. continental tradition) but in their actual hate speech legislation. Whereas no hate speech law exists in the US, Germany legally sanctions certain acts of speech (e.g., Holocaust denial) and has recently introduced a social media law to combat hate speech ("Netzwerkdurchsetzungsgesetz"). Therefore, we expect higher baseline revealed support (in the list experiment) for the free speech statement in the US compared to Germany. We have no prior expectations regarding baseline differences in support for the offensive statement toward Muslim immigrants. However, we have no a priori expectations about country differences in the effect of hate speech regulation. We expect H1 and H2 to hold in both contexts.

Furthermore, we will test for heterogeneous effects in the above hypotheses in various subgroups made explicit below.

In the following section we make explicit how we plan to identify the hypothesized effects in the experimental setup.

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 will be 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 list experiment will be embedded will be launched in December 2018 in both Germany and the United States.

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 reweighted 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 four components:

1. Two double list experiments
2. An experimental manipulation priming hate speech regulation (preceded by a attention check)
3. A matching battery of direct items
4. Items for subgroup analyses (already part of the core survey)

Double list experiment. A difficulty in empirically testing the effects of hate speech regulation is the problem of "preference falsification" (Kuran 1997): How can we separate what people say from what they actually believe? We propose carefully designed list experiments to measure people's willingness or reluctance to openly express discriminatory or controversial views. While this unobtrusive method is intended to avoid the ubiquitous problem of social desirability bias in sensitive survey questions (Tourangeau and Yan 2007), we rely on it to elicit survey item misreporting as a quantity of substantive interest (Gilens, Sniderman

and Kuklinski 1998). In other words, it is precisely the misreporting of sensitive opinions that lets us learn about the presence of preference falsification' and self-censoring behavior.

In a double list experiment, both, the treatment and the control group get two lists of items. For one group, the sensitive item is included in the first list; and for the second group the sensitive item included in the second list. We therefore obtain two estimates of the sensitive item—one estimate from comparing the first list and the second from comparing the second list. Averaging across these two estimates produces the final estimate. Importantly, the variance of this combined estimator is considerably smaller than for the single estimators.

To ensure that the variance of our estimator is minimized, we have to carefully design the control items. A particular challenge is to balance the use of low (or high) prevalence items that reduce the variance but invite the possibility of floor (or ceiling) effects. One solution is to use negatively correlated items within a list and positively correlated lists (Glynn 2013). Our strategy is to design two lists of $J = 3$ items each. To select these items, we rely on items that were tested in a pretest using an MTurk sample of U.S. MTurk workers.

Table 1 provides an overview of the setup. For the first list pair 1A and 1B, respondents are randomly assigned to either Group A or Group B. Respondents randomly assigned to Group A first receive List 1A containing the sensitive item S_1, which reads: "Muslims out of USA. Protect the American People!", and List 1B afterwards without the sensitive item. Respondents randomly assigned to Group B first receive List 1A without the sensitive item and directly afterwards List 1B containing the sensitive item S_1. For the second list pair, respondents are again randomly assigned to either Group C or Group D. Respondents randomly assigned to Group C first receive List 2A containing the sensitive item S_2, which reads: "People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people.", and List 2B afterwards without the sensitive item. Respondents randomly assigned to Group D first receive List 2A without the sensitive item and directly afterwards List 2B containing the sensitive item S_2. In addition, we randomize the order of the items within each list.

The assignment of the treatment and control lists is blocked on the variables gender, age, and education.

Figures 1, 2, 3, and 4 report the wording of the lists in both the German and the U.S. survey. The treatment item as well as the answering option "4" were exclusively shown to the respective treatment group.

Figure 1: Design of List 1A

LIST 1A, US SURVEY

In the following you will find statements posted on social media that some people support or oppose. After you have read all statements, please tell us HOW MANY of them you support. We do not want to know which ones, just HOW MANY.

- "Marijuana is harmless. Legalize it!"
- "Vaccines save lives. Mandatory vaccination for all children!"
- "Stop killing babies. Make Abortion illegal in ALL cases!"
- [**treatment item**] "Muslims out of USA. Protect the American People!"

○ 0 ○ 1 ○ 2 ○ 3 ○ [4]

LIST 1A, GERMAN SURVEY

Im Folgenden sehen Sie einige Aussagen, die in den sozialen Medien gemacht wurden und die manche Leute unterstützen, andere ablehnen. Nachdem Sie alle Aussagen gelesen haben, sagen Sie uns bitte WIE VIELE davon sie unterstützen. Uns interessiert nicht welche, nur WIE VIELE.

- "Marihuana ist harmlos. Legalisierung jetzt!"
- "Impfungen retten Leben. Impfpflicht für alle Kinder!"
- "Stoppt das Töten von Babies. Macht Abtreibung in ALLEN Fällen illegal!"
- [**treatment item**] "Muslime raus aus Deutschland. Schützt das Deutsche Volk!"

○ 0 ○ 1 ○ 2 ○ 3 ○ [4]

Hate speech regulation prime. In our experimental manipulation, we prime half of the respondents with a fictitious hate speech law. While the law is fictitious, it closely resembles actual legislation in the UK. In fact, it is composed from a mix of fragments from the

Table 1: Design of double list experiment

List pair 1		List pair 2	
Group A	Group B	Group C	Group D
List 1A + sensitive item S_1	List 1A	List 2A + sensitive item S_2	List 2A
List 1B	List 1B + sensitive item S_1	List 2B	List 2B + sensitive item S_2

Figure 2: Design of List 1B

LIST 1B, US SURVEY

Here is another set of statements posted on social media that some people support or oppose. Please tell us again HOW MANY of them you support. We do not want to know which ones, just HOW MANY.

- "More Women in Tech. Affirmative Action now!"
- "Guns dont kill people. People kill people."
- "Save the planet. Raise the taxes on gasoline!"
- **[treatment item]** "Muslims out of USA. Protect the American People!"

○ 0 ○ 1 ○ 2 ○ 3 ○ [4]

LIST 1B, GERMAN SURVEY

Hier sind weitere Aussagen, die in den sozialen Medien gemacht wurden und die manche Leute unterstützen, andere ablehnen. Nachdem Sie alle Aussagen gelesen haben, sagen Sie uns bitte WIE VIELE davon sie unterstützen. Uns interessiert nicht welche, nur WIE VIELE.

- "Mehr Frauen in Technik-Berufen. Frauenquote sofort!"
- "Nicht Schusswaffen töten Menschen. Menschen töten Menschen!"
- "Rettet den Planeten. Höhere Steuern auf Benzin!"
- **[treatment item]** "Muslime raus aus Deutschland. Schützt das Deutsche Volk!"

○ 0 ○ 1 ○ 2 ○ 3 ○ [4]

Public Order Act 1986, the Communications Act 2003 and the Racial and Religious Hatred Act 2006. Yet, we simplified the legalistic language to reduce the cognitive burden on the respondents. After instructing them to carefully read the text, we ask whether respondents oppose or support the law on a five-point scale. The wordings of the prime in the two languages are documented in Figures 6) and 7). To minimize deception, we do not suggest that such a law is in place but tell respondents that the law is currently discussed. In fact, Germany already has a new social media law targeting hate speech (NetzDG) that came into effect in early 2018, but the law is still discussed in the public and among policymakers. The assignment of the treatment and control lists is blocked on the variables gender, age, and education.

To check whether respondents actually take the time to carefully read the fictitious hate

Figure 3: Caption

LIST 2A, US SURVEY

In the following you will find statements that some people support or oppose. After you have read all statements, please tell us HOW MANY of them you support. We do not want to know which ones, just HOW MANY.

- "People should be able to make statements that criticize the government publicly."
- "Media organizations should be able to publish information about large political protests in our country."
- "Government should be able to stop a news media outlet from publishing biased or inaccurate information."
- [**treatment item**] "People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people."

○ 0 ○ 1 ○ 2 ○ 3 ○ [4]

LIST 2A, GERMAN SURVEY

Im Folgenden sehen Sie einige Aussagen, die manche Leute unterstützen, andere ablehnen. Nachdem Sie alle Aussagen gelesen haben, sagen Sie uns bitte WIE VIELE davon sie unterstützen. Uns interessiert nicht welche, nur WIE VIELE.

- "Die Leute sollten die Regierung öffentlich kritisieren dürfen."
- "Die Medien sollten über große politische Proteste im Land berichten dürfen."
- "Die Regierung sollte die Medien davon abhalten dürfen, einseitige oder falsche Informationen zu veröffentlichen."
- [**treatment item**] "Die Leute sollten unbeliebte Meinungen öffentlich äußern dürfen, selbst wenn andere diese Meinungen zutiefst anstößig finden."

○ 0 ○ 1 ○ 2 ○ 3 ○ [4]

speech law, we run an attention check just before the experimental manipulation (Berinsky, Huber and Lenz 2012). In this attention check, respondents are asked to ignore the initial question (about daylight saving time) and to just type 'read' into the open text field (Fig. 5). 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.

Figure 4: Design of List 2B

LIST 2B, US SURVEY

Here is another set of statements that some people support or oppose. Please tell us again HOW MANY of them you support. We do not want to know which ones, just HOW MANY.

- "Marijuana should be legalized, even if it may be harmful for some people."
- "Vaccination should be mandatory for all children, even if parents oppose it."
- "Abortion should be illegal, even if there is a health risk for the mother."
- [**treatment item**] "People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people."

○ 0 ○ 1 ○ 2 ○ 3 ○ [4]

LIST 2B, GERMAN SURVEY

Hier sind weitere Aussagen, die manche Leute unterstützen, andere ablehnen. Nachdem Sie alle Aussagen gelesen haben, sagen Sie uns bitte WIE VIELE davon sie unterstützen. Uns interessiert nicht welche, nur WIE VIELE.

- "Marijuana sollte legalisiert warden, auch wenn es für manche Leute schädlich ist."
- "Impfungen sollte für alle Kinder verpflichtend sein, auch wenn die Eltern dagegen sind."
- "Abtreibung sollte verboten sein, auch bei Gesundheitsrisiken für die Mutter."
- [**treatment item**] "Die Leute sollten unbeliebte Meinungen öffentlich äußern dürfen, selbst wenn andere diese Meinungen zutiefst anstößig finden."

○ 0 ○ 1 ○ 2 ○ 3 ○ [4]

Direct items. In order to contrast the support for the two sensitive items elicited indirectly using double list experiments with the answers given in direct questioning, we include a short battery of direct items (Fig. 8). Two of these direct items exactly match the sensitive items used in the list experiments: "Muslims out of USA. Protect the American People!" and "People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people." These will allow us to calculate the degree of preference falsification. In addition, we add another item on religious freedom ("People should be able to practice their religion freely in our country.") and another item on free speech on the internet ("It is important that people can use the Internet without government censorship."). The direct item battery immediately follow the prime (or the attention check for those who

Figure 5: Design of attention check before hate speech regulation prime

ATTENTION CHECK BEFORE PRIME, US SURVEY

People have different opinions about abolishing the switch to daylight saving time. Some would like to get rid of daylight saving time, others to get rid of standard time, others want everything to stay as it is. 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.

- ○ Very much oppose
- ○ Rather oppose
- ○ Neither oppose nor support
- ○ Rather support
- ○ Very much support
- ○ Other: _____

ATTENTION CHECK BEFORE PRIME, GERMAN SURVEY

Leute haben unterschiedliche Meinungen zur Abschaffung der Zeitumstellung. Einige würden gerne die Sommerzeit abschaffen, andere die Winterzeit, andere hätten gerne, dass alles so bleibt, wie es ist. Wir möchten 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 "Andere" unten ein.

- ○ Lehne voll und ganz ab
- ○ Lehne eher ab
- ○ Weder noch
- ○ Unterstütze eher
- ○ Unterstütze voll und ganz
- ○ Andere: _____

are members of the control group). Furthermore, the direct item battery is placed after the double list experiments and with several buffer items in between to avoid potential interference.

Figure 6: Design of hate speech regulation prime, U.S. survey

HATE SPEECH REGULATION PRIME, US SURVEY

As you may have heard, the government is making serious efforts to combat online hate speech. This could mean that a large number of social media posts with offensive or hateful content will be deleted and legally prosecuted.
The content of hate speech legislation that is currently discussed is described in the following text. Please read it very carefully and make sure you understand it. Please read it very carefully and make sure you understand it.

"A person is guilty of an offense if she sends a message over an online platform which

- *uses threatening, abusive or insulting words, or*
- *displays any writing, image or video which is threatening, abusive or insulting,*

if she intends thereby to stir up hatred against a religious group.
A person guilty of an offense under this law is liable for a prison term not exceeding six months or a fine or both.
This law does not prohibit or restrict discussion, criticism or expressions of antipathy, dislike, ridicule, insult or abuse of particular religions or the beliefs or practices of their adherents."

Having carefully read the content of this hate speech legislation, do you favor or oppose this law?

- ○ Very much oppose
- ○ Rather oppose
- ○ Neither oppose nor support
- ○ Rather support
- ○ Very much support

Figure 7: Design of hate speech regulation prime, German survey

HATE SPEECH REGULATION PRIME, GERMAN SURVEY

Wie Sie vielleicht gehört haben, bemüht sich die Regierung sehr ernsthaft Online-Hassrede zu bekämpfen. Das bedeutet, dass eine große Anzahl an Social-Media-Nachrichten mit beleidigenden oder hasserfüllten Inhalten gelöscht und strafrechtlich verfolgt werden.
Der Inhalt eines Hate-Speech-Gesetzes, das derzeit in der Diskussion ist, wird im folgenden Text beschrieben. Bitte lesen Sie den Text sehr sorgfältig und stellen Sie sicher, dass Sie ihn verstehen.

"Eine Person begeht eine Straftat, wenn sie eine Nachricht über eine Online-Plattform versendet, die

- *drohende, abwertende oder beleidigende Worte oder*
- *drohende, abwertende oder beleidigende Texte, Bilder oder Videos enthält*

und die Absicht hat, damit Hass gegen eine religiöse Gruppe zu schüren. Einer Person die sich im Sinne dieses Gesetzes strafbar macht, drohen eine Gefängnisstrafe von maximal sechs Monaten oder eine Geldstrafe oder beides.
Das Gesetz verbietet nicht die Diskussion und Kritik, den Ausdruck von Ablehnung oder das Lächerlichmachen, Beleidigen und Abwerten von bestimmten religiösen Glaubensinhalten oder -praktiken ihrer Anhänger."

Nachdem Sie den Inhalt des Hate-Speech-Gesetzes sorgfältig gelesen haben, unterstützen Sie das Gesetz oder lehnen Sie es ab?

- ○ Stimme überhaupt nicht zu
- ○ Stimme eher nicht zu
- ○ Teils/teils
- ○ Stimme eher zu
- ○ Stimme voll und ganz zu
- ○ Weiß nicht

Figure 8: 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

Additional items. In addition to the three core components of our research design, we will rely on further survey items for subgroup analyses. These are described further below.

3.3 Quantities of interest

Based on our research design ,we are able to identify several quantities of interest (see Table 2). The key design idea of our project is a systematic comparison of indirect and direct question items, i.e. Y_{true} and $Y_{reported}$. This allows us to elicit a behavioral expression of *preference falsification*, which will serve as the main outcome in our study:

$$\text{Preference Falsification} = \text{Y}_{\text{true}} - \text{Y}_{\text{reported}}$$

We refer to preference revealed in the double list experiment as the 'true' preference. Two assumptions are necessary to justify this. First, we have to assume that subjects respond truthfully to the list experiment, which cannot be tested (no liars). Second, we have to assume that the inclusion of the sensitive item does not affect the answer to the control items (no design effect). This assumption can be tested. The answers to the direct items are considered the reported preferences.

Our research design opens up the opportunity to not only assess whether preference falsification occurs and to what degree, but to manipulate this self-censoring behavior in a priming experiment. It is this option that we will leverage to study the effects of hate speech regulation. Comparing respondents who received the prime to those who received no prime, we are able formulate three causal quantities of interest: a) the *difference in true preferences*, b) the *difference in reported preferences*, and c) the *difference in preference falsification*.

Since in our setup the priming experiment comes after the double list experiment, the true preferences are pre-treatment variables. We therefore cannot estimate a causal effect of the prime on true preferences. Instead, we expect to find no difference (assuming that the randomization worked and the primed and non-primed groups are balanced). However, we are able to identify a causal effect of priming hate speech legislation on reported preferences. Our main quantity of interest is how the primed and non-primed groups differ in terms of preference falsification. This quantity is essentially a difference-in-differences and directly operationalizes our main hypotheses. For H1 we will rely on the sensitive 'Muslim' item, and for H2 we will use the sensitive 'Free Expression' item, respectively. In both instances, we expect the difference to be negative.

Table 2: Identification of quantities of interest

	Prime	**No Prime**	**Difference Prime - No Prime**	**QOI**
Indirect	$\mathrm{Y}^{\text{prime}}_{\text{true}}$	$\mathrm{Y}^{\text{no.prime}}_{\text{true}}$	$\mathrm{Y}^{\text{prime}}_{\text{true}} - \mathrm{Y}^{\text{no.prime}}_{\text{true}}$	*Difference in true preference*
Direct	$\mathrm{Y}^{\text{prime}}_{\text{reported}}$	$\mathrm{Y}^{\text{no.prime}}_{\text{reported}}$	$\mathrm{Y}^{\text{prime}}_{\text{reported}} - \mathrm{Y}^{\text{no.prime}}_{\text{reported}}$	*Difference in reported preference*
Difference Indirect - Direct	$\mathrm{Y}^{\text{prime}}_{\text{true}} - \mathrm{Y}^{\text{prime}}_{\text{reported}}$	$\mathrm{Y}^{\text{no.prime}}_{\text{true}} - \mathrm{Y}^{\text{no.prime}}_{\text{reported}}$	$(\mathrm{Y}^{\text{prime}}_{\text{true}} - \mathrm{Y}^{\text{no.prime}}_{\text{true}}) - (\mathrm{Y}^{\text{prime}}_{\text{reported}} - \mathrm{Y}^{\text{no.prime}}_{\text{reported}})$	*Difference in preference falsification*
	Preference Falsification			

4 Pretest

4.1 Setup

To pretest our double list experiments and the hate speech legislation prime, we ran an initial survey experiment 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 relative 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 three main goals. First, we aim to establish the prevalence of the control items used in the double list experiments. This will tell us about the potential threat of floor or ceiling effects. Second, we assess the prevalence and actual sensitivity of our two sensitive items. This will tell whether we are actually capturing controversial opinions. Third, we generate first evidence on whether we are able to experimentally manipulate our respondents by randomly assigning the hate speech regulation prime.

Participants. We recruited a total of $N = 511$ 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. $N = 503$ participants completed the survey and we will only rely on these completed responses. The average completion time was 3.4 minutes with a standard deviation of 2.9 minutes.

To check for respondents attention, we included an attention check just before the hate speech regulation prime. 97 percent of the respondents ($N = 489$) passed the attention check. We will assess whether results change if we exclude inattentive respondents below.

4.2 List experiment: results

Combining the two lists 1A and 1B into a double list experiment yields an averaged support for the statement "Islam out of USA. Protect the American people" of 18 (+/- 4) percent (see Table 3). However, this statement does not appear to be particularly sensitive in our sample.

The difference when compared to the direct item is small and not significant, suggesting that there is not much self-censoring with regard to this statement. This is remarkable given the fact that the sensitive item mirrors an actual statement that lead to a conviction under British hate speech legislation and that was upheld by the European Court of Human Rights (Norwood vs. United Kingdom 2003).

Table 3: Pretest evaluation: outcomes of list experiment 1

Double List Experiment 1: Results		
List 1a	Est	SE
Mean Number of Items Treatment (N=255)	1.88	.05
Mean Number of Items Control (N=248)	1.71	.04
Prevalence of Sensitive Item Indirect	.17	.07
Prevalence of Sensitive Item Direct	.23	.02
Difference (Self-Censorship)	-.06	.07
List 1b	Est	SE
Mean Number of Items Treatment (N=248)	1.74	.06
Mean Number of Items Control (N=255)	1.55	.05
Prevalence of Sensitive Item Indirect	.19	.08
Prevalence of Sensitive Item Direct	.23	.02
Difference (Self-Censorship)	-.04	.08
List 1a and List 1b Combined	Est	SE
Prevalence of Sensitive Item Indirect	.18	.04
Prevalence of Sensitive Item Direct	.23	.02
Difference (Self-Censorship)	-.05	.05

Combining the two lists 2A and 2B into a double list experiment yields an averaged support for the statement "People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people" of 72 (+/- 4) percent (see Table 4). Interestingly, this indirectly elicited support is lower than the support expressed in a direct question where 80 (+/- 2) percent claim to support this statement. Although the difference is not statistically significant in our pretest sample (but see the power analysis below), it may

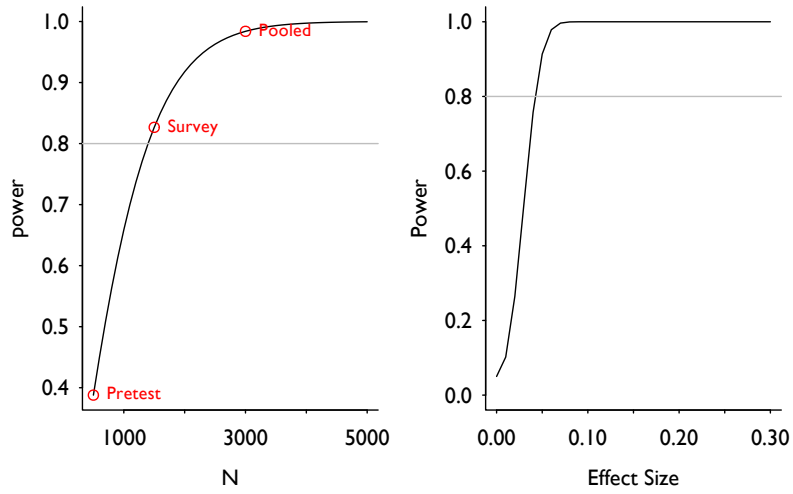
suggest that respondents feel compelled to take a stronger pro free-speech stance because it corresponds to a broadly shared cultural norm in the US context. It will be interesting to contrast this finding with a sample from Germany, where speech norms are generally thought to be different and freedom of speech less sacrosanct.

Table 4: Pretest evaluation: outcomes of list experiment 2

Double List Experiment 2: Results		
List 2a	Est	SE
Mean Number of Items Treatment (N=254)	2.66	.05
Mean Number of Items Control (N=249)	1.95	.04
Prevalence of Sensitive Item Indirect	.71	.06
Prevalence of Sensitive Item Direct	.80	.02
Difference (Self-Censorship)	-.09	.06
List 2b	Est	SE
Mean Number of Items Treatment (N=249)	2.33	.06
Mean Number of Items Control (N=254)	1.60	.04
Prevalence of Sensitive Item Indirect	.73	.07
Prevalence of Sensitive Item Direct	.80	.02
Difference (Self-Censorship)	-.07	.07
List 2a and List 2b Combined	Est	SE
Prevalence of Sensitive Item Indirect	.72	.04
Prevalence of Sensitive Item Direct	.80	.02
Difference (Self-Censorship)	-.08	.05

A preliminary power analysis based on the estimates and variances in the pretest suggests that moving to our actual US and German population surveys of roughly $N = 1,500$ respondents each, we should be able to detect a significant self-censoring effect on the item "People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people" in both countries (see Figure 9).

Figure 9: Power analysis for self-censoring effect



4.3 Priming experiment: results

53 percent of the respondents ($N = 269$) were randomly assigned to the hate speech legislation prime. Of those, 57 percent said they supported the hate speech law ($N = 153$). In addition to the attention check item, we tracked the time the respondents spent on the page containing the hate speech law prime. On average, respondents took 46.9 seconds to read through the text describing the fictitious hate speech law ($SD = 47.6$ seconds). This seems a reasonable amount of time and suggests that, generally speaking, the priming message was successfully delivered to the respondents.

Difference in reported preferences. Next, we assess whether being primed with the hate speech law has an effect on respondents expressed support for four items on religion and freedom of speech. Two of these items are the sensitive items used in the list experiments. Thus, we are also in the position to test whether the prime affects self-censoring behavior. The effects (group differences) are reported in Table 5.

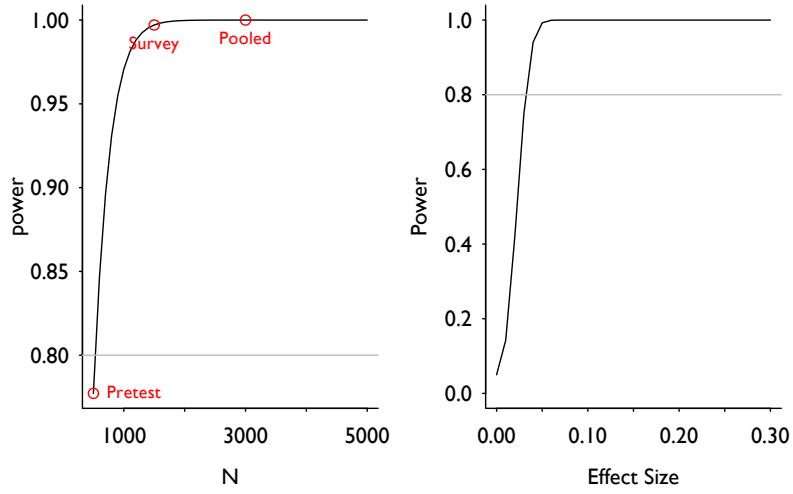
We find that priming hate speech regulation does not affect the expression of statements on religion. In particular, it does not affect the open expression of the controversial item "Islam out of USA. Protect the American people." This is in itself an important result because it suggests that informing people of hate speech legislation does not induce a change in their stated opinion concerning the protected group. Of course, since the law protects religious groups but not religious faiths, it may be worthwhile to check how they would

answer a direct item that says "Muslims out of USA. Protect the American people" when primed.

In contrast to the religious items, we do find that the prime affects support for statements on free expression. Interestingly, respondents are less supportive of free expression when primed with a fictitious hate speech law. They are significantly less likely to say that "it is important that people can use the Internet without censorship" (-6 percentage points, $p < .05$) and significantly less likely to state that "people should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people" (-10 percentage points, $p < .01$). This suggests that hate speech legislation has the potential to change public opinion about the limits of free expression, where people tend to follow the thrust of the law.

We also conducted a power analysis for the priming effect on the direct items, based on the observed effect and variance on the fourth direct item as found in the pretest. The power of the pretest was already at .78, the expected power for a full country survey (assuming a minimum of $N = 1500$) would be 0.99 (see Figure 10). We would even be able to reliably detect smaller priming effects.

Figure 10: Power analysis for priming effect on direct item



Difference in preference falsification. Our main quantity of interest that captures the hypothesized effect is the difference in self-censoring behavior between those primed with

Table 5: Pretest evaluation: effect of prime on direct responses

	No Prime	Prime	Effect
	(N=234)	(N=269)	
People should be able to practice their religion freely in our country.	.90 (.02)	.90 (.02)	.00 (.03)
Islam out of USA. Protect the American people.	.24 (.03)	.22 (.03)	-.02 (.04)
It is important that people can use the Internet without censorship.	.92 (.02)	.86 (.02)	-.06 (.03)*
People should be allowed to express unpopular opinions in public.	.85 (.02)	.75 (.03)	-.10 (.04)**

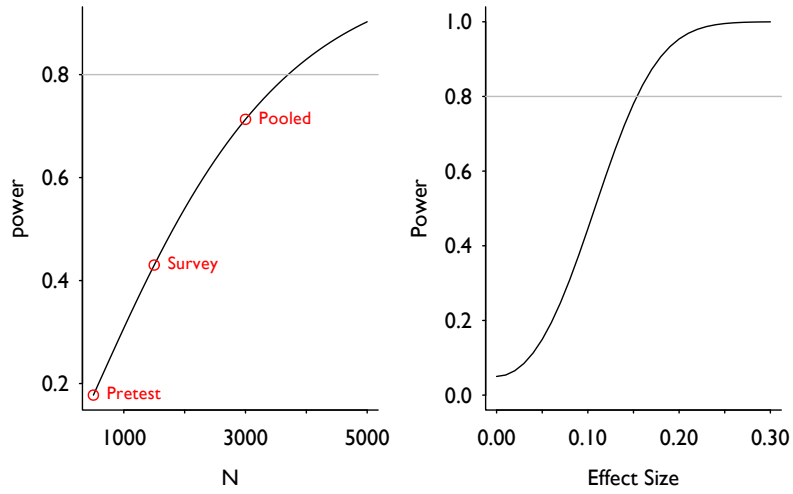
Table 6: Pretest evaluation: effect of prime on preference falsification

	Prime	No Prime	Effect
Indirect Sensitive Item	.67	.77	-.10
	(.06)	(.06)	(.09)
Direct Sensitive Item	.75	.85	-.10
	(.03)	(.02)	(.04)
Difference ("Self-censoring")	-.08	-.08	.00
	(.07)	(.07)	(.09)

hate speech legislation vs. those that did not receive this prime. This is essentially a simple difference-in-differences estimate (see Table 6). Since we found prime effect on the 'Muslim' item, we will restrict ourselves to the 'Free Expression' item. However, in our pretest we were unlucky because already the pre-treatment indirect double-list estimates were highly imbalanced such that they cancelled out the differences in the direct item *exactly*. (It should be noted that we still find a significant difference in the direct item when controlling for this pre-treatment imbalance, i.e. they are themselves not spurious.)

To conduct a power analysis, we fixed the pre-treatment indirect estimates to be equal at .72 (thus yielding a diff-in-diff of around .10 with standard error of .09) and else relied on the observed prevalence of the direct item as well as all as the observed variances and covariances (see Figure 11). The power analysis suggests that we may need considerable more observations than our two country surveys provide, although pooling the German and US sample (assuming roughly $N = 3000$ respondents) would yield a power of greater than .70. Alternatively, the minimum effect of priming hate speech on self-censoring we would be able to detect with a power of .80 in a single country is roughly .15.

Figure 11: Power analysis for priming effect on preference falsification



5 Analysis Plan

This section describes how we plan to analyze the data collected through the two panel surveys in the US and Germany.

5.1 Design

Experimental setup. In light of the pretest results, we have adjusted the sensitive item in the first double list experiment. It now reads "Muslims out of USA. Protect the American people." This should be more sensitive and lead to higher levels of preference-falsification. In addition, we have slightly changed the wording of the hate speech legislation prime to avoid deception. The scale for the support or opposition to the proposed hate speech law, which was binary in the pre-test, is now a five-point scale.

Planned sample. We include the experiment in two existing panel surveys and will rely on the final samples provided by YouGov. A power analysis (see further above) suggests that these samples will be big enough to detect effects that are equal to or greater than ten percentage points difference in item support between treatment and control group. 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.

Exclusion criteria. We will only exclude respondents with missings in the outcome measures elicited through the two double list experiments. Missingness in pre-treatment covariates will be dealt with using 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.

5.2 Confirmatory analyses

Establishing base rates of preference-falsification. In a first step, we will establish the extent of preference-falsification related to the two sensitive items. To do this, we first calculate the prevalence of the sensitive items using a simple differences-in-means estimator between treatment and control groups, averaging over the two lists of each item. The level of preference-falsification is derived by the difference between this prevalence and the support for the matching direct items. We will calculate these key quantities separately for Germany and the US and test for differences between these two country contexts.

Test of main hypotheses. In order to test the two main hypotheses, we will first compare support for the four direct items between the primed and non-primed using simple differences-in-means. Statistical inference will be based on t-tests at the 5% α level. We will calculate this separately for the German and the US sample as well as for the pooled sample by using a country dummy.

H1 will be tested by comparing the level of preference-falsification for the item "Muslims out of USA. Protect the American people" between the primed and non-primed groups using simple differences-in-means (essentially resulting in a difference-in-differences). Statistical inference will be based the 5% α level. We will calculate this separately for the German and the US sample as well as for the pooled sample including a country dummy.

H2 will be tested by comparing the level of preference-falsification for the item "People should be allowed to express unpopular opinions in public, even those that are deeply offensive to other people" between the primed and non-primed groups using simple differences-in-means (essentially resulting in a difference-in-differences). Statistical inference will be based on the 5% α level. We will calculate this separately for the German and the US sample as well as for the pooled sample by using a country dummy.

We will also rely on the method proposed by Eady (2017) and include the prime as predictor in the regression models for preference-falsification. We will also run models including the following pre-treatment covariates (see below for the wordings of the items used to measure the variables):

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

These models will be run on separate and combined German and US samples.

Planned subgroup analyses. While we have not explicitly formulated conditional hypotheses, we intend to explore potential treatment heterogeneity in a series of subgroup analyses. In particular, we will assess whether the effect of priming hate speech legislation on preference-falsification is moderated by the pre-treatment covariates listed above. To operationalize these heterogeneous treatment effects, we will construct multiplicative interaction terms with the hate speech prime and include them in separate regression models as proposed by Eady (2017). These models will be run on separate and combined German and US samples. We will correct for multiple testing using Bonferroni correction.

HATE SPEECH EXPERIENCE

"Hate Speech" describes when someone is verbally attacked because of personal attributes, such as religion, ethnic origin, nationality, sex, or opinions. Please select all of the following that apply to you.

- ○ I have personally been verbally attacked with hate speech online.
- ○ I have experienced how others have been verbally attacked with hate speech online.
- ○ None of the above.

HATE SPEECH REGULATION PREFERENCES

Would you support or oppose a law that would make it illegal to make insulting or hateful statements about...

- Germans? (American Citizens?)
- Muslims?
- Jews?
- Women?
- Christians?
- Neo Nazis?
- the Government?

○ Very much oppose

○ Rather oppose

○ Neither/nor

○ Rather support

○ Very much support

○ Don't know

HATE SPEECH IDENTIFICATION

Which of the following would you label as hate speech?

- A person calling an ethnic minority a racial slur.
- A person calling a woman a vulgar name.
- A person who says that illegal immigrants should be deported.
- A person who says Germany/the USA is an evil country.
- A person who says Islam is taking over Europe/the USA.
- A person calling another person with conservative views a Nazi.

○ Hate speech

○ No hate speech

○ Don't know

RESPONSIBILITY FOR ACTION AGAINST HATE SPEECH

To what extent, if at all, do you think each of the following groups should take responsibility in taking steps against online hate speech?

- People who are victims of online hate speech
- Other users who witness the behavior
- Online services such as social media platforms or other websites
- Policymakers
- Law enforcement
- Employers of distributors of hate speech

○ No responsibility at all

○ Rather no responsibility

○ Some responsibility

○ Very much responsibility

○ Don't know

FEELING TOWARDS DISCUSSING POLITICS

When you discuss politics with others, how free or unrestricted do you feel?

○ I dont feel free to discuss it with anyone

○ I dont feel free to discuss it with many people

○ I feel free to discuss it with a few

○ I feel free to discuss it with anyone

○ I never discuss politics with other people

POLITICAL INTEREST

How regularly do you follow politics?

- ○ Most of the time
- ○ Some of the time
- ○ Only now and then
- ○ Hardly at all
- ○ Don't know

POLITICAL IDEOLOGY

In general, would you describe your political views as...

- ○ Very conservative
- ○ Conservative
- ○ Moderate
- ○ Liberal
- ○ Very liberal
- ○ Not sure

CONGRESS CONTROL PREFERENCE

Which party would you prefer to control Congress after the midterm elections?

- ○ Democrats
- ○ Republicans
- ○ Divided between House and Senate
- ○ None of the above

PRESIDENTIAL APPROVAL

Do you approve or disapprove of the way Donald Trump is handling his job as president?

- ○ Strongly approve
- ○ Somewhat approve
- ○ Neither approve nor disapprove
- ○ Somewhat disapprove
- ○ Strongly disapprove

SOCIAL MEDIA USE

Do you have accounts on any of the following social media services? (check all that apply):

- Twitter
- Facebook
- Instagram
- LinkedIn
- Snapchat
- WhatsApp
- Reddit

TWITTER USAGE FREQUENCY

In the last survey you told us that you have a Twitter account. Today we want to learn more about your Twitter use. How frequently do you:

1. Check Twitter
2. Post messages on Twitter

○ Almost constantly

○ Several times a day

○ About once a day

○ 3 to 6 days a week

○ 1 to 2 days a week

○ Every few weeks

○ Less often

○ Never

○ Don't know

FACEBOOK USAGE FREQUENCY

In the last survey you told us that you have a Facebook account. Today we want to learn more about your Facebook use. How frequently do you:

1. Check Facebook
2. Post messages on Facebook

○ Almost constantly

○ Several times a day

○ About once a day

○ 3 to 6 days a week

○ 1 to 2 days a week

○ Every few weeks

○ Less often

○ Never

○ Don't know

RACIAL RESENTMENT

Here you can find several statements with which some people agree while others do not. How about you? Please state your view on these issues.

1. Irish, Italians, Jewish and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors.
2. Generations of slavery and discrimination have created conditions that make it difficult for blacks to work their way out of the lower class
3. Over the past few years, blacks have gotten less than they deserve.
4. It's really a matter of some people not trying hard enough; if blacks would only try harder they could be just as well off as whites.

○ Strongly disagree

○ Somewhat disagree

○ Neither/nor

○ Somewhat agree

○ Strongly agree

○ Don't know

GENDER

What is your gender?

○ Male

○ Female

AGE

What is your age? OPEN ANSWER [years]

EDUCATION

Which of the following describes best your education?

- ○ No high school
- ○ High school graduate
- ○ Some college
- ○ 2-year
- ○ 4-year
- ○ Post-grad

RELIGION

What is your religion?

- ○ Protestant
- ○ Roman Catholic
- ○ Mormon
- ○ Eastern or Greek Orthodox
- ○ Jewish
- ○ Muslim
- ○ Buddhist
- ○ Hindu
- ○ Atheist
- ○ Agnostic
- ○ Nothing in particular
- ○ Something else

IMPORTANCE OF RELIGION

As how important do you consider religion?

- ○ Very important
- ○ Somewhat important
- ○ Not too important
- ○ Not at all important

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.

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

Eady, Gregory. 2017. "The Statistical Analysis of Misreporting on Sensitive Survey Questions." *Political Analysis* 25(2):241–259.

Gilens, Martin, Paul M Sniderman and James H Kuklinski. 1998. "Affirmative action and the politics of realignment." *British Journal of Political Science* 28(1):159–183.

Glynn, Adam N. 2013. "What can we learn with statistical truth serum? Design and analysis of the list experiment." *Public Opinion Quarterly* 77(S1):159–172.

Kuran, Timur. 1997. *Private truths, public lies: The social consequences of preference falsification*. Harvard 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.

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.

Tourangeau, Roger and Ting Yan. 2007. "Sensitive questions in surveys." *Psychological bulletin* 133(5):859.