Following the Norms or Following the Crowd?

Go Murakami (Ritsumeikan University)*
Yoshitaka Nishizawa (Doshisha University)**

A paper presented at the "virtual" annual meeting of the American Political Science Association, September 10-13, 2020.

ABSTRACT

We investigated the role of information about social norms in shaping people's policy preferences concerning the current hate speech situation in Japan. More specifically, using a web-survey experiment, we tested whether two types of information, bandwagon message and anti-discrimination message, would influence the respondents' attitudes towards government regulation against hate speech. Our analysis shows that the bandwagon message slightly changed some individuals' responses to hate speech regulation and made them conform to the majority opinion. To be more precise, those who are predisposed to give socially desirable answers were more influenced by the bandwagon information. The anti-discrimination message also slightly changed the individuals' responses but in the opposite direction.

* gmurakam(a)fc.ritsumei.ac.jp

** nishizawa(a)mail.doshisha.ac.jp

1. Introduction

How do anti-discrimination norms evolve in democratic societies? In what way do people conform to such norms? Would differences in types of normative messages matter in influencing public opinion on policies regarding minorities? In this paper, we try to provide some empirical evidence to consider these questions by taking advantage of the recent development of hate speech regulation in Japan.

In June 2016, the Japanese government adopted a hate speech prevention law for the first time. Some legal scholars and activists express doubts about the effectiveness of this law because the law lacks an "enforcement mechanism". It, however, gave Japan's local governments some foundations to legally control some of the vicious activities. Just a few years before the passage of this legislation, newspapers and TV stations, too, started to publicize the hate speech situation in Japan. Due to an increase in publicity both by the national/local governments and by mass media, the hate speech issue became at least "visible" to the general public, even if it did not make many Japanese citizens pay direct attention to it.

A series of questions, therefore, arises. Will anti-discrimination norms ever evolve in Japan? And if they do, do Japanese people conform to such norms? In a country where "group pressure" is said to be strong, do people conform to norms because they believe it is better to follow the crowd? Or do they accept and internalize norms because they believe in the principle of equality and they value human dignity?

We explored these questions by using a web-survey experiment.[1] In our experiment, we randomly assigned the respondents two types of messages, bandwagon and anti-discrimination messages. The first, the bandwagon message, contained only the numeric information indicating that the majority supports a policy regulating hate speech against minority groups. The latter, the anti-discrimination message, suggests that while it is desirable to respect each other mutually, hate speech induces discrimination in society. The bandwagon message minimizes the nature of the anti-discrimination norm, whereas the anti-discrimination message suppresses the majority view.

[1] This work was supported by JSPS KAKENHI, Grant Number JP17KT0005. Title of the project is "A Study on the Mental Foundation and Evolving Legal Norm Regarding Hate Speech in Japan." We would like to thank other research members who contributed to this study: Masaru Kohno, Kentaro Hirose (Waseda University), Ki'ichiro Arai (now Chuo University, and Tokyo Metropolitan University at the time of survey), Miwa Nakajo (Tsuda University), and Hae Kim (Chiba University). Editorial assistance by John McCall is also appreciated.

Our analysis shows that the bandwagon message slightly changed some individuals' responses to hate speech regulation and made them conform to the majority opinion. To be more precise, only those who are predisposed to give socially desirable answers were more influenced by the bandwagon information. The anti-discrimination message also slightly influenced the individuals' responses but in the opposite direction.

2. Literature review and theory

2-1. Social norm in social psychological research

Social norm is such a controversial concept that the debate on its definition generates some confusion (Shaffer 1983). Nevertheless, the idea is generally defined as a set of values or rules about what is (not) expected to be done, and such rules should widely be shared among members of a society (Cialdini and Trost 1998, 152; Elster 2007, 353; Sunstein 1996, 914). It is useful for our purpose to note that social psychologists distinguish descriptive norm from injunctive norm, the former meaning the perceptions of what most other people do, and the latter meaning the understandings of what other people approve or disapprove, in other words, what ought to be done (Cialdini, Kallgren, and Reno 1991, 202–3; Cialdini, Reno, and Kallgren 1990, 1015; Rimal and Lapinski 2015). Scholars seem to agree that both types of norms guide people to specific attitudes and behaviors, making people conform to what many others think or do (Deutsch and Gerard 1955). Still, they seem to be inconclusive as to which type is more likely to induce conformity. Some suggest a relative effectiveness of exposing a descriptive norm message to a mere injunctive one (Goldstein, Cialdini, and Griskevicius 2008), while many others suggest that either can change people's behavior when it becomes salient (Cialdini, Kallgren, and Reno 1991; Jacobson, Mortensen, and Cialdini 2011; Kallgren, Reno, and Cialdini 2000). Another set of studies suggest that both types of norm often work hand in hand in the same direction and that injunctive norm moderates the effect of descriptive norm (Rimal and Lapinski 2015, 397–98).

2-2. Bandwagon effect and the effect of social desirability in public opinion research

In public opinion research, influences of social norms on respondents have been conceptualized in two phenomena—bandwagon effect and effect of social desirability. They have, however, been studied independently from each other and have not been linked well to the literature on social norms.

In public opinion research the bandwagon effect is often defined as a change, at an individual level, in their political preference to the more popular one (Barnfield 2019; Schmitt-Beck 2015). While the most popular type of studies on the bandwagon effect dealt with the poll results on vote choice, some other studies examined the bandwagon effect on people's policy preferences (Marsh 1985; Mutz 1992; Nadeau, Cloutier, and Guay 1993; Rothschild and Malhotra 2014; Toff 2018). In either case, these works suggest that the bandwagon effect is small at best and that the effect largely depends on one's predispositions regarding the policy (Hardmeier 2008; Rothschild and Malhotra 2014). Another set of studies examined several possible mechanisms in which the bandwagon information changes one's policy preferences. Such mechanisms include feeling enthusiasm of being on the winning side or satisfaction about winning; eliciting one's policy considerations more, as if being "primed" by the bandwagon information; utilizing the bandwagon information as a shortcut for a "better" preference; or making rational or strategic choices to maximize chances of winning (Hardmeier 2008, 509–10).

In another set of public opinion research, the effect of considering social desirability refers to the change in respondents' answer in a survey as a reaction to a specific type of social pressure. This pressure grows when respondents consider their true answer would be socially undesirable (Tourangeau and Yan 2007, 860). A large body of research suggests that some types of questions elicit "fake" answers, the answers that conform to the social norm because respondents would like to form a good impression of themselves (Demaio 1985; Tourangeau and Yan 2007). An over-reporting of turnout in elections is an often-cited example (Silver, Anderson, and Abramson 1986). Contrary to their actual behaviors, abstainers typically report that they voted because they think that turnout is a civic obligation and answering so conforms to this norm.

Although these two types of social "pressures" operate differently with separate logic and mechanisms, their ostensible outcomes are the same: both bandwagon effect and the effect of social desirability work as a pressure upon people to conform to the social norm. As for the bandwagon effect, as if people jump on the bandwagon, they change their political view, during the survey interview, in the direction of the majority opinion (the bandwagon). They do so, trusting that the majority opinion reflects the social norm. As for the effect of social desirability, people provide answers that they think are socially desirable, even if such answers do not necessarily reflect their true opinion, attitude, or action. By giving a socially desirable response, people conform to the social norm.

Despite this similarity, there is an important difference between them: The types of the social norm that these two effects impose on people are different. The bandwagon effect operates with descriptive norms, while the effect of social desirability emphasizes injunctive norms. Descriptive norms mainly refer to the numeric aspect of the norm. Injunctive norms, on the other hand, concern the nature of the norm. For the bandwagon effect to occur, therefore, one only needs to know what the majority prefers on the issue (the descriptive norm) but does not need to comprehend the nature of the norm (the injunctive norm). In contrast, for people to express a socially desirable opinion, they need to understand the nature of the norm and its implications (the injunctive norm), while they do not necessarily be aware of the numeric aspect of the opinion (the descriptive norm).

This does not mean that the bandwagon effect lacks the injunctive information or that descriptive norms cannot elicit social desirability concerns.[2] Our point is that we can theoretically discern these two types of norms, and we need to pay attention to this fact when we discuss the bandwagon effect and the effect of social desirability.

2-3. Anti-discrimination norm and hate speech regulations in Japan

In June 2016, the Japanese government, for the first time, adopted the Hate Speech Prevention Law, formally titled "Act on the Promotion of Efforts to Eliminate Unfair Discriminatory Speech and Behavior against Persons Originating from Outside Japan". While there have been some doubts about the effectiveness of this law because the law was merely what is called a "principle law" and it lacked an "enforcement mechanism", some observers acknowledge that it "seems to have changed the attitude of the national and local governments and to have been functioning in a way that controls the tone of hate speech expressed in the public sphere" (Kotani 2017, 1). In fact, some of the local governments, such as Kawasaki City and Osaka City, have recently introduced a legal mechanism to enforce the "principle" of the law (Kawanishi 2018).[3]

How have ordinary Japanese citizens reacted to the introduction of this law? According to the public opinion survey conducted by the Japanese government in 2017, a majority (57%) of Japanese knew about hate speech activities, including demonstrations,

[2] In fact, they are typically present together. In studying the effect of the social norm of turnout as the socially desirable behavior (civic obligation) for example, a mixture of both descriptive and injunctive norm information is often used to pressure voters to turnout (Gerber and Rogers 2009).

[3] See also, for example, “Osaka enforces Japan's first ordinance against hate speech, threatens to name names,” *Japan Times*, July 1, 2016; and "Kawasaki enacts Japan's first bill punishing hate speech" Kyodo News, December 12, 2019.

meetings and vans driven with agitating messages blaring through loudspeakers in public. This figure, however, does not indicate the degree to which an anti-discrimination norm is accepted among Japanese. In fact, the term "hate speech" is relatively new to Japanese citizens. Prior to 2013, the term rarely appeared in Japanese newspapers, even for the major ones with nation-wide circulation, such as Asahi and Yomiuri (Hirose, Kim, and Kohno 2020; Kohno and Nishizawa 2019). Given that the hate speech issues have caught the attention of the general public just recently, it is fair to say that a consensus of what consists of hate speech has yet to be reached in Japan. Naturally, many Japanese citizens are still ambivalent about how much hate speech ought to be regulated by the government.

The idea that one should not discriminate against others by their race/ethnicity, gender, culture, religion, and other backgrounds, is widely understood, accepted, and even entrenched in many democratic states by their laws and the International Convention. Several studies show that recognition of such a norm can change people's attitudes towards minorities[4] (Blinder, Ford, and Ivarsflaten 2019; Zitek and Hebl 2007). These findings suggest two questions worth investigating: First, how much does exposure to the anti-discrimination norm change the public opinion on *policy preferences* involving minorities in order for their expressed opinion to be consistent with the norm? And second, which type of anti-discrimination norm message—descriptive or injunctive—contributes more in changing their policy preferences?

The first question arises because of the unique nature of policy preference discussed here. Accepting an anti-discrimination norm does not directly translate into a specific policy preference. Preference on hate speech regulations for example, can vary from cherishing freedom of speech without any restrictions on one hand, to thorough restrictions on hate speech against minorities by law on the other. It is well possible that those who internalize the anti-discrimination norm may still oppose legal restrictions. In the end, conformity to the anti-discrimination norm is conceptually distinct from preferring restrictions on hate speech. Whether prompting the norm leads people to prefer restrictions on hate speech is an open empirical question.

[4] On the one hand, such an effect can be the result of people's internalization of the anti-discrimination norm, but on the other hand, people may just express their view in accordance with the norm without accepting it. While these two dimensions of the conformity—compliance and private acceptance—should be distinguished in order to better understand what it means to conform to the norm (Kiesler and Kiesler 1969), it is extremely difficult to empirically distinguish the two in a survey. Thus in our study, we presume that conformity can take both forms, when people change their views as the result of their exposure to the injunctive norm.

The second question rests on the assumption that the type of message to which people are exposed can be descriptive, injunctive, or both. A bandwagon type of message focuses on descriptive information by stating that *the majority* prefers imposing restrictions on hate speech. It does not characterize the substantive content of the anti-discrimination norm. On the other hand, an anti-discrimination type of message highlights its injunctive nature of the norm by stressing the point that hate speech produces discrimination and thus ought to be prevented. It does not refer to the majority views on hate speech regulations.

As indicated earlier, the hate speech issue has just begun to penetrate, if at all, into the general public in Japan. Whether or not the link between the anti-discrimination norm at the society level and the policy preference about the hate speech regulation at an individual level has been established is far more unclear. In a sense, the Japanese are "naive" concerning the hate speech issue. To that extent, the Japanese must be more susceptible to the anti-discrimination norm than the citizens in some other countries where hate speech problems have been discussed much longer. For that reason, we believe that Japan would be an ideal setting to test these two questions empirically.

3. Propositions to be tested

3-1. Effect of bandwagon message and anti-discrimination message

To address the above two questions prompted by the theoretical discussion, we will test the following two empirical propositions.

Proposition 1: Respondents who received a bandwagon message which emphasizes that the majority supports hate speech regulations are more likely to favor such policies than those who did not receive this message.

Proposition 2: Respondents who received an anti-discrimination message which emphasizes the injunctive norm are more likely to favor policies to regulate hate speech than those who did not receive this message.

By estimating the equations that correspond to each proposition, we can test if these two normative messages would influence the likelihood of supporting government regulations against hate speech. By comparing the magnitudes of the estimated effects between the two, we will know which type of norm, descriptive or injunctive, is more influential.

3-2. Moderated effects

In addition, we will consider the following propositions.

Proposition 3: People have different predispositions with respect to their susceptibility to the different types of norm, and the effects of both bandwagon messages and anti-discrimination messages are moderated by these predispositions.

While there are many possible moderating variables to be considered, we will focus on those predispositions that, we believe, directly influence the effects of two types of messages. As for the bandwagon message, we expect that its impact should be larger among those who, in general, tend to follow the majority opinion. In other words, those who prefer siding with the majority opinion, or who tend to perceive popular views as more attractive would be more likely to conform to the norm due to the bandwagon message.

On the other hand, the effect of the anti-discrimination message should be larger among those who, in general, tend to give a socially desirable response when they are asked to express their opinion. Some social psychology studies suggest that specific personality traits increase a readiness to give socially desirable responses (Demaio 1985). One noteworthy characteristic, among such personality traits, is impression management (Paulhus 1984, 1991). We expect that those who care more about their own impression would be more likely to conform to the norm due to the anti-discrimination message.

4. Data and methods

We conducted a web-based experimental survey between 10 and 14 April 2019.[5] While 4,940 Japanese adults aged between 20 and 69 completed this survey, we excluded "satisficers" from these samples based on two screening questions, which resulted in 4,241 participants as valid samples. We designed and programmed the survey using Qualtrix, whereas Nikkei Research Inc. conducted and supervised the administration of the survey. Nikkei Research Inc. recruited the participants from its associated online panel based on their stratified random sampling procedure so that the participants approximate the distribution of the Japanese population by gender, age group, and prefecture of residence. In the valid sample, men (57% compared to 43% female participants) and those in 50s (25%) and 60s (23%) are slightly over-represented compared to the distribution in the Japanese population (Statistics Japan 2020), and the

[5] The title of the Web survey is: "Public Opinion towards Japan and the International Society."

age group of over 70s are not included at the design stage.

4-1. Experimental protocol

Our survey started with a few questions that asked participants about their demographic backgrounds. We then measured their propensity to give socially desirable responses by using a set of 12 questions from Paulhus's (1991) Balanced Inventory of Desirable Responding (BIDR).[6] It is followed by a series of questions we developed to measure the respondents' tendency to follow the majority opinion or popular political choices. These two sets of questions represent the moderating variables, which we discuss in section 3-2. After answering a few more questions on morality, law and order, free speech, and trust in groups and institutions, the respondents read the experimentally manipulated text.

The experiment is a two-by-five factorial design with a total of 10 conditions. While the respondents are reading a brief description of the hate speech situation in Japan, they are randomly assigned into two experimental groups. The first group reads the description about hate speech against "Korean residents" (hereafter *Zainichi* Koreans), while the second group reads a similar description against "those with disabilities" (see Appendix for this detail). Immediately after this, the respondents were randomly assigned to one of five sub-experimental groups by seeing (1) no message (control), (2) no message but describing hate speech as abusive words (abuse), (3) bandwagon-poll message, (4) no bandwagon-poll message and (5) anti-discrimination message. The first two groups did not read any more texts, whereas the last three groups read an additional text that implies a different type of social norm. The bandwagon-poll message (the third group) says:

> By the way, last year we conducted a public opinion survey of voters in all prefectures in the country (some 3,000 people) and asked whether the government should regulate such activities. We obtained, among those who expressed a “yes/no” opinion to this question, the following result:
>
> In favor of regulation: 78%
> Against regulation: 22%
>
> This result shows that there are quite a few people who think that it should be regulated.

[6] We used the Japanese version, BIDRJ, translated by Tani (2008).

This text focuses on the descriptive norm information that the majority favors regulation. The respondents in the fourth group saw the almost identical script, but the poll results were modified as "for regulation: 54%". They then read the revised description saying, "[t]his result shows that regulation is supported by a little over half of the people, and not necessarily by a great majority". The fourth experimental group was necessary because the poll information itself can encourage people to reconsider their policy preferences, even though the specific opinion was not supported by the majority.

The respondents in the fifth group saw the anti-discrimination message,

> By the way, some people point out that such words and behaviors can create a sense of discrimination against the targeted groups. In order to foster a good society, “acknowledging differences among people and respecting each other” is considered desirable, even if it may not represent majority opinion in Japan.

The message clearly suggests what is socially desirable, while it also hints that this idea may not be supported by the majority. In order to increase the external validity of the message, a substantive portion of this message was adopted from the wordings in the public relations poster, "Hate Speech is Unacceptable," created and distributed by the Japanese Ministry of Justice (Ministry of Justice 2016).[7]

After reading the scripts on the hate speech situation in Japan, all respondents answered the same question on hate speech regulations:

> We ask you how the government should respond to this issue. Do you think that the government should impose restrictions on insults and incitements to violence against [*Zainichi* Koreans] / [those with disabilities]?[8] Or do you think that the government should not impose any restrictions?

The answer options range from zero (“should not impose any restrictions”) to six (“should impose thorough restrictions”) while three (“cannot say”) is a neutral position. This serves as a dependent variable in our study. By this question, we measure the

[7] We did not tell or imply however, that this message was associated with the Ministry of Justice in order to prevent from any booster effect of the message by some political authority (Cialdini 1985).
[8] The same target minority group that appeared in the experimentally manipulated descriptions of the situation on hate speech in Japan is assigned here. In other words, if a respondent saw *Zainichi* Koreans in the text that they read before this question, they see *Zainichi* Koreans as the target minority group in this question again.

respondents' policy preference for the level of government-imposed regulations on hate speech, which includes educating more about, censoring, outlawing and criminalizing hate speech. This policy preference can conceptually be discernable from the injunctive norm that hate speech should not be tolerated. The most frequent answer was four (about 31.5%) with the average of 3.65 and the standard deviation of 1.33. For this detail, see Appendix.

After this question, a few more questions followed, which included asking about respondents' perceptions of the hate speech problem in Japan, liberal-conservative ideology, partisanship and other social background information (education, income and job status). At the end of the survey, respondents were thanked for their participation, and only those who completed all the questions were included as valid samples.

4-2. Statistical models

We use the OLS estimation of linear regression models to estimate the average treatment effects (hereafter ATE) of our randomly assigned scripts and messages.[9] First, the ATEs are estimated by the coefficient β of each dummy variable for the experimental treatment in the following model:

$$Y = \beta_0 + \sum_{i=1}^{8} \beta_i X_i$$

where Y is the dependent variable, β_0 is the constant, X are eight independent variables, distinguished by the subscript i. Five of the eight independent variables are dummies representing each treatment of (1) those with disabilities, (2) abuse, (3) bandwagon-poll message, (4) no bandwagon-poll message and (5) anti-discrimination message in the manipulated scripts. This makes "the script" with *Zainichi* Koreans and without any additional message as the reference group.

The other three independent variables are moderators to estimate moderated effects.[10] As the moderator variables that measure the level of favoring the popular opinion or choice, we constructed two scores, election bandwagon and opinion bandwagon. Another moderator that measures the respondents' susceptibility to the social

[9] We adopted the linear model instead of the ordered logit model for the ease of interpretation of the results. The results in the ordered logit model are too detailed to summarize in short, but the overall interpretation of them remains the same.

[10] Including these moderator variables in the model did not change the estimated ATEs compared to the model without them.

desirability effect was constructed as the impression management score from the 12 items in BIDR questions. All three moderator variables are factor scores produced from the confirmatory factor analyses (see Appendix for this detail). The election bandwagon and opinion bandwagon scores are highly correlated (correlation coefficient, r of 0.86), while they are not correlated with the impression management score (r=-.06 and -.05 respectively). A positive and larger value for the election and the opinion bandwagon scores represents favoring winning candidates and preferring the popular view, respectively. A positive and larger value for the impression management score suggests a tendency to manage their impression more. Moderated effects are calculated by adding an interaction term between a treatment variable and the relevant moderator variable(s) to the statistical model described above. For example, in order to estimate the moderated effects of the bandwagon-poll message, two interaction terms were included: (1) a bandwagon-poll message dummy and the election bandwagon score, and (2) a bandwagon-poll message dummy and the opinion bandwagon score. Details of each statistical model are available in Appendix.

5. Analysis

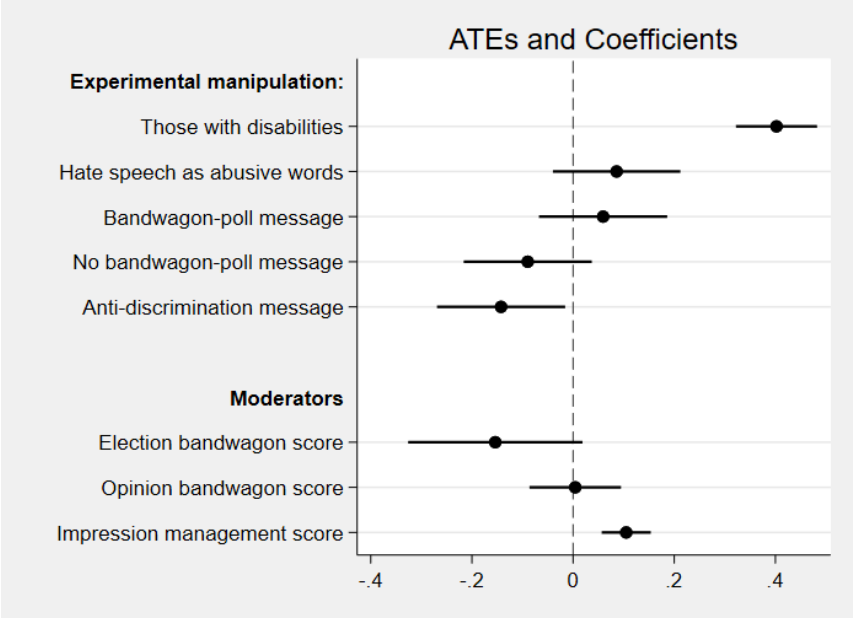
5-1. Average treatment effect of showing the normative messages

The result of the regression analysis for the ATEs is summarized in Figure 1. The five plots at the top represent the ATE of each treatment, and the three plots at the bottom are the coefficients of the three moderator variables. While we find a positive treatment effect of manipulating the target minority group from *Zainichi* Koreans to those with disabilities, due to the space limitation, we solely focus on the ATEs relevant to our propositions. First, we find that the estimated average effect of showing the bandwagon-poll message (third from the top) is so tiny that it is statistically indistinguishable from zero (β=.06, p=.36). In other words, the respondents who saw the bandwagon-poll message favored the same level of the hate speech regulations as those who did not see any message. This rejects our first proposition.[11]

Second, we find that the estimated ATE of showing the anti-discrimination message (fifth from the top) is *negative* (β=-.14, p=.03), which suggests that the respondents who saw the anti-discrimination message favored fewer restrictions on hate speech than those who did not see any messages. This result runs contrary to our preposition 2.

[11] Because the ATE of no bandwagon message is estimated negative (β=-.09, p=.17), we also examined the difference between bandwagon-poll group and no bandwagon-poll group, and found it was positive and statistically significant (β=.15, p=.02).

Figure 1. ATEs of experimental treatments and coefficients of three moderators



This analysis shows that neither message, bandwagon type or the anti-discrimination type, could change the respondents' policy preferences to the direction of their conforming to the norm. The effect of showing the anti-discrimination message was opposite to our prediction, but the overall effect of exposing the individual with the normative message is small (within plus or minus .02, while our dependent variable ranges from 0 to 6).

5-2. Moderated effects

Even if the average effect is zero, the normative message may have influenced the respondents' preference differently: some may have changed it in the direction of conformity, while others may have changed it in the opposite direction. If this is the case, the effect of the normative message on respondents' policy preferences cancelled each other, which resulted in the null finding of the ATEs. Accordingly, we should examine the possible respondent heterogeneity by analyzing the moderated effects by respondents' predisposition to conformity.

Figures 2 and 3 show the different effects of the bandwagon-poll message moderated by the respondents' propensity to favor winning candidates or party (election bandwagon score) and their propensity to favor the opinion supported by the majority (opinion bandwagon score) respectively. In each figure, dots in the left panel show the predicted values of the dependent variable by experimental conditions (bandwagon-poll message

Figure 2. Effects of bandwagon-poll message moderated by the election bandwagon score

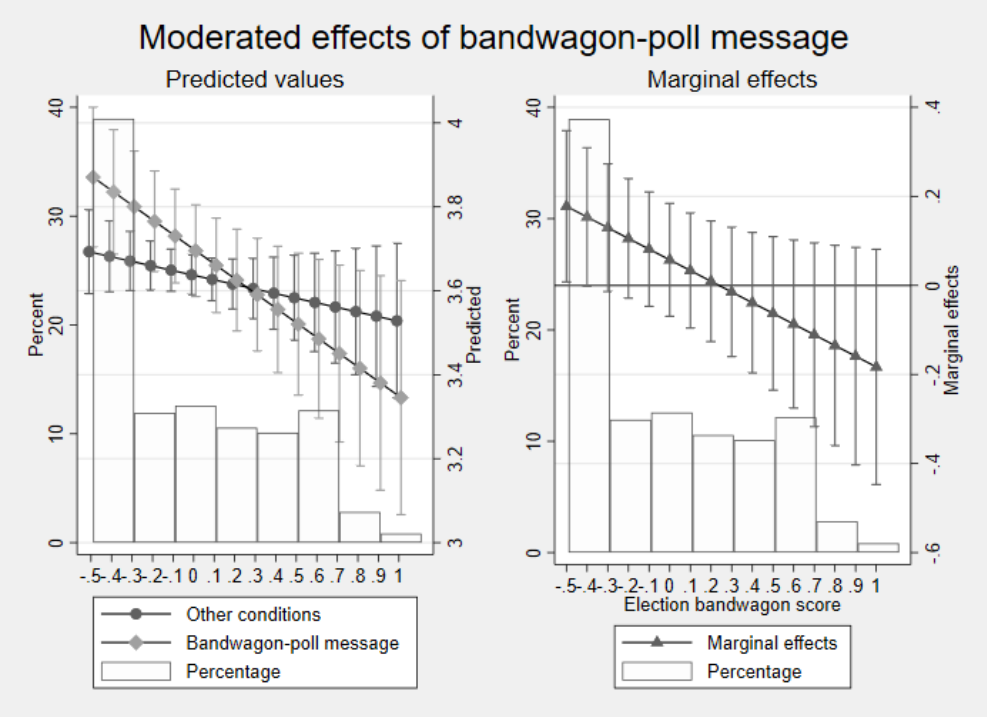
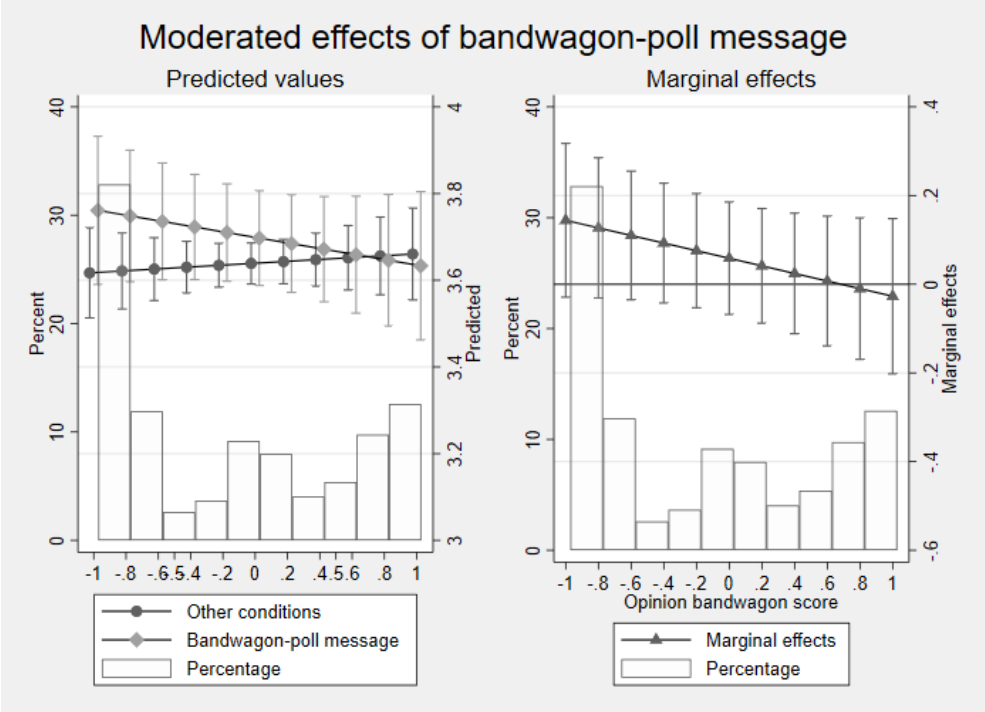


Figure 3. Effects of bandwagon-poll message moderated by the opinion bandwagon score



and others), whereas the dots in the right panel show marginal effects, both by different level of bandwagon score (horizontal axis). The distribution (percentage) of the election or opinion bandwagon score is also shown by histogram bars in each panel. About 30 to 40 percent of the respondents concentrate on the lower side of the scale, which suggests that many answered that they do not like to conform to the majority view or to side with the winning candidates.

If our third proposition holds, then the positive marginal effect should be magnified, as the respondents' election or opinion bandwagon score increases, because those who favor the majority views should be most likely influenced by our bandwagon message. Figures 2 and 3 suggest quite the opposite to this prediction: among the higher score of election or opinion bandwagon predisposition, the respondents in the bandwagon-poll message group preferred the same level of (or even fewer) hate speech restrictions than the respondents in the other experimental conditions. For example, for those with the election bandwagon score of 0.8, the marginal effect of the bandwagon-poll message was -.13 (p=.24) in Figure 2 (the right panel), and for those with the opinion bandwagon score of 0.8, the marginal effect was -.01 (p=.09) in Figure 3 (the right panel). Contrary to our prediction, some positive effects of showing the bandwagon-poll message are observed among those with the lowest election/opinion bandwagon score.

We also examined the moderated effects of the bandwagon-poll message by the impression management in Figure 4, even though this is outside our original proposition. The right panel of this figure shows a positive marginal effect of the bandwagon-poll message among the respondents with the higher score of impression management. For example, the marginal effect of the bandwagon message was .18 (p=.05) among those with the impression management score of 1.0, and .30 (p=.03) among those with the score of 2.0 (the right side of the spectrum in Figure 4). In short, the bandwagon-poll message made only the respondents with the higher impression management score favor hate speech restrictions more.

What about the moderated effect of the anti-discrimination message by the impression management score? Figure 5 illustrates such effects, but they do not endorse our proposition, either. The right panel of Figure 5 suggests that the marginal effect gets closer to zero, as the respondents' impression management score increases. This means that among those with the highest score for impression management, preferences for hate speech regulations were estimated the same between those who saw the anti-discrimination message and those who did not. Although the average effect of showing

Figure 4. Effects of bandwagon-poll message moderated by impression management score

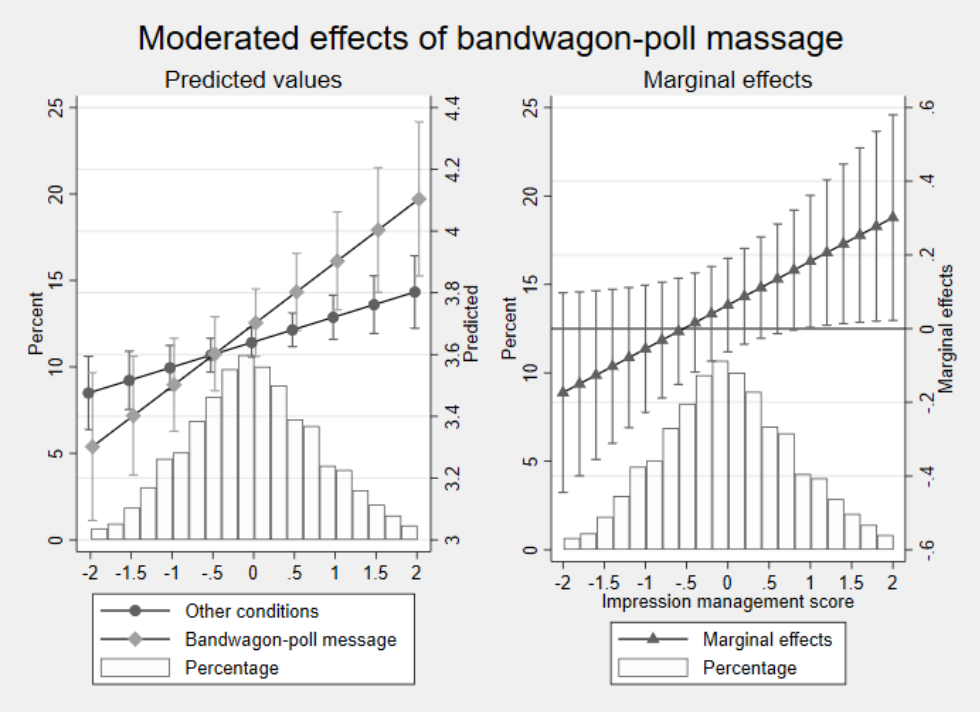
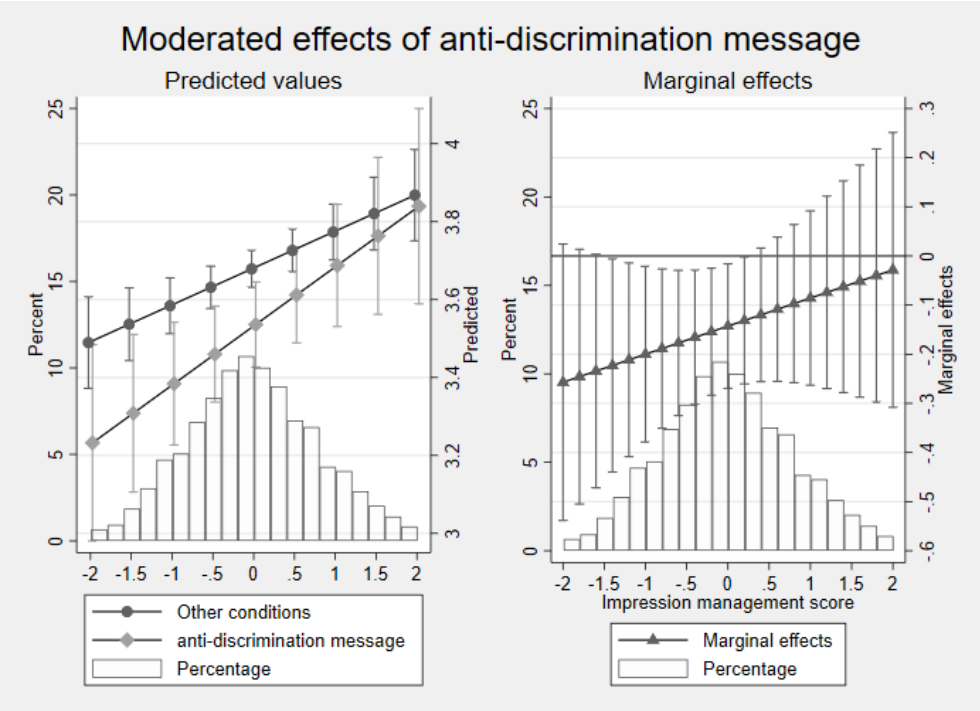


Figure 5. Effects of anti-discrimination message moderated by impression management score



this message is negative, the net zero effect among the highest level of this score rejects our third proposition.

In summary, we found no significant ATE of the bandwagon-poll message, but a small negative ATE of the anti-discrimination message. In both cases, contrary to our propositions, the respondents did not change their policy preferences in the direction of greater support for hate speech regulations. Our analysis of the moderated effects shows that neither message changed the respondents' preferences even among those who are considered more susceptible to each normative message: among the respondents with the higher election or opinion bandwagon score, the level of support for hate speech regulations was the same between bandwagon-poll message group and other (most notably no message) groups. Furthermore, as the respondents' score of impression management increased, their policy preference for hate speech regulations became closer between the anti-discrimination message group and other groups.

6. Discussion

Before discussing implications of our findings, we note some possible interpretations of the negative effect of the anti-discrimination message. Even if the effect size is small, the result shows that some respondents who saw the anti-discrimination message, the expression of which the Japanese Ministry of Justice used for its promotion flyer against hate speech, preferred fewer restrictions on hate speech than those who did not see the message. If this implies that the exposure to an injunctive message encourages people to be contrarian, we should reconsider if our prediction of conformity is theoretically valid. But an alternative interpretation is possible. The respondents who saw the message may have affirmed its anti-discrimination norm, but subsequently thought that if everyone "acknowledges differences among people and respects each other", the government-imposed restrictions would not be necessary anymore. If this was the case, we should re-examine the theoretical relationship between social norm and policy preferences: for individuals, implications of complying with some injunctive norm message for their policy preferences may be open to their interpretations.

The null findings of our analysis leave us several implications for both the theory and hate speech conditions in Japan. First, while the conceptual distinction between a descriptive and an injunctive norm is important, a simple descriptive or injunctive message alone may not be enough to change one's policy preferences. Because the

descriptive and injunctive norms often go together in reality (Rimal and Lapinski 2015), the next obvious question is whether the mixed message of both elements would produce more changes in policy preferences in the public opinion. Alternatively, in order to change one's policy views, a normative message may need some more additional information—source que of the message, identity of the "majority" who supports specific policy positions, effectiveness of the restrictions. Future research should explore and identify under what conditions the descriptive and injunctive norms message changes individuals' policy views in the direction of conformity.

Second, the null findings may reflect that the link between the general anti-discrimination norm and hate speech regulations is weak or ambiguous for many Japanese citizens, especially when the issue of hate speech is relatively new to them. This implies that the context of how well the specific policy issue (in this case hate speech regulations) is understood in relation to the anti-discrimination norm may condition the effect of norm messages on policy preferences in public opinion. In other words, in a society where the people quickly understand that complying with the anti-discrimination norm leads to imposing restrictions on hate speech, an exposure to the injunctive norm may make individuals favor a policy with more restrictions on hate speech. In future research, the implication of the anti-discrimination norm to the minority policies should be closely examined in different societies upon examining the effect of either type of normative message on policy preferences.

References

Barnfield, Matthew. 2019. "Think Twice before Jumping on the Bandwagon: Clarifying Concepts in Research on the Bandwagon Effect." *Political Studies Review*: 1478929919870691.

Blinder, Scott, Robert Ford, and Elisabeth Ivarsflaten. 2019. "Discrimination, Antiprejudice Norms, and Public Support for Multicultural Policies in Europe: The Case of Religious Schools:" *Comparative Political Studies*. https://journals.sagepub.com/doi/10.1177/0010414019830728 (December 31, 2019).

Cialdini, Robert B. 1985. *Influence: How and Why People Agree to Things*. New York: Quill.

Cialdini, Robert B., Carl A. Kallgren, and Raymond R. Reno. 1991. "A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior" ed. Mark P. Zanna. *Advances in Experimental Social Psychology* 24: 201–34.

Cialdini, Robert B., Raymond R. Reno, and Carl A. Kallgren. 1990. "A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places." *Journal of Personality and Social Psychology* 58(6): 1015–26.

Cialdini, Robert B., and Melanie R. Trost. 1998. "Social Influence: Social Norms, Conformity and Compliance." In *The Handbook of Social Psychology*, New York, NY, US: McGraw-Hill, 151–92.

Demaio, Theresa J. 1985. "Social Desirability and Survey Measurement: A Review." In *Surveying Subjective Phenomena*, Russell Sage Foundation, 257–82.

Deutsch, Morton, and Harold B. Gerard. 1955. "A Study of Normative and Informational Social Influences upon Individual Judgment." *The Journal of Abnormal and Social Psychology* 51(3): 629–36.

Elster, Jon. 2007. *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.

Gerber, Alan S., and Todd Rogers. 2009. "Descriptive Social Norms and Motivation to Vote: Everybody's Voting and so Should You." *The Journal of Politics* 71(1): 178–91.

Goldstein, Noah J., Robert B. Cialdini, and Vladas Griskevicius. 2008. "A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels." *Journal of Consumer Research* 35(3): 472–82.

Hardmeier, Sibylle. 2008. “The Effects of Published Polls on Citizens.” In *The SAGE Handbook of Public Opinion Research*, London: SAGE Publications Ltd, 504–14.

Hirose, Kentaro, Hae Kim, and Masaru Kohno. 2020. “Indignity or Offense? A Survey-Experimental Inquiry into Behavioral Foundations of Hate Speech Regulations.” A paper presented at the “virtual” annual meeting of the American Political Science Association, September 10-13, 2020.

Jacobson, Ryan P., Chad R. Mortensen, and Robert B. Cialdini. 2011. “Bodies Obliged and Unbound: Differentiated Response Tendencies for Injunctive and Descriptive Social Norms.” *Journal of Personality and Social Psychology* 100(3): 433–48.

Kallgren, Carl A., Raymond R. Reno, and Robert B. Cialdini. 2000. “A Focus Theory of Normative Conduct: When Norms Do and Do Not Affect Behavior.” *Personality and Social Psychology Bulletin* 26(8): 1002–12.

Kawanishi, Akihiro. 2018. “Hate Speech Regulation in Japan: Around Hate Speech Elimination Law.” *The Reference* 807: 51–73.

Kiesler, Charles A, and Sara Kiesler. 1969. *Conformity*. Reading, Mass: Addison-Wesley Pub. Co.

Kohno, Masaru, and Yoshitaka Nishizawa. 2019. “How Attitudes towards Hate Speech Regulation Is Determined [Heito Supiichi Kisei Eno Sanpi Ha Do Kimaru].” *Chuokoron* 133(4): 166–80.

Kotani, Junko. 2017. “A Comment on Hate Speech Regulation in Japan after the Enactment of the Hate Speech Elimination Act of 2016.” *The Journal of Law and Politics* 21(3–4): 228–218.

Marsh, Catherine. 1985. “Back on the Bandwagon: The Effect of Opinion Polls on Public Opinion.” *British Journal of Political Science* 15(1): 51–74.

Ministry of Justice. 2016. “Promotion Activities Focusing on Hate Speech.” http://www.moj.go.jp/ENGLISH/m_jinken04_00001.html (September 9, 2019).

Mutz, Diana C. 1992. “Impersonal Influence: Effects of Representations of Public Opinion on Political Attitudes.” *Political Behavior* 14(2): 89–122.

Nadeau, Richard, Edouard Cloutier, and J.-H. Guay. 1993. “New Evidence About the Existence of a Bandwagon Effect in the Opinion Formation Process.” *International Political Science Review* 14(2): 203–13.

Paulhus, Delroy L. 1984. “Two-Component Models of Socially Desirable Responding.” *Journal of Personality and Social Psychology* 46(3): 598–609.

———. 1991. “Measurement and Control of Response Bias.” In *Measures of Personality and Social Psychological Attitudes*, Measures of social psychological

attitudes, Vol. 1., eds. John P. Robinson, Phillip R. Shaver, and Lawrence Wrightsman. San Diego, CA, US: Academic Press, 17–59.

Rimal, Rajiv N., and Maria K. Lapinski. 2015. "A Re-Explication of Social Norms, Ten Years Later." *Communication Theory* 25(4): 393–409.

Rothschild, David, and Neil Malhotra. 2014. "Are Public Opinion Polls Self-Fulfilling Prophecies?" *Research & Politics* 1(2): 2053168014547667.

Schmitt-Beck, Rüdiger. 2015. "Bandwagon Effect." In *The International Encyclopedia of Political Communication*, American Cancer Society, 1–5. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118541555.wbiepc015.

Shaffer, Leigh S. 1983. "Toward Pepitone's Vision of a Normative Social Psychology: What Is a Social Norm?" *The Journal of Mind and Behavior* 4(2): 275–93.

Silver, Brian D., Barbara A. Anderson, and Paul R. Abramson. 1986. "Who Overreports Voting?" *American Political Science Review* 80(2): 613–24.

Statistics Japan. 2020. "Population by Age (Five-Year Groups) and Sex for Prefectures - Total population, Japanese population, October 1, 2019." *Population Estimates, Annual Report*. https://www.e-stat.go.jp/en/stat-search/files?page=1&layout=datalist&toukei=00200524&tstat=000000090001&cycle=7&year=20190&month=0&tclass1=000001011679 (July 1, 2020).

Sunstein, Cass R. 1996. "Social Norms and Social Roles." *Columbia Law Review* 96(4): 903–68.

Tani, Iori. 2008. "Development of Japanese Version of Balanced Inventory of Desirable Responding (BIDR-J)." *The Japanese Journal of Personality* 17(1): 18–28.

Toff, Benjamin. 2018. "Exploring the Effects of Polls on Public Opinion: How and When Media Reports of Policy Preferences Can Become Self-Fulfilling Prophesies." *Research & Politics* 5(4): 2053168018812215.

Tourangeau, Roger, and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133(5): 859–83.

Zitek, Emily M., and Michelle R. Hebl. 2007. "The Role of Social Norm Clarity in the Influenced Expression of Prejudice over Time." *Journal of Experimental Social Psychology* 43(6): 867–76.

Appendix

1. Experimental manipulations and question wordings

1-1. Experimental stimulus (manipulated scripts)

Descriptions. The respondents first read either "*Zainichi* Korean" or "those with disabilities" version of the introductory script. One fifth of the respondents (1), randomly chosen, did not read any additional message and moved directly to the hate speech regulation question (Section 1-2 below); (2) another one fifth did not read any additional message, but read the script that described hate speech with abusive words (shown in brackets in "*Experimentally manipulated scripts*" below); (3) another read the bandwagon-poll message; (4) another read the no bandwagon-poll message; and (5) the other group read the anti-discrimination message. This is a two-by-five factorial design, which results in 10 experimental groups. See Table A1 for this detail.

Table A1. Ten experimental groups

Group ID	Target minority group	Additional message
1	*Zainichi* Korean	(No message)
2	*Zainichi* Korean	(No message, description of hate speech with abusive words)
3	*Zainichi* Korean	Bandwagon-poll message
4	*Zainichi* Korean	No bandwagon-poll message
5	*Zainichi* Korean	Anti-discrimination message
6	Those with disabilities	(No additional message)
7	Those with disabilities	(No message, description of hate speech with abusive words)
8	Those with disabilities	Bandwagon-poll message
9	Those with disabilities	No bandwagon-poll message
10	Those with disabilities	Anti-discrimination message

Experimentally manipulated scripts

1) "*Zainichi* Korean" version.

In present day Japan, you can find insults and incitements to violence [such as "get lost", "go to hell", and "kill them"]* against *Zainichi* Koreans in some areas or on

the internet. Here, *Zainichi* Koreans means Korean nationals who reside in Japan (including special long-term residents who have lived in Japan before the Second World War, or their descendants).

* Appeared only for Group ID=2

2) “Those with disabilities” version.

In present day Japan, you can find insults and incitements to violence [such as "get lost", "go to hell", and "kill them"]* against those with disabilities in some areas or on the internet. Here, those with disabilities means people with some physical, intellectual, or mental disabilities.

* Appeared only for Group ID=7

Additional message

1) Bandwagon-poll message.

By the way, last year we conducted a public opinion survey of voters in all prefectures in the country (some 3,000 people) and asked whether the government should regulate such activities. We obtained, among those who expressed a “yes/no” opinion to this question, the following result:

In favor of regulation: 78%

Against regulation: 22%

This result shows that there are quite a few people who think that it should be regulated. Keeping that in mind, please answer the following question.

2) No bandwagon-poll message

By the way, last year we conducted a public opinion survey of voters in all prefectures in the country (some 3,000 people) and asked whether the government should regulate such activities. We obtained, among those who answered this question, the following result:

Support for regulation: 54%

This result shows that regulation is supported by a little over half of the people, and not necessarily by a great majority. Keeping that in mind, please answer the following question.

3) Anti-discrimination message.

By the way, some people point out that such words and behaviors can create a sense of discrimination against the targeted groups. In order to foster a good society, “acknowledging differences among people and respecting each other” is considered desirable, even if it may not represent majority opinion in Japan. Keeping that in mind, please answer the following question.

1-2. Hate speech regulations question (dependent variable).

We ask you how the government should respond to this issue. Do you think that the government should impose restrictions on insults and incitements to violence [such as "get lost", "go to hell", and "kill them"]* against [*Zainichi* Koreans / those with disabilities]? Or do you think that the government should not impose any restrictions?
* Appeared only for Group ID=2 and 7

[Response options]
0. Should not impose any restrictions; 1.; 2.; 3. Cannot say; 4.; 5.; 6. Should impose thorough restrictions; 7. I don’t want to answer (treated as "missing")

1-3. Index of Conformity towards a Majority Opinion[1]

Next, what do you think about each of the following statements about people, politics, and elections? For each statement, please choose one answer from “Agree” to “Disagree.”

Q1) It is useless to vote when a party or a candidate you support has no chance of winning.*
Q2) It is better to vote for a candidate with a good chance to win than to vote for a candidate without any chance.*
Q3) It is better to vote for a large party that is likely to gain many seats than to vote for a small party that is unlikely to secure many seats.*

[1] The questions with an asterisk (*) were used to construct the election bandwagon index, whereas those with a dagger (†) were used to construct the opinion bandwagon index. The fifth question was not used in the analysis, because it did not load significantly on either factor. For this detail, See Section 3. Confirmatory factor analyses of the moderator variables in this Appendix.

Q4) It is better to follow the majority opinion, when it comes to political issues.†
Q5) When two opposing politicians are on a debate, it makes me want to back up the losing side.
Q6) When I hear that the majority of people support a view, it makes me think that the view is correct.†

[Response options]
1. Agree; 2. Somewhat agree; 3. Neither; 4. Somewhat disagree; 5. Disagree;
9. I don’t want to answer (treated as "missing")

1-4. Impression management score from the Balanced Inventory of Desirable Responding, BIDR (Paulhus 1991, Japanese version, BIDR-J Tani 2008).

Next, we will ask you about yourself. How true are the following statements about you? For each item, please choose one answer from “Not true at all” to “Very true.”

Q1) I have done things that I don't tell other people about.
Q2) I sometimes tell lies if I have to.
Q3) There have been occasions when I have taken advantage of someone.
Q4) I never swear.
Q5) I have some pretty awful habits.
Q6) I have taken sick-leave from work or school even though I wasn't really sick.
Q7) I have never dropped litter on the street.
Q8) I sometimes try to get even rather than forgive and forget.
Q9) I always obey laws, even if I'm unlikely to get caught.
Q10) I have received too much change from a salesperson without telling him or her.
Q11) I have never damaged a library book or store merchandise without reporting it.
Q12) I worry a lot about how other people think about my behavior and speech.

[Response options]
1. Not true at all; 2. Not true; 3. Somewhat not true; 4. Neither;
5. Somewhat true; 6. True; 7. Very true; 9. I don’t want to answer (treated as "missing")

2. Descriptive statistics and distribution of the variables in the model

2-1. Descriptive statistics

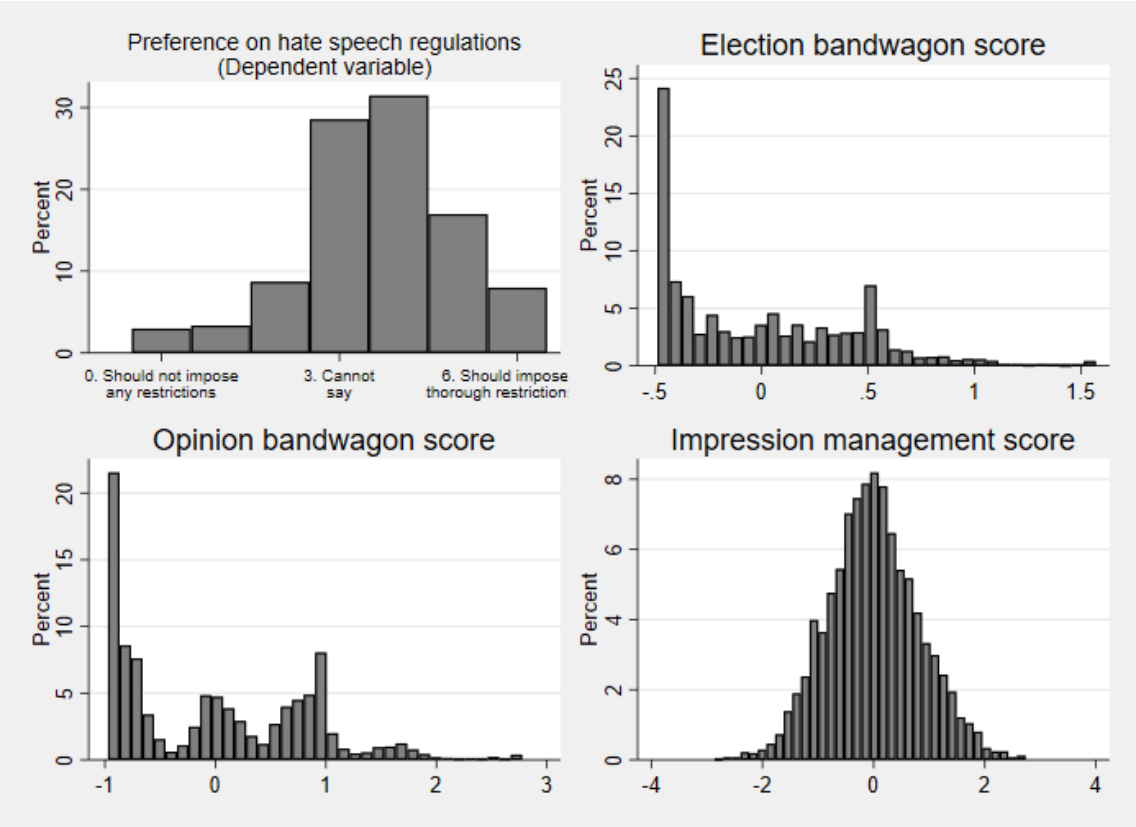
Table A2. Descriptive statistics of the variables in the model

Variable	Mean	SD	Range
Dependent variable:			
Hate speech regulations	3.65	1.33	0 to 6
Moderator variables:			
Election-bandwagon score	-0.005	0.45	-0.49 to 1.57
Opinion-bandwagon score	-0.008	0.85	-0.97 to 2.78
Impression management score	-0.004	0.83	-2.86 to 2.73
Experimental treatment variables:			
Minorities in the script: those with disabilities*	0.51	0.50	0 or 1
Message: No message, but describe hate speech as abusive words†	0.20	0.40	0 or 1
Message: Bandwagon-poll message†	0.20	0.40	0 or 1
Message: No bandwagon-poll message†	0.20	0.40	0 or 1
Message: Anti-discrimination message†	0.20	0.40	0 or 1
N= 4,113			

Notes: * Reference group for the minority groups in the script is *Zainichi* Koreans.
† Reference group for the message is "no message" and "without hate speech abusive words description".

2-2. Distributions of the variables

Figure A1. Distribution of the dependent and moderator variables in the model



3. Confirmatory factor analyses of the three moderators

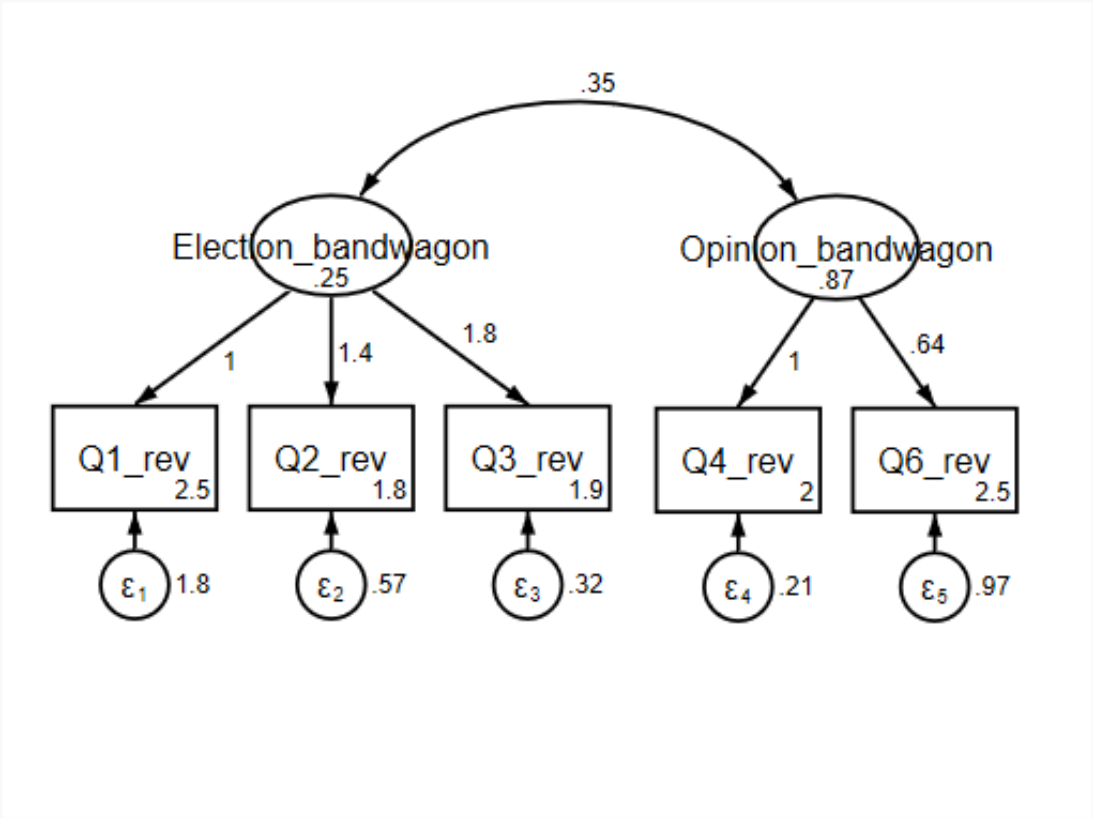
3-1. Election and opinion bandwagon scores

Table A3. Confirmatory factor analysis of the election and opinion bandwagon scores

	Unstandardized			Standardized		
Item	Loading	Intercept	Error var.	Loading	Intercept	Error var.
Election bandwagon score						
Q1	1.000	2.512	1.800	0.347	1.756	0.879
Q2	1.399	1.756	0.570	0.678	1.711	0.541
Q3	1.827	1.928	0.323	0.848	1.799	0.281
Opinion bandwagon score						
Q4	1.000	2.005	0.212	0.896	1.929	0.196
Q6	0.645	2.520	0.970	0.521	2.185	0.729
(Election bandwagon score)			0.247			1.000
(Opinion bandwagon score)			0.868			1.000
Covariance between two scores			0.350			0.755
Likelihood ratio χ^2			62.82			
RMSEA			0.059			
CFI			0.988			
N			4,172			

Notes. Both the election bandwagon score and opinion bandwagon score were constructed from the factor scores of the two-factor model above, with the higher (positive) values representing preference for winning candidates or parties (election) or views that the majority supported (opinion). Cronbach's alpha of 3 items for the election bandwagon score is 0.62. For this distribution, see Figure A1 in Section 2-2 in this Appendix.

Figure A2. Illustration of the confirmatory factor analysis of the election and opinion bandwagon scores



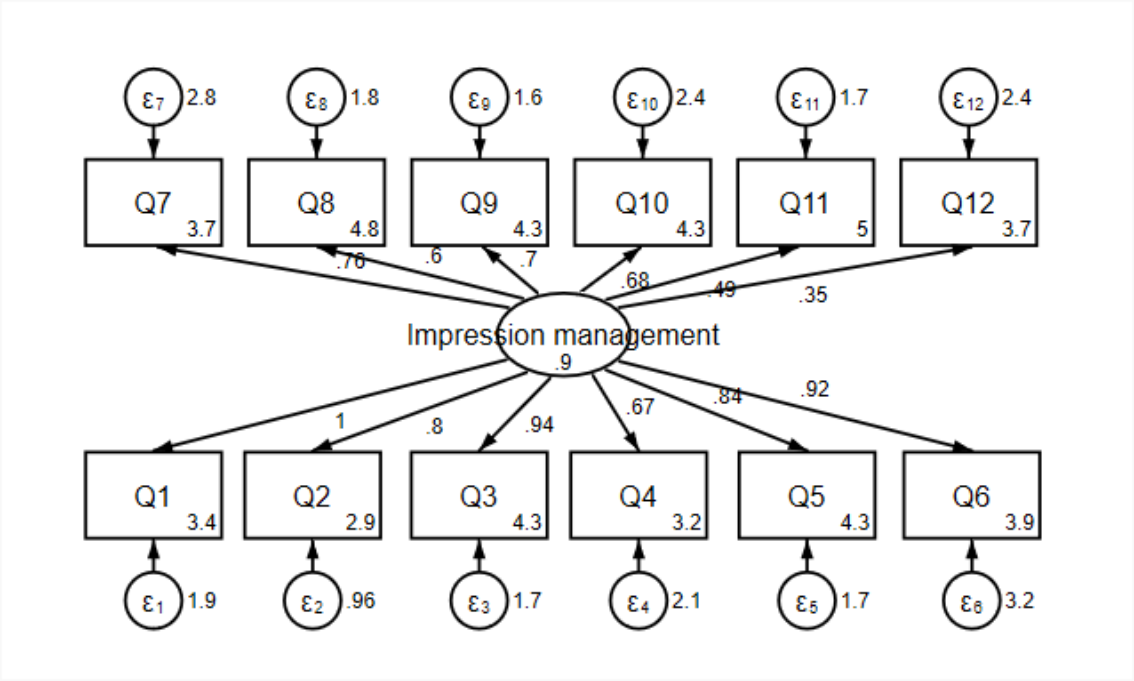
3-2. Impression management score

Table A4. Confirmatory factor analysis of the Impression management score

	Unstandardized			Standardized		
Item	Loading	Intercept	Error var.	Loading	Intercept	Error var.
Q1	1.000	3.413	1.926	0.564	2.031	0.682
Q2	0.798	2.923	0.96	0.61	2.363	0.627
Q3	0.935	4.299	1.67	0.565	2.744	0.681
Q4	0.670	3.183	2.132	0.399	1.999	0.841
Q5	0.836	4.303	1.739	0.515	2.798	0.735
Q6	0.923	3.862	3.200	0.439	1.940	0.808
Q7	0.759	3.739	2.823	0.393	2.046	0.845
Q8	0.600	4.754	1.778	0.392	3.281	0.846
Q9	0.699	4.269	1.617	0.462	2.978	0.787
Q10	0.685	4.274	2.402	0.386	2.544	0.851
Q11	0.492	5.047	1.693	0.337	3.651	0.886
Q12	0.353	3.747	2.407	0.210	2.361	0.956
(Impression management score)			0.896			1.000
Likelihood ratio χ^2			1450.289			
RMSEA			0.08			
CFI			0.802			
N			4,143			

Notes. Cronbach's alpha of these 12 items are 0.74. The impression management score was constructed from the factor scores produced in the analysis above, with the higher (positive) values representing higher likelihood of socially desirable responding based on impression management. For this distribution, see Figure A1 in Section 2-2 in this Appendix.

Figure A3. Illustration of the confirmatory factor analysis of impression management score



4. Models for the calculating moderated effects

As we described in Section 4-2 in the manuscript, the base regression model from which we calculated the ATEs in Figure 1 can be written as

$$Y = \beta_0 + \sum_{i=1}^{8} \beta_i X_i$$

where Y is the dependent variable, β_0 is the constant, X are eight independent variables, distinguished by the subscript i. This model can be collapsed with each term,

$$\begin{aligned} Y = {} & \beta_0 \\ & + \beta_1 X_{disability} + \beta_2 X_{abuse} + \beta_3 X_{bandwagon} + \beta_4 X_{no-bandwagon} + \beta_5 X_{anti-dis.} \\ & + \beta_6 X_{election} + \beta_7 X_{opinion} + \beta_8 X_{impression} \end{aligned}$$

Here, the numeric subscripts for the variables (X) are replaced by descriptions: those in the second line of the equation represent experimental treatments/groups, and those in the third line represent moderating variables.

We calculated the predicted values and marginal effects in Figures 2, 3, 4 and 5 from different regression models, respectively. Each model contained only one interaction term between one of the experimental treatment/group variables and one of the relevant moderator variables concerning the proposition we examined. Generally, the model can be written as:

$$Y = \beta_0 + \sum_{i=1}^{8} \beta_i X_i + \beta_{j,k} X_j X_k$$

where $i = \{1 \leq i \leq 8 \mid i \in N\}; j = \{3, 5\};\ k = \{6, 7, 8\}$.

More specifically, the regression model for Figure 2 (moderated effects of bandwagon-poll message by the election bandwagon score) is described as

$$Y = \beta_0 + \sum_{i=1}^{8} \beta_i X_i + \beta_{3,6} X_{bandwagon} X_{election}$$

Similarly, the model for Figure 3 is

$$Y = \beta_0 + \sum_{i=1}^{8} \beta_i X_i + \beta_{3,7} X_{bandwagon} X_{opinion}$$

The model for Figure 4 is

$$Y = \beta_0 + \sum_{i=1}^{8} \beta_i X_i + \beta_{3,8} X_{bandwagon} X_{impression}$$

And finally the model for Figure 5 is

$$Y = \beta_0 + \sum_{i=1}^{8} \beta_i X_i + \beta_{5,8} X_{anti-dis.} X_{impression}$$

5. Regression analysis results

Table A5. Results of regression analyses for Figures 1 to 5

Independent variables	Model 1	Model 2	Model 3	Model 4	Model 5
Experimental treatment variables:					
Those with disabilities	0.402***	0.403***	0.403***	0.402***	0.402***
	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)
Hate speech as abusive words	0.086	0.084	0.085	0.086	0.086
	(0.064)	(0.064)	(0.064)	(0.064)	(0.064)
Bandwagon-poll message	0.059	0.057	0.059	0.064	0.059
	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)
No bandwagon-poll message	-0.09	-0.092	-0.091	-0.09	-0.09
	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)
Anti-discrimination message	-0.142*	-0.143*	-0.143*	-0.143*	-0.143*
	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)
Moderator variables:					
Opinion bandwagon score	0.004	0.005	0.022	0.001	0.005
	(0.046)	(0.046)	(0.048)	(0.046)	(0.046)
Election bandwagon score	-0.154	-0.11	-0.155	-0.151	-0.154
	(0.088)	(0.090)	(0.088)	(0.088)	(0.088)
Impression management score	0.105***	0.106***	0.106***	0.082**	0.095***
	(0.025)	(0.025)	(0.025)	(0.028)	(0.027)
Interaction terms between experimental treatment and moderator variable:					
Bandwagon-poll message		-0.240*			
* Election bandwagon		(0.117)			
Bandwagon-poll message			-0.086		
* Opinion bandwagon			(0.061)		
Bandwagon-poll message				0.119	
* Impression management				(0.062)	
Anti-discrimination message					0.057
* Impression management					(0.064)
Constant	3.463***	3.465***	3.464***	3.464***	3.464***
	(0.050)	(0.050)	(0.050)	(0.050)	(0.050)
R-squared	0.035	0.036	0.035	0.035	0.035
N	4113	4113	4113	4113	4113

Notes: Standard errors are shown in parentheses. Asterisks represent the level of *p*-value, * $p<0.05$, ** $p<0.01$, *** $p<0.001$. Each model number corresponds to the Figure in the manuscript: Model 1 is used to calculate the ATEs in Figure 1, Model 2 is used to calculate the marginal effects in Figure 2, and so on.