Analyses of Social Issues and Public Policy, Vol. 00, No. 0, 2020, pp. 1–28

When Descriptive Norms Backfire: Attitudes Induce Undesirable Consequences during Disaster Preparation

Taku Ozaki *

Doshisha University

Kazuya Nakayachi

Doshisha University

*This study aimed to examine the adverse effects of descriptive norms in the context of disaster preparation, although the cumulative evidence has demonstrated that descriptive norms desirably promote social behavior. Descriptive norms inevitably inform that some people do not prepare; therefore, the secondary meaning can be utilized to confirm the receiver's already possessed attitudes, and it results in a backfire. We conducted two preregistered experiments examining preparation behaviors for natural disasters while manipulating descriptive norms. In Study 1 (*N *= 262), the only promotive effect of descriptive norms was obtained among U.S. participants. In Study 2 (*N *= 329), when replicated with a sample of Japanese participants, participants' attitudes moderated the effects of descriptive norms. The patterns of moderation were consistent with our prediction and suggested that descriptive norms suppressed the desirable behavior when participants held negative attitudes. The pitfalls of descriptive normative approaches are also practically important.*

*Correspondence concerning this article should be addressed to Taku Ozaki, Graduate School of Psychology, Doshisha University, 1–3, Miyakodani, Tatara, Kyotanabe-shi, Kyoto, Japan [e-mail: ekp1004@mail2.doshisha.ac.jp].

Acknowledgments: This study was preregistered at the Open Science Framework (OSF) prior to data collection (https://osf.io/9cg4x). All materials, the preregistration manuscript, obtained data, and analyses codes are also available at the OSF storage (https://osf.io/yzdqh/). Doshisha University's institutional review board reviewed and approved the study (No. 16071 for Study 1 and No. 201903 for Study 2). The authors thank Toshiaki Kimura, Mariko Shirai, Yukihide Shibata, and Ryosuke Yokoi for their helpful comments on the draft. This work was supported by JSPS KAKENHI (Grant Number JP15J02505). We would like to thank Editage (www.editage.com) for English language editing.

DOI: 10.1111/asap.12195

The Effectiveness of Social Norm Approaches

Social norm approaches, aiming to nudge people toward desirable behavior, are widely applied in various research areas (as a meta-analysis exploring the effectiveness of social norm approaches in field experiments, Bergquist, Nilsson, & Schultz, 2019; as an opinion letter, Nyborg et al., 2016). One of the most comprehensive frameworks of normative influence is the focus theory of normative conduct (Cialdini, Reno, & Kallgren, 1990), which distinguishes social norms as two main parts: injunctive norms (norms concerning the expectancies of other people) and descriptive norms (norms concerning the actual behavior of other people; i.e., the percentage of others who perform the designated behavior).

Cumulative literature shows that descriptive norms promote desirable behavior in various domains. The literature indicates that informing people that many others commit desirable behaviors makes them behave in line with others. The effectiveness of descriptive norms has been observed in disorderly public settings (Keizer, Lindenberg, & Steg, 2008) and concerning pro-environmental behaviors (Goldstein, Cialdini, & Griskevicius, 2008; Melnyk, Herpen, Fischer, & van Trijp, 2011; Schultz, Nolan, Cialdini, Goldstein, & Griskevicius, 2007; for a meta-analysis, Bergquist, Nilsson, & Schultz, 2019; for a review, Farrow, Grolleau, & Ibanez, 2017). The effectiveness was also discovered in risk domains, such as the promotion of a healthy diet (Plows et al., 2017; Stok, De Ridder, De Vet, & De Wit, 2014), prevention of risky alcohol consumption (Lewis, Neighbors, Oster-Aaland, Kirkeby, & Larimer, 2007; Neighbors et al., 2011), and safe sexual behaviors (for a meta-analysis, Sheeran, Abraham, & Orbell, 1999). Overall, descriptive normative approaches could influence desirable behavior (for reviews, Cialdini & Goldstein, 2004; Dempsey, McAlaney, & Bewick, 2018; Lapinski & Rimal, 2005; for a meta-analysis, Rivis & Sheeran, 2003). Even when personal behavior change is targeted, it is also essential to consider social influences.

Injunctive norms have also been shown to play an essential role in social norm approaches. Not only injunctive norms but also the combination of injunctive and descriptive norms was found to successfully promote desirable social behaviors (e.g., Alcott, 2011; Costa & Kahn, 2013; for a review, Farrow, Grolleau, & Ibanez, 2017).

Descriptive Norms Can Backfire

We investigated not only the promotive effect of descriptive norms but also the pitfalls of them, such as the boomerang effect (Richter, Thøgersen, & Klöckner, 2018; Schultz et al., 2007). Schultz et al. (2007) reported that a descriptive normative message triggered the boomerang effect, but that injunctive norms resolved the negative outcome induced by descriptive norms. Although this research demonstrated the reconstructive power of the injunctive norms, the mechanisms of the

backfire's occurrence were not directory examined. In the current study, we focused only on descriptive norms, not injunctive norms, as descriptive norms were assumed to trigger the boomerang effect. The sole examination of descriptive norms in the current study would allow us to understand the mechanisms of the backfire effect. Although the possible pitfalls of descriptive norms have been stated, the mechanisms behind their backfire are less understood because the results came from field experiments (Farrow, Grolleau, & Ibanez, 2017). There are also contradictory theories. Werch and Owen (2002) reported in their literature review that there is an unignorable possibility of a boomerang effect on the health promotion domain, whereas Prince, Reid, Carey, and Neighbors (2014) denied the possibility based on a series of studies. We thus developed controlled, high-powered experiments to explore the backfire effect and examined the descriptive norms process.

The Double Meanings of Descriptive Norms and Attitude as a Moderator

We focused on the double meanings of descriptive normative information and attitude as a potential moderator to investigate the adverse effects of descriptive norms. Descriptive normative information includes two discrepant meanings: most people behave in a desirable manner; however, a minority of people do not. Usually, the majority is focused on; however, the existence of the minority has been highlighted in some cases.

Attitudes also play an essential role in reasoning descriptive normative information. It is well known that individuals are motivated for attitude consistency: for example, cognitive dissonance theory (Festinger, 1957) and balance theory (Heider, 1958). To maintain consistency, people selectively accept or refuse the presented information. People tend to interpret the information along with their held attitudes (Brock & Balloun, 1967; Fraser-Mackenzie & Dror, 2009; Lord, Ross, & Lepper, 1979; for a review, Frey, 1986). Kunda (1990) summarized motivated reasoning and described bias, in which people utilize information to reach their desired goals. This is also known as confirmation bias—the tendency to ignore information that does not suit one's preferred beliefs and utilize the information that matches one's expectations to confirm their personal beliefs (Snyder & Cantor, 1979; Wason, 1960; for reviews, Baron, 2008; Nickerson, 1998).

People who already possess negative attitudes toward the designated behaviors may focus on the minority who did not act in the designated way. Moreover, they might try to utilize the minority trend, by implying that because some people did not perform the desirable behavior, their previously possessed attitudes are confirmed. When people plan to persuade others using normative information, confirmation bias can be an obstacle (Nickerson, 1998). We considered that individuals' way of reasoning may induce the adverse effects of descriptive norms. By

contrast, those who held positive attitudes would be straightforwardly influenced by the majority, thus fostering the promotive effects of descriptive norms.

We assumed that attitudes may play a moderator role. Melnyk, Herpen, Fischer, and van Trijp, (2011) noted that descriptive normative messages promoted pro-environmental behaviors; however, the promotion only occurred among the participants who believed the normative messages were real. Our reinterpretation of those results is that the believers held relatively positive pre-attitudes compared to the nonbelievers. This resulted in normative information being received without conflict, that is, a promotive effect. Nonbelievers who held negative attitudes, on the other hand, did not utilize the information to confirm their attitudes; thus, there was no influence on the norms. Although Melnyk and colleagues did not directly assess previously possessed attitudes, we believe their results suggest that descriptive norms work differently in marginal conditions.

An Alternative Explanation of the Backfire from Social Identity Theory

The backfire effect of descriptive norms can be predicted using social identity theory (Tajfel, 1982; Tajfel & Turner, 1979; Turner, 1982), which considers not only the unidirectional effect of social norms but also the normative influence as the group processes (as a counterargument to the informational dependence approach: Hogg & Abrams, 1988; Hogg & Turner, 1987; Turner, 1987). Thus, the receivers of social norms would not just be passive but have the motivation to maintain their social identity. This perspective views conformity as an identification contingent. Empirically, the moderating effects of social identity on the effects of social norms have been reported (Neighbors et al., 2010; Rinker & Neighbors, 2014).

Social identity theory also distinguishes the influence of group norms derived from the in-group and out-group because social identity is attributed to the source of the norms. Known as group polarization (Myers & Lamm, 1976; Wetherell, 1987), a group member's opinion becomes polarized when confronting the out-group's opinion as a means of distinguishing the difference between the in-group norm and out-group one (Mackie, 1986; Wilder, 1990). Further, the social identity approach has already applied to analyze the real-world problem, such as the one that we aimed to examine (Reicher, 1987).

In our current studies, positive attitudes holders were those who tended to identify themselves with the majority of people who had already prepare for the disaster because of their similarity in beliefs. Such identification would encourage those with positive attitudes to follow the majority. However, participants with negative attitudes participants were those who could not view themselves as belonging to the majority. They instead identified themselves as being in the minority group who did not prepare. Consequently, they were willing to follow the minority and tended not to prepare: the backfire effect.

Social Norm Approaches on Disaster Risk Reduction

Practically and theoretically, disaster preparation can be investigated using social norm approaches. In recent years, the importance of personal preparation for natural hazards has been widely recognized (for a review, Levac, Toal-Sullivan, & O'Sullivan, 2012). However, there is a limitation to promote disaster defense without social norms because the traditional approaches did not fully consider the social norm effects (Solberg, Rossetto, & Joffe, 2010).

Lack of preparation toward disasters is a key practical issue. In 2015, a United Nations conference—the World Conference on Disaster Risk Reduction—adopted the Sendai Framework, which declared that, not only should the public sectors be prepared, but individual citizens also share a responsibility to survive future natural disasters. The U.S. president also published a proclamation that called for citizens' personal preparation for catastrophe, and he defined September 2016 as "National Preparedness Month." Although Japan is one of the most disaster-prone countries, individual mitigation activities are still insufficient (Onuma, Shin, & Managi, 2017).

Theoretical concerns concerning disaster risk reduction are also associated with social normative approaches. Most commonly, risk perception has been the primary exploratory variable for disaster preparation behavior in conventional psychological models (i.e., the Protective Action Decision Model, Lindell & Perry, 1992, 2012; the Protection Motivation Theory, Rogers, 1975, 1983; and the Person-Relative-to-Event Model, Duval & Mulilis, 1999; Mulilis & Duval, 1995). Their central assumption was that high-risk perception would promote preparation. However, recent review literature showed that the relationship between risk perception and behavior was weak or even null (Bubeck, Botzen, & Aerts, 2012; Ratner & Riis, 2014; Shreve et al., 2014; Solberg, Rossetto, & Joffe, 2010; Wachinger, Renn, Begg, & Kuhlicke, 2013). Furthermore, a meta-analysis showed that the correlation between risk appraisal and behavior was seemingly small (Rivis & Sheeran, 2003). Wachinger, Renn, Begg, and Kuhlicke, (2013) named this expected but not strong relationship as the "risk perception paradox." Their consideration implies the necessity of exploring the factors that promote preparation.

Descriptive norms are one of the potential promotive elements of disaster preparedness. Solberg, Rossetto, and Joffe, (2010) noted that social norms might promote behavior toward natural hazards, even though there have been insufficient empirical investigations. A variety of researchers suggested the potential effects of social influences: for example, subjective norms (Ajzen, 1991), normative beliefs (Paton, 2003), environmental elements (Rogers, 1983), and peer influence (Wachinger, Renn, Begg, & Kuhlicke, 2013. Even private disaster preparation can be influenced by social pressures. Some studies have already shown that descriptive norms play an essential role in disaster preparation (Kunreuther et al.,

1978; Lo, 2013; Morsink & Geurts, 2012; Paek, Hilyard, Freimuth, Barge, & Mindlin, 2010; Vinnell, Milfont, & McClure, 2019).

Notably, the backfire effects of descriptive norms have not been investigated in the disaster risk domain. We thus conducted two experiments to clarify the effectiveness and possible pitfalls of the normative approach in this context.

Study Overview

We hypothesized that the attitudes that were previously possessed (pre-attitudes) would moderate the effects of descriptive norms. The pre-attitudes were measured to identify the participants' default standpoint regarding preparing for future disasters (i.e., the following normative information would be interpreted along with the pre-attitudes). The valence of the pre-attitudes could range from negative to positive. Negative pre-attitudes meant that the participants had unfavorable position toward preparation on the affective, behavioral, and cognitive dimensions, whereas participants with positive pre-attitudes seemed to be in favor of the preparation.

We showed participants descriptive normative information that described that most other participants had already prepared for the disasters and the minority did not. The moderation effect can thus be described as follows: participants' preparation behavior would be promoted by descriptive normative messages when they held positive pre-attitudes. This effect results from their focus on the majority who have already prepared for the disasters. By contrast, among participants who held negative pre-attitudes, the promotive effects of the normative messages would be too small to be detected because the unintended focus on the minority would attenuate the positive influences of the norms. Consequently, those holding negative pre-attitudes might focus on the minority to confirm their attitudes.

Farrow, Grolleau, and Ibanez, (2017) claimed that the information processing of social norms is still unknown, even though the effectiveness of social norm approaches is recognized. Our hypothesis testing will theoretically contribute to clarifying the mechanisms behind descriptive norms and their implications. Additionally, identification of the moderator of social norms will be useful when applying social norm approaches in disaster preparation policy.

The present studies were preregistered using the template presented by van't Veer and Giner-Sorolla (2016). All materials and the preregistration manuscript were uploaded to the Open Science Framework (OSF) prior to data collection. All studies were conducted single blind. During the first study, however, we ran into problems with participants from a U.S. crowdsourcing site. The considerable number of responses suggested repeated participation by the same participants. Thus, we decided to terminate data collection in Study 1 in the middle of the collection process and conducted Study 2, in which we recruited participants

from another trustworthy online survey panel in Japan. The study protocols were approved by the Institutional Review Board.

Study 1

Method

Differences between the present study and the preregistration. This study corresponds to "Study 3" in the preregistration. Although the preregistration described three experiments about the boomerang effect of descriptive norms, we only conducted Study 3 in the preregistration because of the first author's lack of research resources. Despite its incompletion, we decided to conduct Study 3 as Study 1 in this article because it was the central part of our hypothesis testing.

In Study 1, we terminated data collection in the middle because we suspected that some of the participants participated in the experiment more than once. We attempted to restrict multiple participation using tools provided by CrowdFlower and Qualtrics before data collection; however, given the responses, we deemed these efforts to be insufficient. To handle this problem, we excluded the suspect participants as an additional criterion to the preregistered exclusion criteria. The other part of the study was identical to the preregistration.

Sample size calculation. In this subsection, we summarize the description of sample size calculations reported in the preregistration. We set two-tailed significance level (α) at .05 and intended to obtain power ($1 - \beta$) at .80 to detect the estimated effect sizes. To assess effect sizes, we referred to Melnyk, Herpen, Fischer, and van Trijp, (2011), which investigated the effect of descriptive normative messages on the number of thoughts, attitudes, and behavioral intention toward environmentally friendly processed farm products. That study included an online experiment as well as our recent surveys. They reported a medium-sized effect of descriptive norms on participants' behavior intention (odds ratio = 2.08). We assumed a smaller effect on actual behavior than on intention, using the odds ratio of 1.44 (equivalent to Cohen's $d = .20$), which we posited was the minimum effect to detect real-world application.

To detect the effect size in the logistic regression model, the requisite sample size was calculated. The base probability of reading the leaflet was set at 68.2%, which was adapted from a preliminary study. We estimated that the covariates would explain 25% of the variance of the outcome. We considered that 382 participants were necessary. An instructional manipulation check was presented to exclude satisficers (Oppenheimer, Meyvis, & Davidenko, 2009). We assumed 80% of participants would pass the instructional manipulation check (Klein et al., 2014). Furthermore, an attention check was presented to ask participants to confirm their memory of the provided descriptive normative information. We estimated

that 90% of participants would successfully pass the attention check. The pass rate estimation was set based on the difficulty level of the attention check and our previous experiences. Finally, we planned to recruit 531 participants.

Participants. Participants who could understand English and lived in the United States were recruited from CrowdFlower. Participants were paid $1.50 for participation. A total of 428 responses were collected. Retrieved latitude and longitude location data were used to exclude suspected multiple participation ($n = 127$).[1] Identifiable information was removed and replaced with unique flags for replication when the obtained data were uploaded to OSF. We referred to a discussion on an Internet forum regarding exclusion (https://groups.google.com/forum/#!topic/otree/H8_xIeangAk). We also referred to a discussion about recruiting participants on a crowdsourcing site by Paolacci, Chandler, and Ipeirotis (2010).

According to the preregistration, we used latitude and longitude location data to exclude participants who participated in the experiment outside of the United States ($n = 2$). Participants who did not complete the questionnaire ($n = 41$), responded that they participated in a similar survey ($n = 30$), participated for less than 1 minute ($n = 19$), failed to pass the instructional manipulation check ($n = 33$), and failed to pass the attention check ($n = 17$) were excluded. The excluded participants above included duplications.

As a result, 262 participants were finally included in the analyses (179 women, 81 men, and two other). Recalculation indicated statistical power $(1 - \beta) = .64$ based on the actual sample size. In other words, the actual sample size yielded power of .80 to detect effects as small as odds ratio = 1.56. Based on the findings that White women are more sensitive to risks than men (Flynn, Slovic, & Mertz, 1994), we considered sex as a control variable. Sex was coded as women = 1 and not women (men, other, and "prefer not to answer") = 0. Participants' mean age was 37.72 years ($SD = 13.29$).

Procedure. The experiment was conducted on the Internet, and the questionnaire was built using Qualtrics. At the beginning of the questionnaire, we presented a description of a natural disaster in the United States. This description provided details about recent catastrophic disasters in the United States and noted the importance of preparation. Following the description, risk perception of natural disasters was measured by four items on a 7-point Likert scale ranging from 1 (*extremely low*) to 7 (*extremely high*). The items concerned (1) effect on physical risk, (2) probability of physical risk, (3) effect on financial risk, and (4) probability

[1] We observed 16 unique combinations of the exact same latitude and longitude (about 14 digits for each latitude and longitude). In one case, 24 responses were recorded from the same location. We believed that a bot (i.e., a computer program) was used to complete the responses instead of a person. In addition, the responses that did not contain location data were also excluded.

of financial risk. According to the definition of risk (National Research Council, 1989), risk perception was calculated as (1) × (2) + (3) × (4).

After measuring risk perception, an instructional manipulation check item was presented. Participants who failed to follow the instructions could proceed to the next questions; however, their answers were not included in the analyses.

In the next section, pre-attitudes were measured using 10 items on an 11-point Likert scale ranging from 1 (*disagree very strongly*) to 11 (*agree very strongly*). To measure pre-attitudes, participants were asked to rate their thoughts about, "In preparation for natural disasters, what do you think of 'knowing information about preparation.'" The items were as follows (items with an asterisk refer to reverse scored items): "I want to learn preparation routinely," "I think it is interesting to know information about preparation," "I think checking information about preparation spontaneously is important," "I think I need to look for opportunities to know information about preparation," "I am always aware of need for preparation," "It seems time-consuming to learn information about preparation*," "I do not think knowing information about preparation is essential to survive*," "I am negative about knowing information about preparation*," "I think knowing information about preparation is boring*," and "I prefer avoiding information about preparation*." The scale covered the three components of attitudes—cognitive, emotional, and behavioral (Rosenberg & Hovland, 1960). The items were developed based on laypersons' discourse of information-seeking preparation toward natural disasters.

On the next page, an explanation of a 10-page leaflet containing useful tips and information to help households be fully prepared for natural disasters was presented. The leaflet was produced by the Tokyo Metropolitan Government (http://www.metro.tokyo.jp/ENGLISH/GUIDE/BOSAI/index.htm).

After the explanation, an experimental manipulation was conducted between participants. Before participants decided to read it, half (randomly assigned, Qualtrics-generated random numbers) saw descriptive normative information, and the other half did not see normative information (controls). Descriptive normative messages stated that the total number of the past participants in this study was 107, and most other participants (68.2%) read the leaflet during this survey and the minority of them did not (31.8%). The proportion was obtained by a preliminary study using the same sample pool; we did not use deception. The proportion was somewhat moderate when adapting the descriptive norm manipulation (cf. 75% in Goldstein, Cialdini, & Griskevicius, 2008; 64% in Melnyk, Herpen, Fischer, & van Trijp, 2011; 91% or "at maximum degree" in Richter, Thøgersen, & Klöckner, 2018). The moderate level of the descriptive norms was beneficial compared to using strong descriptive norms because it meant that the designated behavior was not prevalent, suggesting that there was ample for social improvement using the social norm approaches. Along with both percentages, the reasons why other participants read or did not read the leaflet were also shown: for example, "seemingly

useful information in the leaflet" and "already well prepared for disasters." Descriptive normative messages explained that other participants made their choice independently. After the manipulation, the actual behavior to read or not read the leaflet was observed. Participants were explicitly informed that they would be paid regardless of their choice.

In the next section, the need for cognition was measured by 18 items. The need for cognition (Cacioppo & Petty, 1982; Cacioppo, Petty, & Feng Kao, 1984) was assumed as a covariate for analysis because the extent that participants examine a message can influence the depth of their reasoning about the behavior of others. Subsequently, we asked about the participants' sex and age. Finally, as an attention check, participants were asked whether the message stated that over 50% of past participants decided to read the leaflet, and then they were thanked. Participants who correctly answered the attention check were included in the analysis.

We did not conduct an experimental manipulation check because the manipulation of descriptive norms seemed obvious. Alternatively, we adopted the attention check to confirm that the participants were under the effects of descriptive norms.

Results and Discussion

Descriptive statistics. All analyses were conducted using R version 3.6.1. The experiment was conducted from January to February 2017. Table 1 indicates the number and proportion of participants who performed disaster preparation behavior in each condition and their combined pre-attitudes (stratified by median). Participants who held positive pre-attitudes and were in the descriptive norms condition had the highest probability of reading the leaflet about disaster preparation. However, contrary to our hypothesis, participants who held negative pre-attitudes also had a high probability of performing the behavior when they were presented with the normative message. The median reading time of the leaflet was 63.65 seconds. Half of the participants who chose to read the leaflet read it for more than 1 minute.

Table 1. The Number and Proportion of Participants Who Read the Leaflet ($N = 262$)

			Behavior	
			No *n* (%)	Yes *n* (%)
Condition				
Control	Pre-attitudes	High	23 (36.5)	40 (63.5)
		Low	44 (59.5)	30 (40.5)
Descriptive norms	Pre-attitudes	High	9 (12.9)	61 (87.1)
		Low	18 (32.7)	37 (67.3)

Note: The pre-attitudes indicating high or low stratified by the median (8.40).

Confirmatory analysis: Model comparison between the main effects model and the interaction effect model[2]. To test the interaction effect between descriptive norms and pre-attitudes, we constructed two logistic regression models. Preparation behavior (dichotomous variable: *read the leaflet* = 1, *not read* = 0) was introduced in the logistic regression models as an objective variable. Descriptive norms (dichotomous variable: *descriptive norm condition* = 1, *control condition* = 0) and pre-attitudes (continuous variable: centered to avoid multicollinearity) were considered as explanatory variables. Pre-attitudes and need for cognition were computed by averaging each scale, and they showed excellent internal consistency (Cronbach's $\alpha = .90$ and .93, respectively).

Sex, need for cognition, and risk perception were entered into the models as covariates. To test the interaction effects, we conducted a likelihood-ratio test between two models. The main effects model involved the explanatory variables and the covariates, which meant there were only main effects. The model was considered as the null hypothesis model. The interaction effect model involved an interaction term between descriptive norms and the pre-attitudes with all covariates and the main effects. The model was considered as an alternative hypothesis model. It is known that the difference between the two models' deviance follows the chi-square distribution. The likelihood-ratio test did not reject the null hypothesis model ($\chi^2(1) = 0.63$, $p = .43$). The interaction model was not adopted. Akaike Information Criterion (AIC) also indicated that the main effects model was valid. The Variation Inflation Factors (VIFs) of the main effects model were less than 1.34, and it seemed that there were no multicollinearity problems. The results of the logistic regression analyses are shown in Table 2. The table was created using R package "texreg" (Leifeld, 2013).

In the main effect model, significant promotive effects of descriptive norms and pre-attitudes were observed. The estimated log odds ratios were converted to odds ratios, risk ratios, and Cohen's *d*s for interpretation (Table 3). When the effect size of descriptive norms corresponded to a Cohen's *d*, it is considered an intermediate effect size. Risk ratio was calculated using the transformation method developed to control its inflation by Zhang and Yu (1998) and Grant (2014). The relative risk ratio represented that showing a descriptive normative message multiplied the chance of taking preparation behavior by 1.33 times as compared to not showing the message. These effect sizes are meaningful for actual application concerning promoting preparedness toward natural disasters.

Exploratory analyses: Examining reading time of normative messages. The results that descriptive norms promoted the designated behavior regardless of their pre-attitudes were contrary to our prediction. To examine the influence of

[2] The preregistered hypothesis test is reported in this confirmatory analysis section. Further analyses not described in the preregistration are reported in the exploratory analyses section.

Table 2. The Estimates and the Standard Errors of the Logistic Regression Models ($N = 262$)

	Main effects model	Interaction effect model
Intercept	0.31	0.37
	(0.83)	(0.83)
Risk perception	0.00	0.00
	(0.01)	(0.01)
Need for cognition	−0.09	−0.11
	(0.21)	(0.21)
Sex	−0.00	−0.01
	(0.30)	(0.30)
Pre-attitudes	0.32	0.27
	(0.11)	(0.12)
Descriptive norms	1.19	1.22
	(0.28)	(0.29)
Pre-attitudes × Descriptive norms		0.15
		(0.19)
AIC	319.60	320.97
BIC	341.01	345.94
Log likelihood	−153.80	−153.48
Deviance	307.60	306.97

Note: The preparation behavior was regressed by the explanatory variables and covariates. Descriptive norms and the pre-attitudes were explanatory variables. Sex, need for cognition, and risk perception were considered covariates. The preparation behavior was coded as 0 = not read the leaflet, 1 = read the leaflet; descriptive norm was coded as 0 = control condition, 1 = descriptive norm condition; and sex was coded as 0 = nonfemale, 1 = female. Pre-attitude was centered. AIC refers to Akaike Information Criterion and BIC refers to Bayesian Information Criterion.
$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Table 3. Effect Sizes of the Main Effects Model

	Coefficient (log odds ratio)	95% confidence interval (CI)	Odds ratio	95% CI	Risk ratio	95% CI	Cohen's *d*	95% CI
Pre-attitudes	0.32	[0.02, 0.51]	1.37	[1.11, 1.70]	1.06	[1.02, 1.09]	.18	[0.06, 0.29]
Descriptive norms	1.19	[0.66, 1.79]	3.30	[1.89, 5.75]	1.33	[1.21, 1.42]	.66	[0.35, 0.96]

descriptive norms, an exploratory analysis was conducted. We analyzed the data concerning the duration on the page of the Qualtrics questionnaire, which informed descriptive norms.

One hundred twenty-five participants were shown descriptive normative messages during the survey. When we stratified participants by the median of pre-attitudes (8.50), we found the following descriptive statistics about the reading time of normative information. Among those who held negative and positive pre-attitudes, median reading time was 34.85 and 46.52 seconds, respectively.

The simple correlation between pre-attitudes and reading time was $r = .15$, 95% confidence interval (CI) [–0.03, 0.31], $p = .11$.

The duration time was regressed by an explanatory variable—pre-attitudes. Sex, need for cognition, and risk perception were entered into the model as covariates. We did not apply the experimental condition and the interaction term because this analysis was conducted on the extracted participants. The duration time did not follow a normal distribution; thus, we adopted generalized linear modeling. The objective variable—reading time—was a nonnegative continuous variable; therefore, we considered it would follow a Gamma distribution (Palmer, Horowitz, Torralba, & Wolfe, 2011), and the inverse function was used as the canonical link function. The generalized linear model showed a significant effect of pre-attitudes, indicating that pre-attitudes correlated with reading time (coefficient = –0.004, 95% CI [–0.006, –0.002], $p = .001$). The minus coefficient meant a positive correlation because we used the inverse link function and the predicted value was the inverse of reading time. The intercept of the model was 27.36 seconds; the inverse of the coefficient of the intercept = 0.037, 95% CI [0.018, 0.055], $p < .001$. If pre-attitudes increased one unit from the intercept, mean reading time increased to 30.43 seconds (3.07 seconds longer than the intercept and the inverse of 0.037 minus 0.004). Reading time decreased with a decrease in pre-attitudes in a hyperbolic way. Those who held negative pre-attitudes read the normative messages for less time than did those who held positive pre-attitudes. Although the former saw normative information for a short time period, descriptive normative messages promoted their behavior. Descriptive norms might induce conformity without reflection of normative information among those who hold negative pre-attitudes.

The hypothesis that pre-attitudes moderate the effects of descriptive norms was not supported in Study 1. We found no evidence that descriptive norms backfire owing to pre-attitudes. The results showed the promotive effects of descriptive norms on overt preparation behavior, and the effect size was considerable. The exploratory analysis suggested that holding negative pre-attitudes might not lead one to process normative messages as deeply as those who hold positive pre-attitudes; therefore, they impulsively conformed to the majority.

However, we could not determine whether descriptive norms cause undesirable consequences because the sample size was inadequate. In Study 2, we adopted a reliable sample pool to replicate Study 1.

Study 2

Although our preregistration only assumed Study 1 during the initial phases of the study, we decided to conduct Study 2 for a direct replication to resolve the sample quality problem. Study 2 had an identical study focus and almost the same procedures. We recruited reliable participants from a Japanese participant panel,

which was retained by a Japanese marketing company (Cross Marketing Inc.). All materials were uploaded to OSF before data collection.

Method

Participants. The requisite sample size was identical to Study 1—382 participants. Considering the failure rate of the instructional manipulation check and the attention check in Study 1, we aimed to recruit 416 participants. This sample would result in statistical power of .80 to detect an effect size odds ratio of 1.44. Cross Marketing Inc. already conducted the duplication checks about participants' registry and ensured the uniqueness of the participation; thus, we did not exclude participants based on geolocation data.

Participants who could understand Japanese and live in Japan participated. Participants were paid nonmonetary incentives, which can translate to some services and goods. A total of 541 responses were collected.

Participants who responded outside of Japan ($n = 3$), did not provide informed consent ($n = 25$), did not complete the questionnaire ($n = 41$), participated for less than 1 minute ($n = 75$), participated for more than 1 hour ($n = 1$), failed to pass the instructional manipulation check ($n = 66$), and failed to pass the attention check ($n = 72$) were excluded.

The number of participants who failed to pass the instructional manipulation check and the attention check exceeded our expectation; therefore, the actual sample size ($N = 329$; 107 women, 209 men, 13 did not report their sex) was insufficient to obtain our desired power. Recalculation of power indicated that the actual power was .74. The actual sample size yielded power of .80 to detect effects as small as odds ratio = 1.48. Participants' mean age was 52.89 years ($SD = 10.34$).

Procedure. We translated the English questionnaire used in Study 1 to Japanese. We did not apply back-translation because the authors' first language was Japanese. The description of natural disasters was modified to fit the Japanese context (i.e., focusing on earthquakes instead of hurricanes and other disasters as in Study 1). The need for cognition was measured by the Need for Cognition scale—Japanese edition (Kouyama & Fujihara, 1991). The other questionnaire parts and the procedure were identical to those of Study 1.

Results and Discussion

Descriptive statistics. The experiment was conducted in May 2019. Table 4 indicates the number and proportion of participants who read the leaflet in each condition and their pre-attitudes (stratified by median). Participants who held positive pre-attitudes and were in the descriptive norms condition showed the highest

Table 4. The Number and Proportion of Participants who Read the Leaflet ($N = 329$)

			Behavior	
			No *n* (%)	Yes *n* (%)
Condition				
Control	Pre-attitudes	High	14 (12.4)	99 (87.6)
		Low	17 (17.9)	78 (82.1)
Descriptive norms	Pre-attitudes	High	4 (7.5)	49 (92.5)
		Low	18 (26.5)	50 (73.5)

Note: The pre-attitudes indicating high or low stratified by the median (8.70).

probability to take preparation behavior. More importantly, participants who held negative pre-attitudes and were in the descriptive norms condition showed the lowest behavior rate. Those who held negative pre-attitudes suppressed their behavior when they were presented with descriptive norms. The median reading time for the leaflet was 42.83 seconds.

Confirmatory analysis: Model comparison between the main effects model and the interaction effect model. As in Study 1, we constructed two logistic regression models: the main effects model and the interaction effect model. Preparation behaviors were regressed by the explanatory variables and covariates. Pre-attitudes and need for cognition were computed by averaging each scale, which showed excellent internal consistency (Cronbach's $\alpha = .91$ and .88, respectively). To test interaction effects, a likelihood-ratio test was conducted to compare the deviance between the main effects model—that only included the main effects of descriptive norms and the pre-attitudes and covariates—and the interaction effect model with the interaction term between descriptive norms and pre-attitudes. Consequently, we rejected the main effects model and supported the interaction effect model ($\chi^2(1) = 11.66, p = 0.001$). This result confirmed our hypothesis that the interaction term significantly affected behavior. Pre-attitudes moderated the effect of descriptive norms. AIC also indicated the superiority of the interaction model. The VIFs of the interaction effect model were less than 1.86, indicating no multicollinearity problems. The results of the logistic regression analyses are shown in Table 5.

The log odds ratio of the interaction term was transformed to other effect sizes to further interpretation (Table 6). These effect size were somewhat larger than the predicted effect size. This effect size can be interpreted as a medium-sized effect. The interaction term consists of a product of dichotomously coded descriptive norms and centered pre-attitude; therefore, the obtained larger-than-one relative risk ratio meant that the probability of reading the leaflet increased by 1.11 times with every one-unit increase in pre-attitudes (reflecting positive attitudes) and decreased by 0.90 times with every one-unit decrease in pre-attitudes (reflecting

Table 5. The Estimates and the Standard Errors of the Logistic Regression Models ($N = 329$)

	Main effects model	Interaction effect model
Intercept	2.49*	2.59*
	(1.04)	(1.07)
Risk perception	0.00	0.00
	(0.01)	(0.01)
Need for cognition	−0.19	−0.26
	(0.22)	(0.23)
Sex	0.31	0.28
	(0.37)	(0.38)
Pre-attitudes	0.57***	0.29*
	(0.13)	(0.14)
Descriptive norms	−0.20	0.41
	(0.32)	(0.42)
Pre-attitudes × Descriptive norms		0.91**
		(0.29)
AIC	274.78	265.12
BIC	297.55	291.69
Log likelihood	−131.39	−125.56
Deviance	262.78	251.12

Note: The preparation behavior was regressed by the explanatory variables and covariates. Descriptive norms and the pre-attitudes were explanatory variables. Sex, need for cognition, and risk perception were considered covariates. The preparation behavior was coded as 0 = not read the leaflet, 1 = read the leaflet; descriptive norm was coded as 0 = control condition, 1 = descriptive norm condition; and sex was coded as 0 = non-female, 1 = female. Pre-attitude was centered. AIC refers to Akaike Information Criterion and BIC refers to Bayesian Information Criterion.
*$p < .05$; **$p < .01$; ***$p < .001$

Table 6. Effect Sizes of the Interaction Effect Model

	Coefficient (log odds ratio)	95% confidence interval (CI)	Odds ratio	95% CI	Risk ratio	95% CI	Cohen's *d*	95% CI
Pre-attitudes × Descriptive norms	0.91	[0.34, 1.48]	2.49	[1.41, 4.40]	1.11	[1.05, 1.14]	.50	[0.19, 0.82]

negative attitudes) in the descriptive norm condition. This direction of the obtained estimation empirically confirmed our prediction. In sum, negative pre-attitudes induced the adverse effects of descriptive norms.

Exploratory analyses: Examining the interaction between descriptive norms and pre-attitudes. A simple slope analysis (Aiken & West, 1991; Cohen & Cohen, 1983) was conducted to examine the interaction.[3] The results of the simple

[3] We preregistered that we would stratify the participants by a median of pre-attitudes and perform logistic regression repeatedly to examine the interaction effect. However, a simple slope analysis is

Table 7. The Effects of Descriptive Norms on Each Pre-Attitudes Point

	Coefficient	95% confidence interval (CI)	Risk ratio	95% CI
Negative Pre-attitudes	−0.82	[−1.58, −0.05]	0.84	[0.63, 0.99]
Positive Pre-attitudes	1.63	[0.25, 3.01]	1.14	[1.03, 1.17]

Note: Negative pre-attitudes was Mean –1 *SD* of pre-attitude. Positive pre-attitudes was Mean +1 *SD* of pre-attitudes. We adopted baseline risk (baseline possibility to act) as the probability of reading the leaflet in the control condition (85%).

slope analysis are shown in Table 7. Descriptive norms affected behaviors in the absolutory opposite direction, along with the values of pre-attitudes. The analysis showed that descriptive norms restrained the behavior where the moderator pre-attitudes were negative. The significant suppression effect was interpreted as evidence of the backfire effect. Descriptive normative messages could cause an unintended harmful result because of pre-attitudes. By contrast, descriptive norms promoted behavior when pre-attitudes were positive.

The results of the simple slope analysis confirmed our hypothesis. However, the suppression was far beyond our prediction. Though we preregistered that the promotion effect of descriptive norms would just be nullified among those who held negative attitudes, the results indicated that the norms even decreased the behavior. The results implied a larger pitfall than we imagined. Not only was the desirable promotion effect of descriptive norms diminished by negative pre-attitudes, but also the desirable behavior decreased. Further, among those who held positive pre-attitudes, descriptive norms worked as a promoter for the preparation behavior, and the effect size was roughly equivalent to the effect size found in Study 1.

We hypothesized the process of the backfire effect in that the participants who held negative pre-attitudes selectively paid attention to the minority who did not adopt preparation behaviors. It was predicted that this unintended focus would attenuate the promotive effects of descriptive norms. However, the obtained interaction effect and the results of the simple slope analysis might support our assumed hypothetical process of the backfire effect. If participants who held negative pre-attitudes completely ignored normative information, the interaction effect should not have emerged, and descriptive norms should have had neither a positive nor a negative effect among these individuals. The suppression effects of the norms implied that even the participants who held negative pre-attitudes

more appropriate for our objective than the stratification method because it allows for omnibus data analysis and does not sacrifice degrees of freedom. We primarily reported the results of the simple slope in this article. The results of the stratified logistic regression analysis are reported in the Appendix. The results showed nonsignificant but similar directed effects of descriptive norms. Among those who held negative pre-attitudes, descriptive norms decreased the chance of taking preparation behavior (log odds ratio <0), whereas the opposite was true for those held positive pre-attitudes (log odds ratio >0).

were subject to the double meaning of the normative messages: most adopted preparation behaviors; however, the minority did not.

Exploratory analyses: Examining reading time of normative messages. Measuring cognitive responses (Petty & Cacioppo, 1977, 1979) or eye-tracking methods to gauge whether participants observe the minority are straightforward ways of examining information processing. However, the recent data might be helpful to understand the mechanisms underlying the backfire effect.

We conducted another exploratory analysis to investigate the information processing of descriptive normative messages using the obtained data concerning the reading time of the normative messages. We presumed reading time as the amount of effort to process normative information. If the pre-attitudes did not relate to reading time, we would capture indirect evidence that participants who held negative pre-attitudes reflected upon the descriptive normative messages as the participants with positive pre-attitudes did.

One hundred twenty-one participants who were shown descriptive normative messages during the survey were extracted for the present analysis. When we stratified the participants by the median of pre-attitudes (8.50), we found the following descriptive statistics about the reading time of normative information. Among those who held negative pre-attitudes, median reading time was 28.41 seconds. Among those who held positive pre-attitudes, median reading time was 29.10 seconds. A simple correlation between reading time and pre-attitudes was $r = .07$, 95% CI [–0.11, 0.24], $p = .46$.

As in Study 1, their reading time was regressed by an explanatory variable—pre-attitudes. Sex, need for cognition, and risk perception were entered into the model as covariates. As in Study 1, a generalized linear model was built where the objective variable followed a Gamma distribution and the default inverse link function was used. The analysis suggested that pre-attitudes did not significantly affect the reading time of the normative messages (coefficient = –0.002, 95% CI [–0.006, 0.002], $p = .30$). The point estimate of the pre-attitudes included zero in the CI. It is presumable that participants processed the normative messages almost equally, regardless of their pre-attitudes.

Even though the nonsignificant results cannot be strong evidence to support our assumption, they provide partial insight into information processing of descriptive norms. The result that pre-attitudes did not strongly affect the duration time of normative information implied that the participants who held negative pre-attitudes reflected upon the normative messages similar to their counterparts. Allocating almost similar effort to the processing of the normative messages regardless of pre-attitudes gives rise to a possibility of focusing on the information of minorities among those who held negative pre-attitudes.

General Discussion

Backfire of Descriptive Norms Caused by Attitudes

The main statistical tests of model selection were inconsistent between the studies. In Study 1, the predicted interaction effect model was not adopted; however, the model was affirmed in Study 2. The results of Study 1 did not support the moderation hypothesis, and the promotive effect of descriptive norms was found regardless of participants' pre-attitudes. However, in Study 2, the moderation effect of pre-attitudes was obtained, and its direction was in line with our prediction. Positive pre-attitudes enhanced the promotive effect of descriptive norms despite that negative pre-attitudes led to the boomerang effects rather than just removing the promotive effects. Among those who held negative pre-attitudes, preparation behavior was suppressed by the normative messages compared to those in the control condition.

The conflicting results can be understood consistently through the exploratory analyses of the duration of reading the normative messages. In Study 1, in which the moderation effect did not occur, participants who held negative pre-attitudes read the normative messages for less time than their positive pre-attitude counterparts. This implies that the participants who held negative pre-attitudes did not examine the normative messages as much as the positive ones and impulsively conformed to the majority. Such a shallow process of normative information might lead to the uniform promotive effects of descriptive norms regardless of individuals' pre-attitudes. By contrast, the results of the exploratory analysis of Study 2 showed that pre-attitudes did not modulate the amount of processing time of the normative messages. We estimated that the participants who held negative pre-attitudes processed normative information relatively carefully. This might bring participants to focus on the minorities who did not adopt the preparation. Consequently, we conclude that descriptive norms backfire because of pre-attitudes under certain conditions. We have considered that the saliency of descriptive norms could serve as a possible marginal condition, as discussed below.

Explanation of the Inconsistent Results

Admittedly, the nationality of the participants differed; however, an influence process of descriptive norms might be a possible source of the heterogeneity. We used the same descriptive normative message in both studies—"68.1% of the participants read the leaflet." This proportion was obtained from a preliminary study with U.S. participants. The actual percentage in Study 1 was 64.1%—roughly equivalent to the presented information—however, 85.1% of the Japanese participants read the leaflet in Study 2. The effect of normative messages was thus relatively high in Study 1 as compared to Study 2. When the U.S. participants

faced the normative information, they might have estimated their own possibility of acting as if they had read the leaflet without the normative messages; in other words, they put themselves in the other participants' shoes. This possibility might be equal to or lower than the presented percentage; in fact, 51.1% of the participants in the control condition read the leaflet without descriptive norms. Comparing their estimation and the real information, they might assume that the normative information was salient. By contrast, the Japanese participants might have favored reading the leaflet and estimated that their own possibility was high. The obtained results in the control condition supported our assumption (i.e., 85.1% of the participants decided to read). The comparison between the estimation and the actual percentage was thus unsurprising. Consequently, the participants in Study 2 might think that the normative information was reasonable. We suggest that the distinct saliency of the normative messages played a key role.[4]

The saliency of descriptive norms might be an additional moderator. When the saliency was considerably high, participants processed normative information first because the features were prominent, and they terminated further information processing. This coincides with compulsive conformity to the majority, regardless of individuals' attitudes. When the saliency was low, however, people might consider, in parallel, both normative information and other traits like their personal beliefs or attitudes.

Ando, Ohnuma, and Chang (2007) found that U.S. university students were more sensitive to descriptive norms than Japanese students. The sensitivity difference also partially supports our speculation. U.S. participants might be uniformly affected by the normative messages because of their high sensitivity to descriptive norms and the relatively high saliency of the normative messages in Study 1; however, Japanese participants may be more resistant to descriptive norms. Previous cross-cultural examinations regarding the relationships between group norms and social identity provide clarification on these discrepant results. Bagozzi and Lee (2002) found that the group norm promoted behavioral intention among American participants, but not among Korean participants. Rather, the Korean participants were influenced by the social identification of social norms. The authors theorized that while American participants were directly subject to social pressures, Korean participants were indirectly affected by the effects of social pressure through their internalization of the pressure. This previous empirical difference between individualistic and collectivistic cultures bolsters our results that suggest that the descriptive norms directly influenced U.S. participants, but attitudes moderated

[4] It is assumed that individuals can infer other people's tendency accurately. However, there can be a misperception concerning how descriptive norms occur, especially when people become involved with an issue or interpersonal communication about a topic is activated (for a review, Lapinski & Rimal, 2005). By contrast, in these studies, in which we incidentally presented descriptive norms, neither feature was present. We posit that the degree of misperception was low in the current studies; thus, participants could infer others' behavior accurately.

this effect in Japanese participants. Furthermore, Bergquist et al. (2019) indicated that social norms were more influential for those living in individualistic countries than in collectivistic countries, contrary to their initial assumption.

The difference in the circumstances surrounding the natural disaster risk might contribute to inconsistent results between the two studies. Japan is more likely to experience natural hazards compared to the United States. Thus, Japanese participants might believe that Japanese citizens ought to prepare for natural disasters. In such circumstances, Japanese participants might be likely to assume that most other people would read the leaflet. The estimation could be somewhat higher than we reported in the current study. Consequently, the descriptive norms appeared to moderate and then cause the backfire.

From the social identity perspective, the internalization of norms could be another possible explanation for the lack of the backfire effect in Study 1. Hogg and Turner (1987) discriminated compliance without the agreement of the norms from conformity with the total acceptance of group norms (i.e., internalization of norms; Sherif, 1961). Two of their experiments demonstrated that conformity occurred among those who identified with the group. Another in their series of experiments did not report a significant interaction between norms and social identity; however, the authors suggested that extrinsic categorization was easily overridden by the natural categorization derived from the response feedback by other participants. In our experimental context, the extrinsic categorization was offered by stating that the majority of the ex-participants read the leaflet. However, the participants could easily find the subcategories of readers and nonreaders, making the natural categorization dominant. In Study 1, the participants with negative attitudes found it particularly challenging to identify with the majority who were prepared, but they complied without changing their attitudes. The participants with positive attitudes, however, could identify with most other people and internalized the normative information, which could also lead to the promotive effect of descriptive norms, although their motivation regarding following the majority could differ between those with negative and positive attitudes. This interpretation should be confirmed in future research that examines post-attitudes, which could clarify if negative pre-attitudes participants only conform to the majority, without changing their post-attitudes.

The Promotive Effects of Descriptive Norms

In line with prior research, descriptive norms promoted desirable activities. In Study 1, we obtained medium-sized promotive effects targeting the actual behavior rather than a behavioral intention. We posit that descriptive normative approaches can be applied to promote disaster preparation.

In Study 2, an interaction effect between descriptive norms and pre-attitudes was obtained, although descriptive norms did not solely affect behavior. The results

also suggested the effectiveness of descriptive normative messages among those who held positive pre-attitudes. The obtained effect size of descriptive norms on those who held positive pre-attitudes was also equivalent to the effects found in Study 1. We concluded that the promotive effects of the descriptive norms was consistently found in both studies—with both U.S. and Japanese participants—implying the universality of the results.

Practical Implications

Farrow, Grolleau, and Ibanez, (2017) illustrated a decision tree for applications of the normative intervention. In their research, they were concerned about the undesirable consequences of descriptive normative information because the type of intervention was "scatter-shot" (Miller & Prentice, 2016). They also recommended that the approaches should set targets accurately. Our finding—that attitudes moderated the effects of descriptive norms—elucidates a practical application. When we conduct descriptive normative interventions to promote desirable behaviors, it is crucial to consider individuals' attitudes that can induce unintended outcomes.

In addition, the absolute pre-attitudes value was high in both studies, especially in Study 2. Although participants were favorable toward disaster preparation, relatively negative pre-attitudes (although still above the midpoint) induced the adverse effects of descriptive norms. Notably, the pitfalls of normative approaches are difficult to avoid.

The suppression effect in Study 2 was the first evidence that descriptive norms backfire even when the majority act in a desirable way. Richter, Thøgersen, and Klöckner, 2018 found that the boomerang effect occurred when descriptive normative message stated that less than half of others performed the designated behavior; however, we showed that even descriptive norms in which more than half perform the behavior can induce adverse effects.

Limitations

We measured participants' behavior—whether they read the leaflet during the online experiments. This measurement was selected mainly because of the feasibility of measuring actual behavior, and because we believed that the index was sensitive to our experimental manipulation, it worked as a metaphor of the preparation behavior. However, for the effectiveness of practical preparation, the behaviors should link directly to the protection of people's lives and property. Future researchers should investigate whether descriptive norms work as well as they did in the present studies when examining costly behaviors such as an antiseismic reinforcement, obtaining disaster insurance, and disaster preparation education. Further, the present results only focused on individuals'

information-seeking behavior concerning disaster preparation. Multifaceted research should be conducted in the future.

In the current study, we examined only one level of descriptive norms. The moderate level of descriptive norms implied the swinging state of others and that the designated behavior was not firmly present. Based on our results, when future researchers try to replicate this study, stronger descriptive norms (e.g., norms from 70% or more of others) might not show the backfire effect, as the norms seem to overwhelm other psychological factors, such as attitudes.

Conclusions

Descriptive norms backfired when participants held negative attitudes. Attitudes modified the influence of descriptive norms in Study 2. Among those who held negative attitudes, showing descriptive norms reduced the desirable behaviors. We assumed this result is evidence of the backfire of a descriptive normative intervention. We believe this result is not an artifact because the direction of the moderation effect was as predicted, and the effect size was considerable. The exploratory analyses of the duration time of the normative messages also supported our assumption.

In Study 2, participants' attitudes did not affect the reading time of the normative messages in contrast to the results of the same analysis in Study 1. Even if participants held negative attitudes, they considered the descriptive norms and might be affected by the minority who did not prepare for the disasters. Comparing the discrepant results might help elucidate the underlying mechanisms behind the backfire effect. Perhaps, as we suggested, there is another marginal condition that underlies the backfire effect. We finally conclude that descriptive norms backfire in a certain condition, especially when descriptive norms were not strongly salient.

Open Research Badges

This article has earned Open Data, Open Materials and Preregistered Research Design badges. Data, materials and the preregistered design and analysis plan are available at https://osf.io/yzdqh/ and https://osf.io/9cg4x.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: SAGE.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, *95*(9–10), 1082–1095. https://doi.org/10.1016/j.jpubeco.2011.03.003

Ando, K., Ohnuma, S., & Chang, E. C. (2007). Comparing normative influences as determinants of environmentally conscious behaviours between the USA and Japan. *Asian Journal of Social Psychology*, *10*(3), 171–178. https://doi.org/10.1111/j.1467-839X.2007.00223.x

Bagozzi, R. P., & Lee, K.-H. (2002). Multiple routes for social influence: The role of compliance, internalization, and social identity. *Social Psychology Quarterly*, *65*(3), 226–247. https://doi.org/10.2307/3090121

Baron, J. (2008). *Thinking and deciding* (4th ed.). New York, NY: Cambridge University Press.

Bergquist, M., Nilsson, A., & Schultz, W. P. (2019). A meta-analysis of field-experiments using social norms to promote pro-environmental behaviors. *Global Environmental Change*, *59*, 101941. https://doi.org/10.1016/j.gloenvcha.2019.101941

Brock, T. C., & Balloun, J. L. (1967). Behavioral receptivity to dissonant information. *Journal of Personality and Social Psychology*, *6*(4, pt 1), 413–428. https://doi.org/10.1037/h0021225

Bubeck, P., Botzen, W. J. W., & Aerts, J. C. J. H. (2012). A review of risk perceptions and other factors that influence flood mitigation behavior. *Risk Analysis*, *32*(9), 1481–1495. https://doi.org/10.1111/j.1539-6924.2011.01783.x

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306–307. https://doi.org/10.1207/s15327752jpa4803_13

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*, 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (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), 1015–1026. https://doi.org/10.1037/0022-3514.58.6.1015

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lowrence Erlbaum.

Costa, D. L., & Kahn, M. E. (2013). Energy conservation "nudges" and environmentalist ideology: Evidence from a randomized residential electricity field experiment. *Journal of the European Economic Association*, *11*(3), 680–702. https://doi.org/10.1111/jeea.12011

Dempsey, R. C., McAlaney, J., & Bewick, B. M. (2018). A critical appraisal of the social norms approach as an interventional strategy for health-related behavior and attitude change. *Frontiers in Psychology*, *9*, 2180. https://doi.org/10.3389/fpsyg.2018.02180

Duval, T. S., & Mulilis, J. P. (1999). A person-relative-to-event (PrE) approach to negative threat appeals and earthquake preparedness: A field study. *Journal of Applied Social Psychology*, *29*(3), 495–516. https://doi.org/10.1111/j.1559-1816.1999.tb01398.x

Farrow, K., Grolleau, G., & Ibanez, L. (2017). Social norms and pro-environmental behavior: A review of the evidence. *Ecological Economics*, *140*, 1–13. https://doi.org/10.1016/j.ecolecon.2017.04.017

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Flynn, J., Slovic, P., & Mertz, C. K. (1994). Gender, race, and perception of environmental health risks. *Risk Analysis*, *14*(6), 1101–1108. https://doi.org/10.1111/j.1539-6924.1994.tb00082.x

Fraser-Mackenzie, P., & Dror, I. (2009). Selective information sampling: Cognitive coherence in evaluation of a novel item. *Judgment and Decision Making*, *4*(4), 307–316.

Frey, D. (1986). Recent research on selective exposure to information. *Advances in Experimental Social Psychology*, *19*, 41–80. https://doi.org/10.1016/S0065-2601(08)60212-9

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research*, *35*(3), 472–482. https://doi.org/10.1086/586910

Grant, R. L. (2014). Converting an odds ratio to a range of plausible relative risks for better communication of research findings. *BMJ*, *348*, f7450. https://doi.org/10.1136/bmj.f7450

Heider, F. (1958). *The psychology of interpersonal relations*. Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1037/10628-000

Hogg, M. A., & Abrams, D. (1988). *Social identifications: A social psychology of intergroup relations and group processes*. New York, NY: Routledge.

Hogg, M. A., & Turner, J. C. (1987). Social identity and conformity: A theory of referent informational influence. In W. Doise & S. Moscovici (Eds.), *Current issues in European social psychology* (Vol. 2, pp. 139–182). Cambridge, UK: Cambridge University Press.

Keizer, K., Lindenberg, S., & Steg, L. (2008). The spreading of disorder. *Science*, *322*(5908), 1681–1685. https://doi.org/10.1126/science.1161405

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs'' replication project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Kouyama, T., & Fujihara, T. (1991). A basic study of the need for cognition scale. *Research in Social Psychology*, *6*(3), 184–192.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498.

Kunreuther, H. C., Ginsberg, R., Miller, L., Sagi, P., Slovic, P., Borkan, B., & Katz, N. (1978). *Disaster insurance protection*. New York, NY: Wiley-Interscience.

Lapinski, M. K., & Rimal, R. N. (2005). An explication of social norms. *Communication Theory*, *15*(2), 127–147. https://doi.org/10.1111/j.1468-2885.2005.tb00329.x

Leifeld, P. (2013). texreg: Conversion of statistical model output in R to LATEX and HTML tables. *Journal of Statistical Software*, *55*(8), 1–24. https://doi.org/10.18637/jss.v055.i08

Levac, J., Toal-Sullivan, D., & O'Sullivan, T. L. (2012). Household emergency preparedness: A literature review. *Journal of Community Health*, *37*(3), 725–733. https://doi.org/10.1007/s10900-011-9488-x

Lewis, M. A., Neighbors, C., Oster-Aaland, L., Kirkeby, B. S., & Larimer, M. E. (2007). Indicated prevention for incoming freshmen: Personalized normative feedback and high-risk drinking. *Addictive Behaviors*, *32*(11), 2495–2508. https://doi.org/10.1016/j.addbeh.2007.06.019

Lindell, M. K., & Perry, R. W. (1992). *Behavioral foundations of community emergency planning*. Washington, DC: Hemisphere Press.

Lindell, M. K., & Perry, R. W. (2012). The protective action decision model: Theoretical modifications and additional evidence. *Risk Analysis*, *32*(4), 616–632. https://doi.org/10.1111/j.1539-6924.2011.01647.x

Lo, A. Y. (2013). The role of social norms in climate adaptation: Mediating risk perception and flood insurance purchase. *Global Environmental Change*, *23*(5), 1249–1257. https://doi.org/10.1016/j.gloenvcha.2013.07.019

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098–2109. https://doi.org/10.1037/0022-3514.37.11.2098

Mackie, D. M. (1986). Social identification effects in group polarization. *Journal of Personality and Social Psychology*, *50*(4), 720–728. https://doi.org/10.1037/0022-3514.50.4.720

Melnyk, V., Herpen, E. V., Fischer, A. R. H., & van Trijp, H. C. M. (2011). To think or not to think: The effect of cognitive deliberation on the influence of injunctive versus descriptive social norms. *Psychology and Marketing*, *28*(7), 709–729. https://doi.org/10.1002/mar.20408

Miller, D. T., & Prentice, D. A. (2016). Changing norms to change behavior. *Annual Review of Psychology*, *67*, 339–361. https://doi.org/10.1146/annurev-psych-010814-015013

Morsink, K., & Geurts, P. (2012). *The trusted neighbour effect: Local experience and demand for microinsurance*. Institute of Governance Studies (IGS) Working Paper.

Mulilis, J.-P., & Duval, T. S. (1995). Negative threat appeals and earthquake preparedness: A person-relative-to-event (PrE) model of coping with threat. *Journal of Applied Social Psychology*, *25*(15), 1319–1339. https://doi.org/10.1111/j.1559-1816.1995.tb02620.x

Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin*, *83*(4), 602–627. https://doi.org/10.1037/0033-2909.83.4.602

National Research Council. (1989). *Improving risk communication*. Washington, DC: National Academy Press.

Neighbors, C., Jensen, M., Tidwell, J., Walter, T., Fossos, N., & Lewis, M. A. (2011). Social-norms interventions for light and nondrinking students. *Group Processes & Intergroup Relations*, *14*(5), 651–669. https://doi.org/10.1177/1368430210398014

Neighbors, C., LaBrie, J. W., Hummer, J. F., Lewis, M. A., Lee, C. M., Desai, S., … Larimer, M. E. (2010). Group identification as a moderator of the relationship between perceived

social norms and alcohol consumption. *Psychology of Addictive Behaviors*, *24*(3), 522–528. https://doi.org/10.1037/a0019944

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Nyborg, K., Anderies, J. M., Dannenberg, A., Lindahl, T., Schill, C., Schlüter, M., … de Zeeuw, A. (2016). Social norms as solutions. *Science*, *354*(6308), 42–43. https://doi.org/10.1126/science.aaf8317

Onuma, H., Shin, K. J., & Managi, S. (2017). Household preparedness for natural disasters: Impact of disaster experience and implications for future disaster risks in Japan. *International Journal of Disaster Risk Reduction*, *21*, 148–158. https://doi.org/10.1016/j.ijdrr.2016.11.004

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Paek, H.-J., Hilyard, K., Freimuth, V., Barge, J. K., & Mindlin, M. (2010). Theory-based approaches to understanding public emergency preparedness: Implications for effective health and risk communication. *Journal of Health Communication*, *15*(4), 428–444. https://doi.org/10.1080/10810731003753083

Palmer, E. M., Horowitz, T. S., Torralba, A., & Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 58–71. https://doi.org/10.1037/a0020747

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.

Paton, D. (2003). Disaster preparedness: A social-cognitive perspective. *Disaster Prevention and Management*, *12*(3), 210–216. https://doi.org/10.1108/09653560310480686

Petty, R. E., & Cacioppo, J. T. (1977). Forewarning, cognitive responding, and resistance to persuasion. *Journal of Personality and Social Psychology*, *35*(9), 645–655. https://doi.org/10.1037/0022-3514.35.9.645

Petty, R. E., & Cacioppo, J. T. (1979). Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. *Journal of Personality and Social Psychology*, *37*(10), 1915–1926. https://doi.org/10.1037/0022-3514.37.10.1915

Plows, S. E., Smith, F. D., Smith, J. R., Chapman, C. M., La Macchia, S. T., & Louis, W. R. (2017). Healthy eating: A beneficial role for perceived norm conflict? *Journal of Applied Social Psychology*, *47*(6), 295–304. https://doi.org/10.1111/jasp.12430

Prince, M. A., Reid, A., Carey, K. B., & Neighbors, C. (2014). Effects of normative feedback for drinkers who consume less than the norm: Dodging the boomerang. *Psychology of Addictive Behaviors*, *28*(2), 538–544. https://doi.org/10.1037/a0036402

Ratner, R. K., & Riis, J. (2014). Communicating science-based recommendations with memorable and actionable guidelines. *Proceedings of the National Academy of Sciences*, *111*(Supplement 4), 13634–13641. https://doi.org/10.1073/pnas.1320649111

Reicher, S. D. (1987). Crowd behaviour as social action. In J. C. Turner (Ed.), *Rediscovering the social group: Self-categorization theory* (pp. 171–202). Oxford, UK: Blackwell.

Richter, I., Thøgersen, J., & Klöckner, C. (2018). A social norms intervention going wrong: Boomerang effects from descriptive norms information. *Sustainability*, *10*(8), 2848. https://doi.org/10.3390/su10082848

Rinker, D. V., & Neighbors, C. (2014). Do different types of social identity moderate the association between perceived descriptive norms and drinking among college students? *Addictive Behaviors*, *39*(9), 1297–1303. https://doi.org/10.1016/j.addbeh.2014.03.018

Rivis, A., & Sheeran, P. (2003). Descriptive norms as an additional predictor in the theory of planned behaviour: A meta-analysis. *Current Psychology*, *22*(3), 218–233. https://doi.org/10.1007/s12144-003-1018-2

Rogers, R. W. (1975). A protection motivation theory of fear appeals and attitude change. *The Journal of Psychology*, *91*(1), 93–114. https://doi.org/10.1080/00223980.1975.9915803

Rogers, R. W. (1983). Cognitive and physiological processes in fear appeals and attitude change: A revised theory of protection motivation. In J. T. Cacioppo & R. E. Petty (Eds.), *Social psychophysiology: A sourcebook* (pp. 153–177). New York, NY: Guilford Press.

Rosenberg, M. J., & Hovland, C. I. (1960). Cognitive, affective and behavioral components of attitudes. In M. J. Rosenberg & C. I. Hovland (Eds.), *Attitude organization and change: An analysis of consistency among attitude components* (pp. 1–14). New Haven, CT: Yale University Press.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, *18*(5), 429–434. https://doi.org/10.1111/j.1467-9280.2007.01917.x

Sheeran, P., Abraham, C., & Orbell, S. (1999). Psychosocial correlates of heterosexual condom use: A meta-analysis. *Psychological Bulletin*, *125*(1), 90–132. https://doi.org/10.1037/0033-2909.125.1.90

Sherif, M. (1961). Conformity-deviation, norms, and group relations. In I. A. Berg & B. M. Bass (Eds.), *Conformity and deviation* (pp. 159–198). New York, NY: Harper and Brothers. https://doi.org/10.1037/11122-000

Shreve, C., Fordham, M., Anson, S., Watson, H., Hagen, K., Wadhwa, K., … Karanci, N. (2014). *Report on risk perception and preparedness*. Deliverable 1.1 of the TACTIC Project 31.

Snyder, M., & Cantor, N. (1979). Testing hypotheses about other people: The use of historical knowledge. *Journal of Experimental Social Psychology*, *15*(4), 330–342. https://doi.org/10.1016/0022-1031(79)90042-8

Solberg, C., Rossetto, T., & Joffe, H. (2010). The social psychology of seismic hazard adjustment: Re-evaluating the international literature. *Natural Hazards and Earth System Sciences*, *10*(8), 1663–1677. https://doi.org/10.5194/nhess-10-1663-2010

Stok, F. M., De Ridder, D. T. D., De Vet, E., & De Wit, J. B. F. (2014). Don't tell me what I should do, but what others do: The influence of descriptive and injunctive peer norms on fruit consumption in adolescents. *British Journal of Health Psychology*, *19*(1), 52–64. https://doi.org/10.1111/bjhp.12030

Tajfel, H. (1982). Social psychology of intergroup relations. *Annual Review of Psychology*, *33*(1), 1–39. https://doi.org/10.1146/annurev.ps.33.020182.000245

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–37). Monterey, CA: Brooks/Cole.

Turner, J. C. (1982). Towards a cognitive redefinition of the social group. In H. Tajfel (Ed.), *Social identity and intergroup relations* (pp. 15–40). Cambridge, UK: Cambridge University Press.

Turner, J. C. (1987). *Rediscovering the social group: Self-categorization theory*. New York, NY: Blackwell.

van 't Veer, A., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2–12. https://doi.org/10.1016/j.jesp.2016.03.004

Vinnell, L. J., Milfont, T. L., & McClure, J. (2019). Do social norms affect support for earthquake-strengthening legislation? Comparing the effects of descriptive and injunctive norms. *Environment and Behavior*, *51*(4), 376–400. https://doi.org/10.1177/0013916517752435

Wachinger, G., Renn, O., Begg, C., & Kuhlicke, C. (2013). The risk perception paradox: Implications for governance and communication of natural hazards. *Risk Analysis*, *33*(6), 1049–1065. https://doi.org/10.1111/j.1539-6924.2012.01942.x

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140. https://doi.org/10.1080/17470216008416717

Werch, C. E., & Owen, D. M. (2002). Iatrogenic effects of alcohol and drug prevention programs. *Journal of Studies on Alcohol*, *63*(5), 581–590. https://doi.org/10.15288/jsa.2002.63.581

Wetherell, M. S. (1987). Social identity and group polarization. In J. C. Turner (Ed.), *Rediscovering the social group: Self-categorization theory* (pp. 142–170). New York, NYA: Blackwell.

Wilder, D. A. (1990). Some determinants of the persuasive power of in-groups and out-groups: Organization of information and attribution of independence. *Journal of Personality and Social Psychology*, *59*(6), 1202–1213.

Zhang, J., & Yu, K. F. (1998). What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *Journal of the American Medical Association*, *280*(19), 1690–1691. https://doi.org/10.1001/jama.280.19.1690

Appendix

We conducted two logistic regression analyses to estimate the effects of descriptive norms among those who held negative and positive pre-attitudes. We stratified participants on the median of their pre-attitudes. The behavior—whether they read the leaflet—was regressed by the experimental manipulation and the covariates. Among those who held negative pre-attitudes ($n = 163$), the effect of descriptive norms was negative but nonsignificant, coefficient = –0.50, 95% CI [–1.26, 0.25], $p = .19$). On the other hand, among those who held positive pre-attitudes ($n = 166$), the effect of descriptive norms was positive but nonsignificant, coefficient = 0.54, 95% CI [–0.64, 1.72], $p = .37$. Both p-values were not corrected by Bonferroni adjustment. We posit that the nonsignificant results were because of the sacrifice of the degrees of freedom. However, the directions of the effect of descriptive norms were congruent with our prediction and the simple slope analysis.

TAKU OZAKI is a PhD candidate in the Graduate School of Psychology, Doshisha University. He obtained an M.A. in Psychology at Doshisha University, Japan. His research focuses on the social norm approaches toward natural disasters. His doctoral research includes the examination of the effectiveness and ineffectiveness of the social norm approaches.

KAZUYA NAKAYACHI is a Full Professor at Faculty of Psychology, Doshisha University. He earned his PhD in psychology from Doshisha University in 2003. His research interests focus on risk perception, trust in risk managers, and preparedness for disasters. He has published several articles in *Risk Analysis* as well as in *International Journal of Disaster Risk Reduction*, and received the Award of the Best Paper for 2013 in *Risk Analysis*.