Journal of Behavior Therapy and Experimental Psychiatry 61 (2018) 134–141

Contents lists available at ScienceDirect

Journal of Behavior Therapy and Experimental Psychiatry

journal homepage: www.elsevier.com/locate/jbtep

Trigger warning: Empirical evidence ahead

Benjamin W. Bellet*, Payton J. Jones, Richard J. McNally

Department of Psychology, Harvard University, Cambridge, MA, USA

ARTICLE INFO

Keywords:
Trigger warning
Anxiety
PTSD
Resilience
Vulnerability

ABSTRACT

Background and objectives: Trigger warnings notify people of the distress that written, audiovisual, or other material may evoke, and were initially used to provide for the needs of those with posttraumatic stress disorder (PTSD). Since their inception, trigger warnings have become more widely applied throughout contemporary culture, sparking intense controversy in academia and beyond. Some argue that they empower vulnerable individuals by allowing them to psychologically prepare for or avoid disturbing content, whereas others argue that such warnings undermine resilience to stress and increase vulnerability to psychopathology while constraining academic freedom. The objective of our experiment was to investigate the psychological effects of issuing trigger warnings.

Methods: We randomly assigned online participants to receive ($n = 133$) or not receive ($n = 137$) trigger warnings prior to reading literary passages that varied in potentially disturbing content.

Results: Participants in the trigger warning group believed themselves and people in general to be more emotionally vulnerable if they were to experience trauma. Participants receiving warnings reported greater anxiety in response to reading potentially distressing passages, but only if they believed that words can cause harm. Warnings did not affect participants' implicit self-identification as vulnerable, or subsequent anxiety response to less distressing content.

Limitations: The sample included only non-traumatized participants; the observed effects may differ for a traumatized population.

Conclusions: Trigger warnings may inadvertently undermine some aspects of emotional resilience. Further research is needed on the generalizability of our findings, especially to collegiate populations and to those with trauma histories.

1. Introduction

Is it better to warn people about potentially distressing material, or allow them to deal with it on their own terms? Trigger warnings and other protective measures implemented at institutions of higher learning, such as safe spaces and the dis-invitation of potentially offensive speakers, have become the subject of contentious, widespread debate (Wilson, 2015). In the classroom, a trigger warning is the practice of "teachers offering prior notification of an educational topic so that students may prepare for or avoid distress that is automatically evoked by that topic due to clinical mental health problems" (Boysen, 2017, p. 164). Much support for trigger warnings arises from the desire to provide students with posttraumatic stress disorder (PTSD) and other disadvantaged groups with an inclusive, level academic playing field (Carter, 2015; Stokes, 2014). However, others believe that trigger warnings hamper free academic inquiry and "coddle" students by sheltering them from any stressful material they may encounter (Lukianoff & Haidt, 2015), thereby undermining their preparation for the "real world" beyond the campus gates.

The use of trigger warnings is supported by evidence that individuals with PTSD can experience painful recollections of trauma in response to reminders of their experience (American Psychiatric Association, 2013); trigger warnings may help those with PTSD to choose the time and place of their exposure to reminders, or psychologically brace for them (Boysen, 2017). However, trigger warnings may encourage avoidance of cues related to trauma (McNally, 2014,2016). Avoidance runs counter to the aims of prolonged exposure (PE) therapy, the most efficacious treatment for PTSD (Institute of Medicine, 2008). PE encourages systematic exposure to triggers, enabling patients to habituate to them and regain functioning. Conversely, avoidance of triggers may diminish distress in the short term, but worsens symptom severity in the long term (Rosenthal, Hall, Palm, Batten, & Follette, 2005). Further, receiving trigger warnings about trauma-related cues may enhance the centrality of traumatic events to

* Corresponding author. Department of Psychology, Harvard University, 33 Kirkland St., Cambridge, MA, 02138, USA.
E-mail address: bbellet@g.harvard.edu (B.W. Bellet).

https://doi.org/10.1016/j.jbtep.2018.07.002
Received 25 April 2018; Received in revised form 20 July 2018; Accepted 26 July 2018
Available online 27 July 2018
0005-7916/

survivors' identities (McNally, 2014), reminding them to view material through the lens of trauma. Regarding trauma as central to one's identity is associated with severity of PTSD symptoms (Berntsen & Rubin, 2007; Boelen, 2012; Robinaugh & McNally, 2011). Clearly, the question of whether trigger warnings help or harm trauma survivors has been the subject of much spirited debate, with plausible arguments on both sides of the aisle.

However, the use of trigger warnings has spread beyond efforts to accommodate only trauma survivors; trigger warnings have been used more broadly to shield members of other disadvantaged groups from a wide range of content, including depictions of classism and privilege (Boysen, 2017; Lukianoff & Haidt, 2015). Further, trigger warnings have become normative in settings other than academia, such as online discussion groups (Wyatt, 2016). The question of whether trigger warnings are beneficial or harmful for trauma survivors is an important one. However, because trigger warnings are now applied to a broad range of content in many different settings, another important question is whether they foster attitudes that undermine resilience in people who have not – or not yet – experienced trauma. Despite the timeliness and importance of this question, experimental research has remained silent on the subject.

Concerns about how trigger warnings affect trauma survivors, such as avoidance behaviors and trauma centrality, are distinct from those of interest in trauma-naïve individuals. One area of concern for those not yet traumatized is whether trigger warnings increase individuals' vulnerability to psychopathology, i.e. developing PTSD *in the event of* exposure to trauma. Although trauma is common, PTSD is rare (Breslau & Kessler, 2001; McNally, 2014). Experiencing some symptoms of PTSD in the immediate aftermath of a traumatic event *is* common, but symptoms rarely persist (Rothbaum, Foa, Riggs, Murdock, & Walsh, 1992). Trauma survivors who appraise acute symptoms negatively are at heightened risk for PTSD (Dunmore, Clark, & Ehlers, 2001; Ehring, Ehlers, & Glucksman, 2006). Trigger warnings suggest that trauma survivors will have difficulty with content encountered in daily life, and may lead people to believe that they are likely to develop PTSD should they encounter trauma, causing them to iatrogenically catastrophize acute posttraumatic symptoms. Further, receiving constant reminders of potential emotional harm may contribute to perceptions of heightened vulnerability, fostering a maladaptive self-identification as a victim (Wyatt, 2016).

Similarly, trigger warnings may also change the way that people think about others' vulnerability in the wake of trauma. Trigger warnings may raise awareness of the difficulties of people suffering from PTSD. However, they may also create the impression that the experience of trauma always renders survivors emotionally incapacitated. In reality, most trauma survivors are resilient and show few symptoms of PTSD after an initial period of adjustment (Breslau & Kessler, 2001). The perception of trauma survivors as dysregulated victims may contribute to negative stigma concerning the very individuals trigger warnings are intended to protect.

Trigger warnings may also ironically increase acute anxiety by producing an expectation of negative consequences. Indeed, nocebo effects (detrimental effects produced by negative expectations) are an established phenomenon in psychological research (e.g. Barsky, Saintfort, Rogers, & Borus, 2002). Research provides some support for a nocebo effect of trigger warnings (Bruce, 2017a) indicating that physiological markers of anxiety are heightened in the presence of trigger warnings in comparison to “PG-13” warning and “no warning” conditions. Such an effect may be exacerbated for individuals who already harbor the belief that exposure to offensive words or other media can cause long-lasting emotional harm.

On the other hand, perception of control over stressors reduces stress reactions (Thompson, 1981), and predictable stressors are less distressing than unpredictable ones (Grupe & Nitschke, 2013; Mineka & Kihlstrom, 1978). Distressing physiological sensations produce more anxiety when they violate expectations (Telch, Harrington, Smits, & Powers, 2011). Therefore, trigger warnings may enable people who choose to view the material to brace themselves for disturbing content without being surprised and dysregulated by its presentation. Alternatively, trigger-warning accustomed individuals may develop the implicit assumption that offensive content can always be anticipated, rendering even relatively innocuous content viewed without a warning surprising and more fearful (the cognitive equivalent of Lukianoff and Haidt's “coddling” hypothesis). Such an effect may be exacerbated for individuals who are already have high expectations of controllability and predictability in their daily lives.

1.1. The current study

Taken together, some research suggests that trigger warnings could be conducive to better emotional functioning and lower anxiety levels, whereas other research indicates that they may be anxiogenic and generative of risk for developing PTSD in the event of trauma. Despite these equally plausible hypotheses (and the spirited political debate surrounding trigger warnings), there is a dearth of research on trigger warnings' impact on resilience factors in the non-traumatized population.

1.1.1. Aims

Working within the tradition of experimental psychopathology, we sought to determine whether (and in what way) trigger warnings affect resilience variables specific to those who have not yet experienced potentially traumatic events. We also explored other demographic characteristics that may influence these resilience variables, and examined the reasons that individuals might support the use of trigger warnings, apart from their psychological reactions to them.

To achieve these aims, we recruited participants who had not experienced canonical traumatic events. We restricted our sample to trauma-naïve individuals because we wanted to examine how trigger warnings affect aspects of resilience specific to those who have not yet been traumatized (e.g., perceived emotional vulnerability *in the event of* experiencing trauma), which are distinct from those that concern traumatized individuals (e.g., encouraging avoidance behaviors). We had participants read distressing passages from world literature either with trigger warnings (experimental condition) or without trigger warnings (control condition) prior to reporting their anxiety levels after each passage. Participants then completed measures addressing perceptions of vulnerability in themselves and others. To test whether trigger warnings affect subsequent emotional reactivity to less distressing content, we included moderately distressing passages without a trigger warning at the end of the study. We also wanted to assess traits that may influence one's anxiety response to a trigger warning. Accordingly, we measured participants' strength of belief that words can harm people, enabling us to test whether it affects anxiety in response to potentially distressing material preceded by a trigger warning. We also measured participants' assumptions about how controllable and predictable the world is to test whether such beliefs increase anxiety provoked by less distressing material not preceded by trigger warnings.

1.1.2. Research questions

Due to different sources of indirect evidence suggesting that trigger warnings may be either detrimental or helpful to resilience, and the lack of empirical data on this topic, we formed research questions about whether trigger warning use would influence resilience variables, rather than making a priori hypotheses as to the direction of such effects. Accordingly, we tested whether trigger warnings would (Q1) affect participants' perceptions of their posttraumatic vulnerability, (Q2) affect participants' overall degree of implicit identification as “vulnerable” versus “resilient”, and (Q3) affect participants' perceptions of others' posttraumatic vulnerability. We also tested whether trigger warnings would (Q4) affect immediate anxiety response to potentially

distressing material, and whether the belief that words can cause harm might amplify an anxiety response. We also examined whether (Q5) trigger warnings would affect subsequent anxiety response to less distressing material, and whether stronger beliefs in the world's controllability and predictability might amplify this anxiety response.

2. Method

2.1. Participants

Participants were recruited on Amazon's Mechanical Turk (MTurk; Berinsky, Huber, & Lenz, 2012), then read and acknowledged an institutionally approved informed consent form. A single-item screening question excluded individuals who had experienced a canonical stressor (e.g., rape, natural disaster) qualifying for Criterion A of the PTSD diagnosis in *DSM-5* (American Psychiatric Association, 2013). Three hundred participants completed the study. Four participants were excluded from all analyses because they reported having received a diagnosis of PTSD despite denying exposure to canonical traumatic stressors. An additional 26 participants were excluded because they answered content-based attention check questions incorrectly, indicating inattentive responding. This left 270 participants, 133 in the Trigger Warning condition, and 137 in the No Warning condition.

2.2. Materials

To simulate an academic setting, we chose passages from world literature that commonly appear in high school or college courses. Each passage was standardized in word length, and passage exposures were set to a minimum of 20 s before participants were allowed to continue to the next screen. Transparent attention checks based on the passages' content assessed whether participants were attentively reading the passages (see supplementary materials S1 for an example of a content check question). We used three types of passages. *Neutral* passages were devoid of disturbing content (e.g. a character description from Herman Melville's *Moby-Dick*). *Mildly distressing* passages concerned themes of violence, injury, or death, but lacked graphic details (e.g. a description of a battle from James Bradley's *Flags of Our Fathers*). *Markedly distressing* passages contained graphic descriptions of violence, injury, or death (e.g. the murder scene from Fyodor Dostoevsky's *Crime and Punishment*). See supplementary materials (S1) for a sample passage from each category.

To ensure that our passages elicited levels of anxiety consistent with their categories, we conducted a pilot study involving 50 participants on MTurk to norm each passage's anxiogenic properties. Forty candidate passages were included. Means and inter-quartile ranges (IQRs) of anxiety response were calculated for each passage. Passages with the lowest means and IQRs that did not extend above the grand mean were designated *neutral*. Passages with means closest to the grand mean and IQRs within the grand IQR were designated *mildly distressing*. Passages with the highest means and IQRs that extended above the grand IQR were designated *markedly distressing*.

2.3. Measures

2.3.1. Perceived posttraumatic vulnerability scale-self (PPVS-S)

The PPVS-S is a 19-item questionnaire which assesses belief in the likelihood of long-term adverse emotional effects of trauma exposure. These perceived vulnerabilities include developing a mental disorder, being unable to effectively regulate emotions, or functional disability. Participants are asked to imagine themselves experiencing a hypothetical traumatic event, and to indicate their level of endorsement for each statement concerning its effects (e.g. *I would lose my grip on reality.*) on a 100-point scale (1 = *disagree*, 100 = *agree*). These responses are averaged for a composite score. Higher scores indicate stronger belief in vulnerability. The PPVS-S displayed excellent internal consistency in our sample ($\alpha = 0.95$). All measures devised for this experiment are in the supplementary materials, S2.

2.3.2. Perceived Posttraumatic Vulnerability Scale-other (PPVS-O)

Analogous in format to the PPVS-S, the PPVS-O assesses the degree to which individuals believe that trauma survivors are vulnerable to long-term negative emotional events. Participants are asked to imagine a hypothetical "average" person experiencing a traumatic event, and indicate their level of endorsement for each statement in reference to this person (e.g. *He/she would feel isolated and alone.*) on a 100-point scale (1 = *disagree*, 100 = *agree*). These responses are averaged for a composite score. Higher scores indicate stronger beliefs that trauma survivors will experience persistent and debilitating negative emotional effects. The PPVS-O displayed excellent internal consistency in our sample ($\alpha = 0.96$).

2.3.3. Implicit association test (IAT), vulnerability vs. resilience

The IAT (Greenwald, McGhee, & Schwartz, 1998) measures the degree to which a participant implicitly associates concepts with each other. Response latencies when sorting items between different categories formed by concept pairs determines which concepts are more strongly associated with each other (Greenwald, Poehlman, Uhlmann, & Banaji, 2009). The IAT has displayed good internal consistency and test-retest reliability (Bosson, Swann, & Pennebaker, 2000) and convergent validity in its agreement with measures of explicit preferences (Greenwald et al., 1998). Our IAT assessed strength of implicit association between the self (i.e., *me* versus *not me*) and vulnerability (i.e., *vulnerable – resilient)*. More positive *d*-scores on the IAT indicate a greater implicit association of the self with the resilient attribute versus the opposite configuration; more negative scores indicate a greater association of the self with the vulnerable attribute. The IAT had adequate reliability in our sample ($\alpha = 0.85$). While calculating these *d*-scores, we identified 25 participants whose response latencies were implausibly fast, signifying invalid responding. Indeed, several participants reported difficulties with the IAT arising from online connectivity glitches. We eliminated these 25 participants from analyses involving the IAT. Our IAT was created using, iatgen, an open-source IAT builder for online surveys, and was analyzed using the *iatgen* package for R (Carpenter et al., 2017).

2.3.4. Words-can-harm scale (WCHS)

The WCHS (see supplementary materials S1) is a 10-item scale that assesses the degree to which an individual believes that exposure to offensive words has the potential to cause serious harm to themselves or other people. Participants indicated their level of endorsement for each statement (e.g. *I could be traumatized without ever being touched, just through someone's hurtful words*) on a 100-point scale (1 = *disagree,* 100 = *agree*). Responses were averaged for a composite score. Higher scores indicate stronger beliefs that words can harm people. The WCHS displayed excellent internal consistency in our sample ($\alpha = 0.92$).

2.3.5. World assumptions scale (WAS)

The WAS (Janoff-Bulman, 1989) is a measure that assesses a participant's degree of belief in different underlying assumptions about the world and themselves. For the purposes of our experiment, we used only the 3 subscales from this measure that pertain to our proposed moderator, which deals with controlling and predicting stressful events (Controllability, Randomness, and Self-Controllability Subscale), a total of 12 items. Participants indicate their level of agreement with statements about their underlying assumptions (*Through our actions we can prevent bad things from happening to us*) on a scale from 1 (*strongly disagree*) to 7 (*strongly agree*). These subscales were averaged (with the Randomness Subscale reverse-scored) to create a composite score reflecting perceptions of the world's controllability and predictability. Higher scores indicate stronger beliefs that one's world is predictable and controllable. Our controllability/predictability scale had good

internal consistency (α = 0.83). The subscales of the WAS have satisfactory convergent validity in their sensitivity to whether individuals have experienced trauma (Janoff-Bulman, 1989), and correlate with posttraumatic symptom severity (Elklit, Shevlin, Solomon, & Dekel, 2007), indicating that different life events can affect global assumptions about the world.

2.3.6. Trigger warning attitudes assessment (TWAA)

The TWAA (see supplementary materials, S2) is a two-item scale that assesses attitudes toward trigger warnings. First, participants receive a short definition of trigger warnings, and are asked "Do you think that trigger warnings should be used?" If participants agree with this statement, they are then asked "Why do you think that trigger warnings should be used?" Participants view a list of potential reasons for trigger warning use (e.g. protection of vulnerable populations, fairness, psychological harm) and are asked to select all that apply. An "other" category is also provided, and participants can add reasons not listed.

2.3.7. Demographics questionnaire

This questionnaire asks for non-identifying information on participants' backgrounds. The questionnaire assesses gender, self-reported race and ethnicity, and age. Religiosity is assessed using a 5-point Likert scale (1 = *not religious*, 5 = *extremely religious*), as is political orientation (1 = *very liberal*, 5 = *very conservative*).

2.3.8. Psychiatric history questionnaire

This questionnaire asks "Have you ever been diagnosed with a psychiatric or psychological problem?" If participants answer *yes*, they are asked to choose diagnoses from a list, including an "other" option that allows them to add any not included on the questionnaire.

2.4. Procedure

After undergoing institutional review and receiving approval, our online experiment was posted as a Human Intelligence Task (HIT) on MTurk. The HIT description indicated that our survey involved reading and providing feedback on passages from literature. The consent form also mentioned that the readings would "cover a diverse range of emotional and dramatic content." After providing informed consent, participants were screened for exposure to traumatic events. We excluded those reporting trauma exposure, and randomly assigned the others to either the No Warning or Trigger Warning condition.

Participants in both conditions then read three mildly distressing passages in random order. After each passage, they used slider bar scales ranging from 0 (*not at all*) to 100 (*very much*) to rate their response on the following measures: *sad, happy, afraid, anxious, angry, content, disgusted,* degree of unpleasant emotion overall, and degree of anticipated long-term negative emotion. The target emotion was anxiety; the other items were fillers included to diminish demand effects. The average of these three passages' anxiety responses served as the baseline anxiety response for each participant.

Next, participants read another series of 10 passages in random order. Five were neutral, and the other five were markedly distressing. In the Trigger Warning condition, each of the markedly distressing passages was preceded by a trigger warning screen which had to be acknowledged by clicking a radio button, i.e. *TRIGGER WARNING: The passage you are about to read contains disturbing content and may trigger an anxiety response, especially in those who have a history of trauma*. (Although we screened out individuals who experienced events likely to constitute Criterion A traumas, we included the phrase concerning trauma victims because it unmistakably qualifies the statement as a trigger warning.) The No Warning condition participants viewed a screen that indicated they were about to view the next passage, which was also acknowledged by clicking a radio button. Participants rated the intensity of their reactions after each markedly distressing passage; the difference between the average of these anxiety ratings and the baseline average anxiety rating constituted the "immediate anxiety change" for each participant.

After completion of condition-specific passage presentations, participants read three more mildly distressing passages and rated the intensity of their reactions. The difference between the average of these anxiety responses and baseline anxiety responses constituted the "follow-up anxiety change" score for each participant. Next, participants completed measures presented in random order that tapped the outcomes of perceived vulnerability to posttraumatic symptoms for themselves (PPVS-S) and others (PPVS-O) and a measure of implicit self-identification as vulnerable versus resilient (IAT). Participants also completed measures that tapped constructs hypothesized to moderate the relationship between trigger warning presentation and immediate anxiety change (Words Can Harm Scale; WCHS) and trigger warning presentation and follow-up anxiety change (Controllability Scale derived from the World Assumptions Scale; WAS). Participants also responded to the demographics questionnaire, the psychiatric diagnosis history questionnaire, and the TWAA. Finally, participants were provided with a debriefing form which explained the deception implemented and purpose of the experiment. The debriefing form also informed participants that "the graphic nature of some of the passages may have caused you discomfort or emotional distress. Such feelings, although unpleasant, usually subside fairly quickly." The form provided information on available free resources for any persistent distress that participants might have experienced. Participants were each compensated $3.00 for completing the entire survey based on a projected completion time of no more than 1 h, a rate of payment consistent with ethical guidelines for crowdsourced remuneration (Chandler & Shapiro, 2016).

2.5. Planned analyses

Our sample size (N = 270) provided sufficient power (1 – β error probability = .96) to detect a small effect size (f^2 = 0.10) in our planned interaction analyses, which had the greatest number of possible predictors of all our analyses (maximum possible predictors = 9). We based our sample size requirements on a small effect, as there is no precedent for experiments involving trigger warnings. We first planned to examine demographic characteristics to determine whether participants had been effectively randomized to condition. As an exploratory analysis, we also planned to determine the reasons why participants might favor the use of trigger warnings. Next, we planned to conduct bivariate analyses between our demographic and outcome variables in order to determine which characteristics of the sample should be controlled for in the main analyses. For our main analyses, we planned to conduct multiple regressions to determine the effects of trigger warnings on each outcome variable while controlling for relevant demographic characteristics. Following the suggestion of an anonymous peer reviewer, we also conducted uncontrolled regression analyses in order to account for the possibility of statistical overcontrol (Meehl, 1971). We also planned follow-up analyses to examine our proposed moderated relationships between trigger warnings and anxiety changes using regression-based interaction detections and simple slopes analyses. See supplementary materials (S3) for all R code used in our analyses.

3. Results

3.1. Sample characteristics

The sample contained a majority of females (n = 156, 57.8%), and the mean age was 37 years old (SD = 12.4 years). Race was predominantly Caucasian (n = 191, 70.7%), with 9.6% African American (n = 26) and 9.3% Asian/Pacific Islander (n = 25) participants. Ethnicity was predominantly non-Hispanic (n = 250, 92.6%), and political orientation was predominantly at least "somewhat liberal" (n = 146, 54.0%). The majority of participants identified as at least

Table 1
Bivariate correlations between demographic variables and outcome variables (N = 270).

	IAC	FAC	PPVS-S	PPVS-O	IAT[e]
Gender[a,b]	−.10	−.01	−.20**	−.05	.11
Race[a,c]	−.03	−.01	.13*	.16**	−.02
Ethnicity[a,c]	−.02	−.04	.00	.07	−.06
Psychiatric Diagnostic Status[a,d]	−.06	−.10	.24***	.10	−.09
Religiosity	−.04	−.02	−.01	−.01	−.06
Political Orientation	−.03	−.01	−.14*	−.16**	.03
Age	−.08	.09	−.18**	−.20**	.07

Note. IAC = Immediate Anxiety Change, FAC = Follow-Up Anxiety Change, PPVS-S = Perceived Posttraumatic Vulnerability Scale – Self, PPVS-O = Perceived Posttraumatic Vulnerability Scale – Other, IAT = Vulnerable/Resilient IAT *d*-score.
***$p < .001$, **$p < .01$, *$p < .05$.
[a] Correlation coefficients depicted for these dichotomous variables are point-biserial correlations (r_{pb}).
[b] Dichotomized as *female* = 0, *male* = 1.
[c] Dichotomized as *non-minority* = 0, *minority* = 1.
[d] Dichotomized as *no diagnosis* = 0, *at least one diagnosis* = 1.
[e] $n = 245$ due to 25 IAT scores identified as invalid.

"somewhat religious" ($n = 156$, 57.8%). A total of 42 participants (15.6%) endorsed one or more psychiatric diagnoses other than PTSD. A majority of participants ($n = 216$, 80.0%) believed that trigger warnings should be used. Of these, a large majority based their belief on the need to protect psychologically vulnerable populations (such as those with PTSD) ($n = 192$, 88.9%), with roughly half believing that protection should be afforded to any minority group member ($n = 109$, 50.5%) or to people in general ($n = 112$, 51.9%). When measured as a continuous variable, political orientation differed by condition ($r_{pb} = .13$, $p < .05$), indicating that the Trigger Warning condition participants were slightly more conservative than those in the No Warning condition. Therefore, political orientation was included as a covariate in all regression analyses (both controlled and uncontrolled).

3.2. Bivariate associations

Bivariate associations between demographic characteristics and outcome variables appear in Table 1. Women, racial minorities, liberals, younger individuals, and those with at least one psychiatric diagnosis perceived themselves as more vulnerable to persistent negative emotional effects in the event of trauma than did men, Caucasians, conservatives, older participants, and those without a psychiatric diagnosis. Liberals, younger individuals, and racial minorities perceived trauma survivors in general as more vulnerable. Accordingly, we included these variables in those outcomes' controlled regression analyses.

3.3. Multiple regression analyses

Table 2 shows the results of the multiple regressions for each outcome variable, with relevant demographic characteristics entered as control variables and condition (Trigger Warning or No Warning) entered as predictors. Relative to participants who received no trigger warnings, those receiving them perceived themselves as more vulnerable to suffering persistent negative emotional effects in the event of experiencing trauma (i.e., a 5.2% increase in the strength of this belief, $B = 5.17$, t (263) = 2.12, $p < .05$). The results of the uncontrolled analysis for this outcome were similar; $B = 5.48$, t (263) = 2.13, $p < .05$. Relative to participants who received no warnings, those who received trigger warnings had stronger beliefs that trauma survivors would suffer persistent negative emotional effects, (i.e., a 5.4% increase in the strength of this belief, $B = 5.38$, t (265) = 2.35, $p < .05$). The results of the uncontrolled analysis for this outcome were similar; $B = 5.09$, t (263) = 2.19, $p < .05$. No significant effect of condition was found for participants' implicit identification of self with the attributes of vulnerable versus resilient. The groups with and without trigger warnings did not differ in their immediate anxiety change in response to "markedly distressing" content during the experimental paradigm. Similarly, exposure to trigger warnings did not display a global effect on participants' follow-up anxiety change in response to "mildly distressing" content.

Table 2
Multiple regression analyses of the effect of condition on outcome variables, controlling for relevant demographic characteristics (N = 270).

Predictor	Outcome Variable									
	PPVS-S		PPVS-O		IAT[a]		IAC		FAC	
	B	*SE B*	*B*	*SE B*	*B*	*SE B*	*B*	*SE B*	*B*	*SE B*
Condition[b]	5.16*	2.43	5.38*	2.29	.02	.05	.98	1.93	−3.59	2.01
R^2	.17**		.09**		.00		.00		.01	

Note. PPVS-S = Perceived Posttraumatic Vulnerability Scale – Self, PPVS-O = Perceived Posttraumatic Vulnerability Scale – Other, IAT = Vulnerable/Resilient IAT *d*-score, IAC = Immediate Anxiety Change, FAC = Follow-Up Anxiety Change.
**$p < .001$ *$p < .05$.
[a] $n = 245$ due to 25 IAT scores identified as invalid.
[b] Control predictors entered included gender (PPVS-S), race (PPVS-S, PPVS-O), psychiatric diagnostic status (PPVS-S), age (PPVS-S, PPVS-O), and political orientation (all outcomes).

3.4. Moderation analyses

For our moderation analyses, we entered political orientation, condition, moderator variable scores, and the cross-product of condition and the moderator variable scores as independent variables in a multiple regression predicting the outcome of interest. If the cross-product's coefficient was significant, we considered it support for the presence of an interaction. Table 3 shows the results of these interaction detections. For significant interactions, we conducted a simple slopes analysis that tested for the conditional effect of the predictor variable on the outcome variable at 1 *SD* above and below the mean of the moderator variable.

3.4.1. Words can harm belief

The analyses suggest that trigger warnings increase acute anxiety to the extent that participants believe that words can cause harm. A simple slopes analysis indicated that for participants who do not have a strong

Table 3
Interaction detection analyses for the prediction of anxiety change variables (N = 270).

Model 1: Immediate Anxiety Change	F (4, 265) = 4.67	R^2 = .07**
Variable	*B*	*SE*
Condition	−12.52*	5.07
WCHS Score	-.01	.06
Condition × WCHS Score	.24**	.09
Model 2: Follow-Up Anxiety Change	**F (4, 265) = 2.79**	**R^2 = .04***
Variable	*B*	*SE*
Condition	.27	11.27
Controllability Score	.90*	.41
Condition × Controllability Score	-.19	.63

Note. WCHS = Words Can Harm Scale. Political orientation was controlled for in both analyses.
**$p < .01$, *$p < .05$.

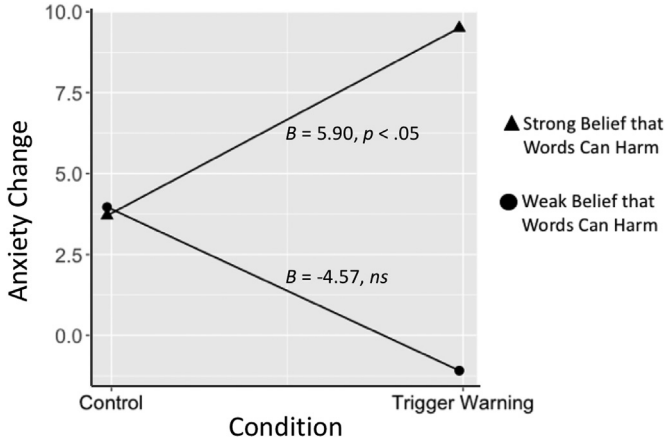

Fig. 1. Simple slopes of condition predicting change in immediate anxiety response from baseline at high (M + 1 SD) and low (M – 1 SD) values of the belief that words can harm.

belief that words can cause harm (M_{WCHS} – 1 SD), receiving a trigger warning does not significantly increase anxiety from baseline (B = −4.57, t (265) = −1.73, ns). However, if participants have a strong belief that words can harm (M_{WCHS} + 1 SD), trigger warnings significantly increase anxiety from baseline (B = 5.90, t (265) = 2.20, p < .05). Fig. 1 depicts this moderating function.

3.4.2. Assumptions of controllability

This analysis indicated that assumptions about the world's controllability do not change the relation between trigger warnings and subsequent anxiety change, as the cross-product term was non-significant. However, assumptions of controllability did display a small but significant main effect on follow-up anxiety change (B = .90, t (265) = 2.16, p < .05), such that higher controllability beliefs increased anxiety change by a very small percentage.

4. Discussion

This study is the first to examine the effects of trigger warnings on individual resilience factors via a randomized controlled experiment. Our results indicate that trigger warnings affect some specific domains of resilience relevant to trauma-naïve individuals, but seem to matter less for other domains. We will now address each of our questions and discuss implications for resilience to stress and trauma.

4.1. Perceived vulnerability of the self (Q1, Q2)

Trigger warnings increased people's perceived risk of suffering long-term debilitating emotional harm (such as PTSD) in the wake of a traumatic event (Q1). This effect, albeit small, is notable. Beliefs about the self are generally quite stable (Church et al., 2012); a significant change based on such a small manipulation is somewhat surprising. Trigger warnings may increase perceptions of self-vulnerability by sending an implicit message about the long-term harm caused by trauma; extensive exposure to trigger warnings may amplify this effect. This result has implications for resilience, as pathogenic appraisal of one's emotional reactions to stressors increases risk for PTSD (Dunmore et al., 2001; Ehring et al., 2006). Importantly, the effect of trigger warnings on perceptions of vulnerability appear to apply only to explicit beliefs regarding resilience to traumatic events; trigger warnings did not significantly affect implicit identification of the self as resilient versus vulnerable (Q2).

4.2. Perceived posttraumatic vulnerability of others (Q3)

Our results also indicate that trigger warnings enforce a "soft stigma" concerning trauma survivors, implying their inability to function as other people can. This effect was also small, but may be additive over the long term. This finding suggests trigger warnings may have unintended yet potentially deleterious consequences for those they aim to protect.

4.3. Anxiety response to potentially distressing material (Q4)

Trigger warnings did not affect anxiety responses to potentially distressing material in general. However, trigger warnings may foster a self-fulfilling prophecy (Merton, 1948) that increases anxiety for those individuals who believe that words can harm them. Trigger warnings themselves do not appear to generate the belief that words can harm, as the strength of this belief was not significantly related to condition. Rather, trigger warnings may confirm this belief in those who already harbor it. Hence, such warnings may increase acute anxiety by fostering an expectancy of harm (Barsky et al., 2002; Reiss & McNally, 1985).

4.4. Subsequent anxiety response to less distressing material (Q5)

Trigger warnings did not affect reactivity to mildly distressing material viewed without a warning, indicating that their anxiogenic effects are limited to immediate reactions for a specific subset of people. Additionally, assumptions that one's world is controllable and predictable do not appear to affect this relationship, failing to support the notion that trigger warnings exacerbate an expectancy of predictability that sensitizes people to less severe unexpected stressors (the "coddling" hypothesis; Lukianoff & Haidt, 2015).

4.5. Relationships between resilience factors and demographic variables

To some extent, our outcome variables correlated in expected directions with demographic variables. For example, perceived vulnerability to posttraumatic impairment is associated with demographic factors associated with greater risk of PTSD among people exposed to trauma, such as being female (Tolin & Foa, 2006) and having pre-existing psychiatric disorders (Breslau, Davis, Andreski, & Peterson, 1991). Additionally, racial minority status and younger age were associated with higher levels of perceived vulnerability. In this case, it is possible that these variables are proxies for more meaningful third variables. For example, the relation between age and perceived risk for impairment could signify a cohort effect; younger participants may perceive greater risk may because of their upbringing within an especially protective cultural moment (Lukianoff & Haidt, 2015).

4.6. Trigger warning attitudes

A large majority of participants supported the use of trigger warnings, independently of whether they were randomized to the trigger warning condition or to the control condition. A considerable proportion of these participants believed that trigger warnings are needed not only by those with PTSD and other psychological vulnerabilities, but also by minority groups and people in general. These results suggest that trigger warnings are viewed by many people as applicable to a much broader range of concerns than those of accommodating people with PTSD, as others have noted (Boysen, 2017; Lukianoff & Haidt, 2015).

4.7. Limitations and future directions

One limitation of our study is that we had to devise novel and hitherto untested measures to assess perceived vulnerability to posttraumatic impairment (PPVS-S, PPVS-O), belief that words can harm

(WCHS), and attitudes about trigger warnings (TWAA). However, their internal consistencies were high, and several correlated in expected directions with demographic variables. Another limitation was the use of self-report. As with any online crowdsourced study, the validity of responses can be difficult to determine, but our use of content-based attention checks mitigates the effects of this limitation.

Do our findings merely reflect demand effects? Perhaps participants in the trigger warning condition reported themselves and others as vulnerable to posttraumatic impairment after exposure to the explicit warnings embodied in trigger warnings. Perhaps they merely tried to satisfy presumptive experimental expectations rather than conveying their actual *beliefs* about posttraumatic vulnerability. To guard against such demand effects, we did not directly ask participants whether they thought they would develop PTSD following trauma. Rather, we asked them to imagine themselves surviving an attempted murder (the words "trauma" or "PTSD" were not mentioned) and then asked them to rate the likely severity of *specific symptoms* they believed they might experience thereafter. Additionally, trigger warnings did not have a main effect on immediate anxiety responses to passages, but rather only displayed an effect for those who strongly believed that words can harm. This result indicates that trigger warnings are achieving their effects by exacerbating specific iatrogenic *beliefs* about the likelihood of harm rather than by activating participants' desire to be good research subjects.

This study used the written word as stimuli, rather than in vivo stress inductions or the use of film or images. The use of literary passages as stimuli is a strength of this study, as the written word is ubiquitous in educational settings – the center of the trigger warning debate. However, the effortful engagement required in order for the written word to induce an emotional response may have limited the size of our effects when compared to the use of more vivid media. Future research should examine whether the effects generalize to other types of stimuli.

Our study used participants from a crowdsourcing website, as we were primarily interested in the effects of trigger warnings in the general population. It is unclear whether our findings generalize to an exclusively collegiate population. However, the MTurk population is more demographically diverse than the typical undergraduate population (Chandler & Shapiro, 2016; Mason & Suri, 2012), and hence our findings may generalize broadly to American society.

Our study emphasized pre-traumatic resilience, and it remains unclear whether our results pertain to traumatized individuals. Nevertheless, Bruce's (2017a) research indicates that some of our effects may apply to a population with PTSD. She found that physiological markers of anxiety were heightened after the presentation of a trigger warning when compared to "PG-13" and "no warning" conditions, and that this effect was significantly larger for those with more severe PTSD symptoms. More broadly, concerns about trigger warnings as they apply to trauma-naïve individuals are different from the resilience factors at issue in trauma survivors, such as iatrogenically encouraging avoidance of trauma-related cues, and reinforcing the centrality of trauma to individuals' identities (McNally, 2014).

Research to date has lent some plausibility to such concerns, showing significant positive cross-sectional associations between amount of trigger warning use and trauma centrality (Bruce, 2017a, b), and between degree of trigger warning use and avoidance behavior (Bruce, 2017b). Researchers should address trauma survivor-specific concerns about trigger warnings with experimental tests to clarify these issues.

5. Conclusion

Taken together, our findings provide a preliminary look at the effects of trigger warnings on pre-traumatic resilience variables as they apply to the general population, and a step forward in answering the question of whether trigger warnings help or harm. Trigger warnings do not appear to be conducive to resilience as measured by any of our metrics. Rather, our findings indicate that trigger warnings may present nuanced threats to selective domains of psychological resilience. Such consequences are limited to perceived vulnerability to emotional harm, which may increase risk for developing PTSD in the event of trauma, and disability-related stigma around trauma survivors. However, this effect does not apply to implicit self-identification regarding vulnerability. Trigger warnings do not appear to affect sensitivity to distressing material in general, but may increase immediate anxiety response for a subset of individuals whose beliefs predispose them to such a response. These findings do not form the basis for immediate policy changes regarding the use of trigger warnings without subsequent replication, as effect sizes were small. This may partly be due to our use of literary passages that require effortful engagement to induce emotional response, and to our use of a non-traumatized sample. Finally, although many of our subjects were of the same cohort as college students, replication in an entirely collegiate sample is warranted, as this population is especially likely to experience exposure to trigger warnings.

Declaration of interest and funding

None of the authors has conflicts of interest to report. This work was supported by funds from the Elsie Hopestill Stimson Memorial Fund granted to the first author by the Department of Psychology at Harvard University.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jbtep.2018.07.002.

References

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Barsky, A. J., Saintfort, R., Rogers, M. P., & Borus, J. F. (2002). Nonspecific medication side effects and the nocebo phenomenon. *The Journal of the American Medical Association, 287*, 622–627. https://doi.org/10.1001/jama.287.5.622.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis, 20*(3), 351–368. https://doi.org/10.1093/pan/mpr057.

Berntsen, D., & Rubin, D. C. (2007). When a trauma becomes a key to identity: Enhanced integration of trauma memories predicts posttraumatic stress disorder symptoms. *Applied Cognitive Psychology, 21*, 417–431. https://doi.org/10.1002/acp.1290.

Boelen, P. (2012). A prospective examination of the association between the centrality of a loss and post-loss psychopathology. *Journal of Affective Disorders, 137*, 117–124. https://doi.org/10.1016/j.jad.2011.12.004.

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology, 79*, 79–101.

Boysen, G. A. (2017). Evidence-based answers to questions about trigger warnings for clinically-based distress: A review for teachers. *Scholarship of Teaching and Learning in Psychology, 3*(2), 163–177. https://doi.org/10.1037/stl0000084.

Breslau, N., Davis, G. C., Andreski, P., & Peterson, E. (1991). Traumatic events and posttraumatic stress disorder in an urban population of young adults. *Archives of General Psychiatry, 48*, 216–222. https://doi.org/10.1001/archpsyc.1991.01810270028003.

Breslau, N., & Kessler, R. C. (2001). The stressor criterion in DSM-IV posttraumatic stress disorder: An empirical investigation. *Biological Psychiatry, 50*(9), 699–704. https://doi.org/10.1016/S0006-3223(01)01167-2.

Bruce, M. J. (2017b). Predictors of trigger warning use: Avoidance or asserting accommodation needs? *Poster presented at the annual meeting of the international society of traumatic stress studies, Chicago, IL*.

Bruce, M. J. (2017a). Does trauma centrality predict trigger warning use? *Physiological responses to using a trigger warning. Poster presented at the annual meeting of the Midwest Psychological Association.* Chicago, IL: Declaration of Interest and Funding.

Carpenter, T., Pogacar, R., Pullig, C., Kouril, M., LaBouff, J., Aguilar, S., et al. (2017, August 7). *Building and analyzing implicit association tests for online surveys: A tutorial and open-source tool.* Retrieved from: osf.io/preprints/psyarxiv/6xdyj.

Carter, A. M. (2015). Teaching with trauma: Trigger warnings, feminism, and disability pedagogy. *Disability Studies Quarterly, 35*, 9.

Chandler, J., & Shapiro, D. N. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology, 12*, 53–81.

Church, T. A., Alvarez, J. M., Katigbak, M. S., Mastor, K. A., Cabrera, H. F., Tanaka-Matsumi, J., et al. (2012). Self-concept consistency and short-term stability in eight

cultures. *Journal of Research in Personality, 46*(5), 556–570.
Dunmore, E., Clark, D. M., & Ehlers, A. (2001). A prospective investigation of the role of cognitive factors in persistent posttraumatic stress disorder (PTSD) after physical or sexual assault. *Behaviour Research and Therapy, 39*, 1063–1084. https://doi.org/10.1016/S0005-7967(00)00088-7.
Ehring, T., Ehlers, A., & Glucksman, E. (2006). Contribution of cognitive factors to the prediction of post-traumatic stress disorder, phobia, and depression after motor vehicle accidents. *Behaviour Research and Therapy, 44*, 1699–1716. https://doi.org/10.1016/j.brat.2005.11.013.
Elklit, A., Shevlin, M., Solomon, Z., & Dekel, R. (2007). Factor structure and concurrent validity of the world Assumptions scale. *Journal of Traumatic Stress, 20*, 291–301. https://doi.org/10.1002/jts.20203.
Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480.
Greenwald, A., Poehlman, T., Uhlmann, E., & Banaji, M. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*(1), 17–41.
Grupe, D. W., & Nitschke, J. B. (2013). Uncertainty and anticipation in anxiety: An integrated neurobiological and psychological perspective. *Nature Reviews Neuroscience, 14*, 488–501. https://doi.org/10.1038/nrn3524.
Institute of Medicine of the National Academies (2008). *Treatment of posttraumatic stress disorder: An assessment of the evidence.* Washington, DC: The National Academies Press.
Janoff-Bulman, R. (1989). Assumptive worlds and the stress of traumatic events: Applications of the schema construct. *Social Cognition, 7*, 113–136.
Lukianoff, G., & Haidt, J. (2015, September). *The coddling of the American mind.* The Atlantic. Retrieved from: https://www.theatlantic.com/magazine/archive/2015/09/the-coddling-of-the-american-mind/399356/.
Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's mechanical turk. *Behavior Research Methods, 44*, 1–23. https://doi-org.ezp-prod1.hul.harvard.edu/10.3758/s13428-011-0124-6.
McNally, R. J. (2014, May 20). *Hazards ahead: The problem with trigger warnings, according to the research.* Pacific Standard. Retrieved from: https://psmag.com/education/hazards-ahead-problem-trigger-warnings-according-research-81946.
McNally, R. J. (2016, September 13). *If you need a trigger warning, you need P.T.S.D. treatment.* New York Times. Retrieved from: https://www.nytimes.com/roomfordebate/2016/09/13/do-trigger-warnings-work/if-you-need-a-trigger-warning-you-need-ptsd-treatment.
Meehl, P. E. (1971). High school yearbooks: A reply to Schwarz. *Journal of Abnormal Psychology, 77*, 143–148. https://doi.org/10.1037/h0030750.
Merton, R. K. (1948). The self-fulfilling prophecy. *Antioch Review, 8*, 1930210.
Mineka, S., & Kihlstrom, J. (1978). Unpredictable and uncontrollable events: A new perspective on experimental neurosis. *Journal of Abnormal Psychology, 87*, 256–271. https://doi.org/10.1037/0021-843X.87.2.256.
Reiss, S., & McNally, R. J. (1985). Expectancy model of fear. In S. Reiss, & R. R. Bootzin (Eds.). *Theoretical issues in behavior therapy* (pp. 107–121). San Diego, CA: Academic Press.
Robinaugh, D. J., & McNally, R. J. (2011). Trauma centrality and PTSD symptom severity in adult survivors of childhood sexual abuse. *Journal of Traumatic Stress, 24*(4), 483–486. https://doi.org/10.1002/jts.20656.
Rosenthal, M. Z., Hall, M. L., Palm, K. M., Batten, S. V., & Follette, V. M. (2005). Chronic avoidance helps explain the relationship between severity of childhood sexual abuse and psychological distress. *Journal of Child Sexual Abuse, 14*, 25–41. https://doi.org/10.1300/J070v14n04_02.
Rothbaum, B. O., Foa, E. B., Riggs, D. S., Murdock, T., & Walsh, W. (1992). A prospective examination of post-traumatic stress disorder in rape victims. *Journal of Traumatic Stress, 5*(3), 455–475. https://doi.org/10.1002/jts.2490050309.
Stokes, M. (2014, May 29). *In defense of trigger warnings.* The Chronicle of Higher Education. Retrieved from: http://chronicle.com/blogs/conversation/2014/05/29/in-defense-of-trigger-warnings/.
Telch, M. J., Harrington, P. J., Smits, J. A., & Powers, M. B. (2011). Unexpected arousal, anxiety sensitivity, and their interaction on CO2-induced panic: Further evidence for the context-sensitivity vulnerability model. *Journal of Anxiety Disorders, 25*, 645–653. https://doi.org/10.1016/j.janxdis.2011.02.005.
Thompson, S. C. (1981). Will it hurt less if I can control it? A complex answer to a simple question. *Psychological Bulletin, 90*, 89–101. https://doi.org/10.1037/0033-2909.90.1.89.
Tolin, D. F., & Foa, E. B. (2006). Sex differences in trauma and posttraumatic stress disorder: A quantitative review of 25 years of research. *Psychological Bulletin, 132*, 959–992. https://doi.org/10.1037/0033-2909.132.6.959.
Wilson, R. (2015, September 14). *Students' requests for trigger warnings grow more varied.* The Chronicle of Higher Education. Retrieved from: http://chronicle.com/article/Students-Requests-for/233043.
Wyatt, W. (2016). The ethics of trigger warnings. *Teaching Ethics, 16*, 17–35. https://doi.org/10.5840/tej201632427.