Journal of Experimental Social Psychology 79 (2018) 252–263

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

Trigger warnings as an interpersonal emotion-regulation tool: Avoidance, attention, and affect depend on beliefs☆

Izzy Gainsburg*, Allison Earl*

Department of Psychology, University of Michigan, United States of America

A B S T R A C T

Trigger warnings-warnings of anticipated negative affect in response to distressing content-are increasingly common and debated, but no empirical research has tested their effects on anticipated affect, emotion-regulation behavior, or experienced affect. The present research explores trigger warnings as an interpersonal emotion-regulation strategy, introducing a temporal dimension to interpersonal emotion-regulation by regulating others' future, anticipated emotions. In a descriptive survey, Study 1 demonstrated that anticipated anxiety for warned-of content predicts intentions to avoid information. Furthermore, beliefs about trigger warnings as protective (versus coddling) best predicted anticipated anxiety for warned-of content and subsequent intentions to avoid. In Study 2, participants had higher anticipated negative affect for videos with trigger warnings, compared to those without, and this mediated increased avoidance for warned-of videos. In Study 3, trigger warnings preceding essays increased anticipated negative affect and attentional-regulation strategies, but reduced experiences of negative affect. Across studies, believing that trigger warnings are protective (versus coddling) increased their effect on anticipated negative affect, but weakened their effect on experienced negative affect. Implications for policy and future research are discussed.

Although most emotion-regulation literature has focused on the regulation of one's own emotions, recent research has increasingly explored interpersonal emotion-regulation (Rimé, 2007; Zaki & Williams, 2013). Strategies to regulate others' emotions include empathic concern (i.e., matching one's emotional state to that of others; Batson, 2017), coregulation (i.e., dynamic, bidirectional regulation of affective state in dyads; Butler & Randall, 2013), and social support (i.e., providing resources to aid others in emotion-regulation; Cohen & Wills, 1985). The above strategies focus on regulating another person's current emotions but do not address regulating another person's anticipated emotional experiences. The present research addresses this gap by exploring how *trigger warnings*—warnings of anticipated negative affect in response to distressing content—influence anticipated affect, emotion-regulation, and experienced affect.

1. Background on trigger warnings

Trigger warnings are statements that warn of a negative emotional response to potentially distressing subject matter. An example of a trigger warning is: "TRIGGER WARNING: This article or section, or pages it links to, contains information about sexual assault and/or violence which may be triggering to survivors" ("Trigger Warning," n.d.-b). Trigger warnings are sometimes intended to help people with posttraumatic stress disorder (PTSD) avoid severe distress from "triggering" content, but are often intended for broader populations (Boysen, 2017). As such, trigger warnings are increasingly prevalent in universities and news media (Kamenetz, 2016; Schmidt, 2015). Debate has arisen about whether trigger warnings cause avoidance or facilitate emotion-regulation, inviting commentary from prominent psychologists (Lukianoff & Haidt, 2015; McNally, 2014). However, no empirical research exists on trigger warnings to date. Furthermore, research has yet to identify stimuli that are widely triggering for those with or without PTSD, making it unclear whether warnings of potentially triggering stimuli would effectively facilitate emotion-regulation.

2. Why trigger warnings might be different

Because the word "trigger" in "trigger warnings" refers to a severe, emotional reaction associated with PTSD, the phrase "trigger warning" itself calls attention to a potentially negative emotional experience (Boysen, 2017). Separate literatures in psychology have explored warnings of television violence, physical danger, and counter-attitudinal appeals, however, trigger warnings are unique in that they explicitly warn of negative affect, whether or not they contain specific information about the content (Bushman & Stack, 1996; Cox et al., 1997; Wood & Quinn, 2003). In other words, the primary feature that

☆ This research was supported in part by the Department of Psychology and Rackham Graduate School at the University of Michigan.

Special thanks to Ethan Kross and the Emotion and Self-Control Lab and the HAILab at the University of Michigan for feedback and Amena Khan for help with developing materials.

* Corresponding authors at: Department of Psychology, 530 Church Street, University of Michigan, Ann Arbor, MI 48109, United States of America.

E-mail addresses: izzyg@umich.edu (I. Gainsburg), anearl@umich.edu (A. Earl).

https://doi.org/10.1016/j.jesp.2018.08.006

Received 31 January 2018; Received in revised form 25 July 2018; Accepted 10 August 2018

distinguishes trigger warnings from other types of warnings is that trigger warnings specifically signal a potentially negative affective response to subsequent content. For instance, even warnings of television violence that caution of potential "harmful effects" or that "viewer discretion is advised" are broad and do not explicitly call attention to a potential negative emotional experience (Bushman & Stack, 1996). These more common warnings found in the media that might seem similar (e.g., movie ratings, explicit language labels on music albums) are also different from trigger warnings because a) they are sometimes used by adults to protect children, b) they often accompany content intended for mere entertainment, whose content has been normalized, c) such warnings have also been normalized, likely resulting in greater habituation (Rooke, Malouff, & Copeland, 2012), and d) these warnings do not caution, specifically, of a negative emotional response to warned-of content.[1]

3. Consequences for anticipated emotions

Because trigger warnings explicitly warn of a potentially negative emotional experience, they should generally increase people's anticipated negative affect for engaging with warned-of content. Still, whether recipients of trigger warnings actually anticipate having negative emotional responses to warned-of content is an empirical question. Furthermore, whether trigger warnings affect people's anticipated emotions may depend on characteristics of the individual or the relationship of the warned-of content to the individual. Although a number of factors could moderate the effect of trigger warnings on anticipated negative affect, the present research focuses on the influence of people's beliefs about whether trigger warnings are protective or coddling. These beliefs about whether triggers warnings are protective or coddling capture the core of the polarizing discussion around their utility (e.g., Bass & Clark, 2015; Lukianoff & Haidt, 2015; McNally, 2014; Stokes, 2014). In other words, these particular beliefs seem to be prevalent, strongly held, and varied across the population, making them of practical interest as a moderator.

The belief that trigger warnings are protective (versus coddling) should result in the warning having a stronger effect on anticipated negative affect. If one generally believes that trigger warnings are protective from a credible harm (e.g., negative emotional experience), then belief-consistent expectations would result in warning-induced increases in anticipated negative affect. This rationale is consistent with cognitive models of threat-perception, in which valid cues to potential threat have a greater effect on people's expectations of threat and subsequent action tendencies than invalid cues (van Rooijen, Ploeger, & Kret, 2017), as well as appraisal theory, where beliefs about whether a stimulus is a valid threat moderate threat expectations (Lazarus, 1991). That beliefs about trigger warnings could shape anticipated emotions for warned-of content would be consistent with other psychological research that has focused on the power of beliefs and mindsets to guide expectations (Crum, Salovey, & Achor, 2013; Dweck, 2012; Oyserman, 2015). Taken together, the protective value of trigger warnings is a divisive belief that underlies the debate around their use, and theoretically, should influence how people expect to feel in response to warned-of content.

4. Consequences for emotion-regulation

Although most emotion-regulation research focuses on regulating current emotions (Gross, 1998), some research has focused on how people cope with anticipated negative emotions (Aspinwall, 2011). In the cognitive control literature, for instance, *proactive control* is the activation and maintenance of goal-relevant information that prepares attentional, perceptual, and action systems for anticipated cognitively demanding tasks (Braver, 2012). Trigger warnings, then, might activate emotion-regulation strategies in preparation for distressing content if they increase expectations of negative affect. The present research will focus on whether trigger warnings elicit two broad strategies in emotion-regulation: *avoidance* and *monitoring*.

Research on *information avoidance* suggests that people often avoid information that they expect to induce negative affect (Earl, Crause, Vaid, & Albarracín, 2016; Earl & Hall, in press; Earl & Nisson, 2015; Earl, Nisson, & Albarracín, 2015; Howell & Shepperd, 2012; Sweeny, Melnyk, Miller, & Shepperd, 2010). Several emotion-regulation strategies also involve avoidance, such as *situation selection* (i.e., removing the self from situations that elicit negative emotions) and *attentional deployment* (i.e., directing attention away from stimuli that elicit negative emotions; Gross, 1998; Aspinwall, 2011; Wadlinger & Isaacowitz, 2011; Richards, Benson, Donnelly, & Hadwin, 2014). Finally, hedonic motives can result in preferences for avoidance of attitude-inconsistent in favor of attitude-consistent information (Earl & Hall, in press; Earl & Nisson, 2015). In sum, the above literatures propose that if trigger warnings elicit expectations of negative emotions, people might avoid warned-of content.

Avoidance, however, is not always feasible (e.g., leaving a play when anticipating an unpleasant scene) or desired (e.g., wanting to see the rest of the play). In such cases, people often monitor their environments for potential threat and engage in a reactive emotion-regulation strategy if necessary (Miller, 1987). Thus, by cueing a potential threat, trigger warnings might prompt monitoring for the cued threat when people engage with the warned-of content. Trigger warnings might also ironically elicit approach if they activate conflicting motivations such as reactance-driven approach, which has been observed in response to other warnings such as those of television violence (Bushman & Stack, 1996); smoking hazards (Erceg-Hurn & Steed, 2011); fattening foods (Bushman, 1998); and persuasive appeals (Wood & Quinn, 2003). Taken together, even if trigger warnings elicit expectations of negative affect, there are contexts where they might elicit monitoring or approach, rather than avoidance.

5. Consequences for experienced emotions

Trigger warnings are intended to reduce experiences of negative emotion, but whether or not trigger warnings do so is an untested, empirical question. On the one hand, trigger warnings should decrease experiences of negative affect if they effectively facilitate emotion-regulation. In addition, trigger warnings might reduce people's experiences of negative affect by prompting mental contrasts of one's current affective response to a more severe imagined alternative (Geers & Lassiter, 1999). Similarly, negative affective experiences are perceived as more severe when they are unexpected, such as receiving a bad exam grade (Sweeny & Shepperd, 2010); losing money in a gamble (*decision affect theory*; Mellers, Schwartz, & Ritov, 1999), or tasting low-quality wine (Gneezy, Gneezy, & Lauga, 2014). On the other hand, expectations of negative affect can bias perceptions in line with expectations, such as in perceptions of pain (Berkowitz & Thome, 1987; Leventhal, Brown, Shacham, & Engquist, 1979; Price, Finniss, & Benedetti, 2008) and affective responses to comics and videos (Wilson, Lisle, Kraft, &

[1] We acknowledge that there can be overlaps between trigger warnings and other types of warnings. Like other warnings, trigger warnings can also warn of specific content. Furthermore, sometimes a trigger warning's caution of potential distress is implied by the content or the context in which the warning is issued, rather than being explicitly stated (for instance, some warnings that many would qualify as trigger warnings do not use the specific phrase "trigger warning"). For this reason, some warnings in entertainment contexts could reasonably be considered as trigger warnings, although this would depend on how the warning is interpreted based on the specific warning and the context in which it is issued. In the present research, we explore the effects of "trigger warnings" specifically, for two reasons: 1) to be as precise as possible in testing the warnings that are being debated, and 2) to test the effects of warnings that explicitly warn of negative affect (a defining feature of trigger warnings), which has not been captured in previous literature.

Table 1
Demographic information from Studies 1–3.

Demographic variable	Study 1 (N = 80)		Study 2 (N = 276)		Study 3 (N = 979)	
	N	Percentage	N	Percentage	N	Percentage
Race						
White	57	71.25%	200	72.46%	582	59.45%
African American	1	1.25%	25	9.06%	54	5.52%
Hispanic/Latino	5	6.25%	8	2.90%	41	4.19%
Asian American	2	2.50%	18	6.52%	33	3.37%
Other or multiracial	1	1.25%	15	5.43%	34	3.47%
Missing/no response	14	17.50%	10	3.62%	235	24.00%
Gender						
Male	34	42.50%	129	46.74%	270	27.58%
Female	30	37.50%	133	48.19%	472	48.21%
Other	1	1.25%	2	0.72%	9	0.92%
Missing/no response	15	18.75%	12	4.35%	228	23.29%
Education level						
< High school	0	0.00%	0	0.00%	3	0.31%
High school/GED	18	22.50%	68	24.64%	211	21.55%
Associate's degree	15	18.75%	28	10.14%	105	10.73%
Bachelor's degree	22	27.50%	121	43.84%	294	30.03%
Master's & professional degree	10	12.50%	45	16.30%	123	12.56%
Other	0	0.00%	3	1.09%	14	1.43%
Missing/no response	15	18.75%	11	3.99%	229	23.39%

Wetzel, 1989).

Moreover, beliefs about whether trigger warnings are protective (versus coddling), may moderate their effect on experience affect. As previously mentioned, these beliefs should moderate warning-induced expectations of negative affect, which can guide how warned-of content influences experienced affect in the ways discussed above. Furthermore, these beliefs also might also moderate the direct relationship between exposure to warnings and experienced negative affect via motivated reasoning (Kunda, 1990). For instance, people who believe trigger warnings are protective might modulate their affective responses in ways that allow them to maintain their beliefs that trigger warnings are protective. This could either result in decreased experience of negative emotion, which would support the belief that the trigger warning was helpful, but it could also result in increased experience of negative emotion, which would support the belief that the trigger warning was a valid cue to threatening content. Thus, although beliefs about trigger warnings might moderate their effect on emotional responses to warned-of content, it is unclear exactly how this might unfold.

6. The present research

In a descriptive survey about trigger warnings, Study 1 first aimed to conceptually replicate previous research on information avoidance by showing that anticipated negative emotions for warned-of content predicted intentions to avoid. More importantly, we tested whether the divergent beliefs that characterize the debate around trigger warnings matter for how people respond to trigger warnings, and whether these specific, divergent beliefs about trigger warnings mattered more than other polarizing beliefs (e.g., political orientation). Specifically, we predicted that the belief that trigger warnings are protective (versus coddling) would increase participant's anticipated negative emotions for engaging with warned-of content, mediating increased intentions to avoid warned-of content.

Study 2 tested the effects of trigger warnings preceding videos on anticipated affect for videos and video selection. We predicted participants would expect to be feel more negatively while watching warned-of videos (compared to videos without warnings), and that this difference would be greater for participants who believe trigger warnings to be protective (versus coddling). We also predicted that increased expectations of negative affect for warned-of content would mediate increased avoidance of warned-of videos.

Study 3 tested the effects of trigger warnings preceding an essay on anticipated affective response to the essay, avoidant and monitoring attention during the essay, and experienced affect while reading the essay. In addition to the control condition, two trigger warning conditions were used-one general warning of negative affect ("Trigger Warning Only") and one warning of the specific content ("Trigger Warning with Content")-to explore whether the generality or specificity of the trigger warning mattered for the dependent variables of interest. We predicted that trigger warnings would elicit greater expectations of negative affect for the essay and increased avoidant and monitoring attention, and that these attentional strategies would be mediated by greater expectations of negative affect. Finally, we predicted that participants who receive trigger warnings would feel less negatively than participants who do not receive trigger warnings, and that this would be mediated by increases in anticipated negative affect and attention-regulation strategies. Study 3 also tested whether beliefs that trigger warnings are protective (versus coddling) would moderate the direct effect of trigger warnings on experienced negative affect, but no directional hypotheses were made.

Across studies, we report all measures, manipulations, and exclusions in the main text. In addition, although trigger warnings are sometimes intended for those with PTSD, they are often intended for and deployed in the context of broader populations (Boysen, 2017). Thus, the present research focuses on the effects of trigger warnings in the general population, arguing that future research should carefully consider their effects for those with PTSD.

7. Study 1

7.1. Method

7.1.1. Participants

We recruited 120 individuals from across the United States through Amazon's Mechanical Turk (MTurk), but an additional 28 additional participants signed up despite not finishing the survey. Of these 148 individuals, 4 were excluded on account of English being their second language and 64 participants dropped out of the survey after informed consent but prior to answering any questions, resulting in a final sample of 80 participants. Sample size was decided a priori, with no collection after data analysis, but was done so without formal power analysis due to the descriptive and exploratory nature of the survey. See Table 1 for demographic details for all studies.

7.1.2. Procedure

Participants were told that the survey examined "experiences, beliefs, and feelings about trigger warnings." Participants then freely responded to the question, "What is your definition of a Trigger Warning?" Participants were then provided with the following definition of trigger warnings: "A statement at the start of a piece of writing, video, etc., alerting the reader or viewer to the fact that it contains potentially distressing material (often used to introduce a description of such content)" ("Trigger Warning," n.d.-a). Participants then answered questions about their anticipated emotional and behavioral responses for warned-of content, their beliefs and attitudes about trigger warnings, experiences with trigger warnings, broader beliefs they hold about the world, generalized anxiety, coping style, and demographics. Items that are not directly related to the present research questions are available in the supplemental materials.

7.1.3. Measures

7.1.3.1. Anticipated emotions for warned-of content. Participants answered, "How do you expect to feel when encountering content that was preceded by a trigger warning?" for 10 emotions (anxious, sad, angry, nervous, happy, surprised, scared, confused, apathetic, worried) on a 9-point scale (1 = "Not at all" to 9 = "Extremely").

7.1.3.2. Anticipated avoidance. Participants answered, "Does seeing a trigger warning make you more likely to avoid the warned-of material?" on a 7-point scale (e.g., 1 = "Less likely to avoid" to 7 = "More likely to avoid").

7.1.3.3. Beliefs about trigger warnings as protective (versus coddling). Participants rated agreement with two statements, "Trigger warnings that precede distressing content 'coddle' people, hurting them in the long run," and "Trigger warnings that precede distressing content 'protect' people, helping them in the long run" on 7-point scales (1 = "Strongly disagree" to 7 = "Strongly agree"). These items were highly correlated, $r(65) = -0.83$, $p < .001$. A two-item composite was created by reverse scoring the "coddling" item and averaging the two items.

7.1.3.4. Trigger warning attitudes. Participants responded to the prompt, "In your opinion, trigger warnings preceding some piece of writing, video, etc., are" for five different 7-point scales (1 = Bad to 7 = Good, 1 = Negative to 7 = Positive, 1 = Harmful to 7 = Beneficial, 1 = Foolish to 7 = Wise, 1 = Unnecessary to 7 = Necessary). There was excellent reliability among the five items (Cronbach's $\alpha = 0.95$), and a composite was made from averaging participant responses across the five items.

7.1.3.5. Political correctness attitudes. Participants answered, "How positively or negatively do you feel about a culture that values 'political correctness'?" on a 7-point scale (1 = Very Negative to 7 = Very Conservative).

7.1.3.6. Political attitudes. Participants answered, "What best describes your political orientation?" on a 7-point scale (1 = Liberal to 7 = Conservative).

7.2. Results

7.2.1. Anticipated emotions for warned-of content

First, we tested which emotions participants anticipated feeling while engaging with warned-of content. An exploratory factor analysis, using principal axis factoring and promax rotation, tested the relationships between all the items asking participants about their anticipated emotions for warned-of content. Two factors with eigenvalues > 1 were extracted. The first component (eigenvalue = 5.92) was defined by nine of the ten emotions: Nervous (0.90), Worried (0.88), Scared (0.84), Anxious (0.81), Confused (0.79), Sad (0.74), Angry (0.71), Surprised (0.65), Happy (0.64). This factor, comprised of both positive and negative emotions, appears to be a "general affect" factor. The second component (eigenvalue = 1.35) was defined by two emotions: Happy (0.49) and Confused (0.47). It is less clear as to what construct underlies this component.

Because the first factor was too broad to use as a diagnostic tool for predicting avoidance and the second factor was weak and difficult to interpret, we looked to see which specific emotions participants expected to feel the most when engaging with warned-of content. A repeated measures ANOVA revealed differences among these emotions, $F(9, 576) = 11.01$, $p < .001$, partial $\eta^2 = 0.15$. Pairwise comparisons show that participants reported anticipated anxiety ($M = 3.95$, 95% CI [3.32, 4.58]) more than any other emotion except nervousness ($M = 3.65$, 95% CI [3.05, 4.25]; $M_{\text{difference}} = 0.31$, 95% CI [−0.02, 0.64], $p = .07$, $d = 0.21$). Thus, anticipated anxiety was used as the primary anticipated emotion, and was subsequently analyzed with regards to avoidance.

Table 2
Correlates of anticipated anxiety for warned-of content.

	TW protective beliefs	TW attitudes	Political correctness attitudes	Political orientation
Anticipated anxiety	0.27*	0.11	0.21	−0.02

Values are bivariate correlations. * indicates correlation is significant at $p < .05$ (two-tailed).

7.2.2. Anticipated avoidance

Next, we tested whether anticipated anxiety for warned-of content predicted intentions to avoid warned-of content. A linear regression showed that increased anticipated anxiety for warned of content predicted increased intentions to avoid warned-of material ($\beta = 0.41$, $t(68) = 3.64$, $p < .001$).

7.2.3. Predicting anticipated anxiety

To understand what individual-level factors predict anticipated anxiety for warned-of content, bivariate correlates of anticipated anxiety were explored. Warning-specific beliefs and attitudes were included, as well as broader beliefs that might underlie these warning-specific beliefs (e.g., political orientation). Only beliefs about trigger warnings as protective (versus coddling) emerged as a significant correlate of anticipated anxiety ($r = 0.27$, $p = .03$). See Table 2.[2]

7.2.3.1. Model of intended avoidant emotion-regulation. Model 4 in PROCESS (Hayes, 2012) revealed an indirect effect of beliefs about trigger warnings as protective (versus coddling) on intended avoidance through increased anticipated anxiety for warned-of content, but not a direct effect (indirect effect: b = 0.08, 95% CI [0.0118, 0.1884]; total effect: b = 0.17, 95% CI [−0.0183, 0.3495]). Specifically, stronger beliefs that trigger warnings are protective predicted increased anticipated anxiety for interacting with warned-of material, which in turn increased intentions to avoid warned-of content (see Fig. 1).

7.3. Discussion

In Study 1, participants reported anticipated anxiety for engaging with warned-of content above and beyond other anticipated emotions

[2] A full correlation matrix between all items in Study 1 is available in the supplement, which allows interested readers to see the correlations between the beliefs about trigger warnings and other associated constructs.

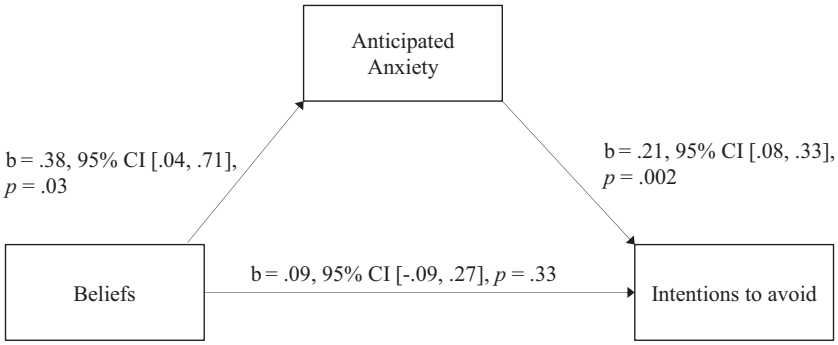

Fig. 1. Beliefs about trigger warnings increase intentions to avoid warned-of content through increased anticipated anxiety for warned-of content.

(except nervousness), and anticipated anxiety predicted avoidance of warned-of content. These findings suggest that trigger warnings influence anticipated affect, and that this anticipated affect might have consequences for emotion-regulation behavior (e.g., avoidance). Critically, participants who believed that trigger warnings were protective (versus coddling) were more likely to anticipate feeling anxious for engaging with warned-of content. Other predictors, such as attitudes about trigger warnings, political orientation, and attitudes toward political correctness, did not predict anticipated anxiety for warned-of content. A mediation model suggested that those who believed trigger warnings to be protective (versus coddling) reported higher expectations of feeling anxious around warned-of material, mediating increased intentions to avoid. Although the bivariate relationship between protective beliefs and intentions to avoid was marginal ($r = 0.22$, $p = .08$), this could be due to insufficient power, or due to the hypothetical nature of recalling warnings and avoidance in this study, rather than exposure to actual trigger warnings and subsequent content. This model is conceptually tested in Studies 2 and 3, where participants actually see trigger warnings and warned-of material.

8. Study 2

8.1. Method

8.1.1. Participants

We recruited 240 individuals from across the United States through MTurk. Only 200 participants were needed to detect a small effect ($d = 0.2$) given the predicted interaction in the present design (two-level within subjects factor by a continuous moderator with a predicted r of 0.5 between the repeated measures, Power = 0.8, and $\alpha = 0.05$), but we oversampled this target to account for potential warning-related attrition. An additional 46 participants provided usable data despite not finishing the survey and 10 participants were excluded on the basis of not selecting a video to watch, resulting in a final sample of 276 participants and the ability to detect an effect size of $d = 0.17$. Sample size was determined a priori and analyses were conducted only after the data collection ceased.

8.1.2. Procedure

Participants were told the study was about "people's thoughts and feelings around videos that are available on the internet." Participants were shown two fictional video titles and asked to choose one to watch. The titles were chosen from a pilot test of eight original fictional titles among 30 participants to determine two titles similar across 14 dimensions known to influence approach and avoidance (e.g., interesting; Earl et al., 2009; Hart et al., 2009; Sweeny et al., 2010). See supplementary materials for details.

For all participants, one of the videos had the following trigger warning (the video with the trigger warning was counterbalanced across participants): "Trigger Warning: This video contains distressing content." Participants then rated how they would expect to feel while watching each video using the Self-Assessment Manikin (SAM; Bradley & Lang, 1994), which allows for quick measurement of the valence (e.g., Please rate how you would expect to feel while watching Video 1 from very negative (1) to very positive (9) as shown on the graphic below) and arousal (e.g., "Please rate how you would expect to feel while watching Video 1 from not at all intense (1) to extremely intense (9) as shown on the graphic below") components of affect. Finally, participants reported reasons for their choice; their attitudes and beliefs about trigger warnings (same items as Study 1), and demographics. Participants never watched the videos.

8.2. Results

8.2.1. Anticipated negative affect

Study 1 demonstrated that trigger warnings were most associated with expectations of feelings anxious. In circumplex models of emotion, anxiety is negative-valence and high-arousal (Bradley & Lang, 1994; Larsen & Diener, 1992). Thus, to capture anxiety on the SAM, valence was zero-centered and weighted by intensity, such that scores above 0 were increasingly negative and intense, and numbers below zero were increasingly positive and intense (hereafter referred to as *anticipated negative affect*). A paired samples t-test shows that participants reported significantly higher levels of anticipated negative affect for videos with trigger warnings ($M_{TW} = 10.51$, 95% CI [8.92, 12.11]) than for videos without trigger warnings ($M_{NoTW} = 6.26$, 95% CI [5.05, 7.47], $t(269) = 6.18$, $p < .001$, $d = 0.38$). Furthermore, a mixed-model 2 (video-type, within-subjects) x Continuous (beliefs, between-subjects) ANOVA demonstrated that the difference between videos with and without warnings for anticipated negative affect was moderated by beliefs about whether trigger warnings are protective (versus coddling), $F(1, 263) = 12.45$, $p < .001$, partial $\eta^2 = 0.05$. Specifically, participants who believed trigger warnings to be protective reported larger differences in anticipated affect between videos with and without warnings ($M_{\text{difference}+1\ SD} = 6.55$, 95% CI [4.66, 8.43]) than those who believed trigger warnings to be coddling ($M_{\text{difference-1 SD}} = 1.77$, 95% CI [−0.11, 3.65]). See Fig. 2a.

8.2.2. Selective exposure

A binomial test revealed that participants chose videos without trigger warnings (0.56, binomial 95% CI [0.50, 0.62]) more than videos with trigger warnings (0.44, binomial 95% CI [0.38, 0.50]), although this difference was only marginal ($p = .06$) using a two-tailed significance test. Logistic regression showed that people avoided warned-of videos more when they believed that trigger warnings are protective (versus coddling) (Wald = 7.25, $p = .007$, Exp(B) = 0.71) and when they anticipated feeling worse while watching warned-of videos relative to videos without warnings (Wald = 5.89, $p = .02$, Exp(B) = 0.73). Model 4 in PROCESS (Hayes, 2012) revealed a direct effect of beliefs about trigger warnings on avoidance (b = −0.20, 95% CI

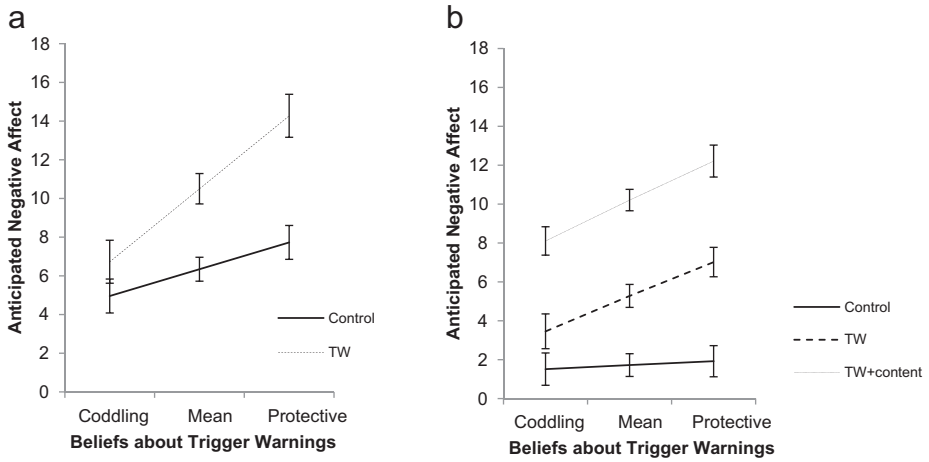

Fig. 2. a. Effects of Trigger Warnings on anticipated negative affect in Study 2, with beliefs plotted at ± 1 Standard Deviation. Error bars represent ± 1 Standard Error.
b. Effects of Trigger Warnings on anticipated negative affect in Study 3, with beliefs plotted at ± 1 Standard Deviation. Error bars represent ± 1 Standard Error.

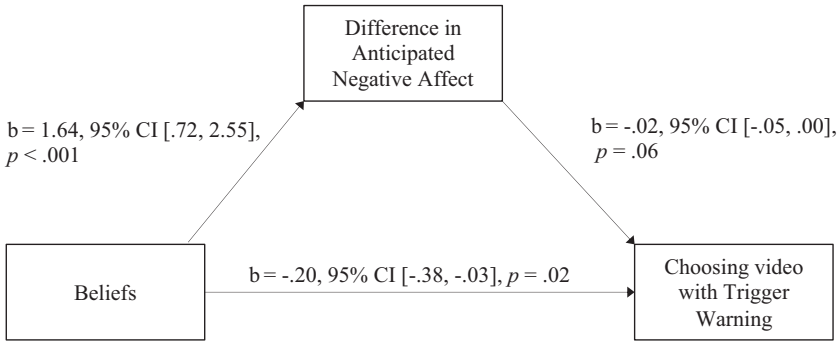

Fig. 3. Beliefs about trigger warnings increase avoidance of videos with trigger warnings through increased anticipated negative affect.

[−0.3797, −0.0289]) as well as an indirect effect of beliefs on avoidance through increased anticipated negative affect for videos with trigger warnings compared to those without (indirect effect: b = −0.04, 95% CI [−0.1003, −0.0025]). See Fig. 3.

8.3. Discussion

Participants anticipated that videos with trigger warnings would elicit higher levels of negative affect than those without. Furthermore, this effect was moderated by beliefs about trigger warnings, such that participants who believed that trigger warnings are protective (versus coddling) reported greater differences in anticipated negative affect between videos with trigger warnings and those without. Videos with trigger warnings were more often avoided than those without, although this effect was marginal using a two-tailed significance test ($p = .06$). Participants who believed trigger warnings are protective (versus coddling) and were also more likely to avoid warned-of videos, and this avoidance was mediated by elevated anticipated negative affect for warned-of videos. Study 2 provides the first evidence that trigger warnings increase people's expectations of feeling negatively and subsequent avoidance, and that these effects vary by people's beliefs about trigger warnings. Because participants did not actually engage with the warned-of content, it remains unclear how trigger warnings affect attention and experienced affect in response to warned-of content. Study 3 addresses these questions.

9. Study 3

9.1. Method

9.1.1. Participants

We recruited 720 individuals through MTurk. The smallest effect size we were aiming to conceptually replicate (differences in anticipated negative affect between content with and without warnings; $d = 0.38$) required 110 participants per cell in a between-subjects design. We recruited 120 participants per cell to account for potential warning-related attrition, resulting in an intended sample of 720 participants given the 3 × 2 between-subjects design. An additional 290 additional people signed up despite not finishing the survey. Of these 1010 participants, 23 non-native English-speakers and 8 people that did not consent were excluded, leaving 979 participants, which was large enough to detect an effect size of $d = 0.31$ for a between-subjects comparison of a control condition and a trigger warning condition given Power = 0.8 and $\alpha = 0.05$. All analyses were conducted only after the data collection ceased.

9.1.2. Materials

Participants read an essay involving domestic violence developed by the authors and a research assistant. Participants read one of two versions of the essay, which were identical excepting three moments that varied in severity (see supplement for full essay text). Two versions of the essay were used to test whether effects of the trigger warnings

would generalize to content of different intensity that could credibly have trigger warnings. We did not use content so mild such that trigger warnings would not provide any emotional benefit (i.e., there are no negative emotions to be reduced), or content that is so severe that it would violate ethical standards for the present research context. Given that the present research does not focus on these extremes, it was predicted that there would be a main effect of severity on experienced negative affect without any interactions with the warning conditions.

In addition, this present study had a control condition and two different trigger warning conditions. One trigger warning condition merely warned that content might be distressing ("Trigger Warning: This article contains distressing content," i.e., Trigger Warning Only) and the other alerted participants to the specific subject matter that might be distressing ("Trigger Warning: This article contains content around domestic violence," i.e., Trigger Warning with Content). The Trigger Warning Only condition, like the control condition, made no mention of specific subject matter, thus making it a more internally valid manipulation. On the other hand, the Trigger Warning with Content condition was higher in external validity, as real-life trigger warnings often occur in the context of recipients being aware of the potentially distressing subject matter.

Furthermore, there is reason to think the two warning conditions could have different effects on anticipated negative affect, attention-regulation, and experienced negative affect. For anticipated negative affect, the Trigger Warning with Content might have a stronger effect due to the warned-of content being relatively severe (domestic violence). On the other hand, it is also possible that the Trigger Warning Only condition will have a stronger effect, given previous research suggesting that more uncertain, general threats produce greater anxiety than specific, predictable threats (Grupe & Nitschke, 2013). Differences in anticipated negative affect between these two conditions would also result in subsequent differences in attention-regulation that follow from anticipated negative affect. For experienced negative affect, the Trigger Warning with Content could result in lower levels of experienced negative affect given prior research suggesting that anticipating specific threats allows for better emotion-regulation than anticipating more general threats (Grupe & Nitschke, 2013); on the other hand, if the Trigger Warning Only elicits weaker anticipated negative affect, it could result in weaker expectancy effects and thus lower levels of experienced negative affect. Study 3 was designed to test these competing hypotheses.

9.1.3. Procedure

After providing informed consent, participants were randomly assigned to one of three conditions: One where participants saw only the essay's title; one where they saw the title and a sentence below that read "Trigger Warning: This article contains distressing content;" and one where they saw the title and a sentence below that read "Trigger Warning: This article contains content around domestic violence." Next, participants reported how they anticipated feeling while reading using the Self-Assessment Manikin. Participants were then randomly assigned to read the "less severe" or "more severe" essay. After reading, participants reported their experienced affect while reading using the SAM; their thoughts about the essay (Cacioppo, Glass, & Merluzzi, 1979); their memory for the essay's content; a ten-item scale of self-reported attention; their attitudes, beliefs, and experiences around trigger warning (same items as in Study 1 and Study 2); their coping style; and demographics. Because the present report focused on anticipated and experienced affect, as well as self-regulation, analyses of the thought-listing and memory measures are not reported (although these measures are available in the supplement).

9.2. Results

9.2.1. Anticipated negative affect

We first tested the effects of trigger warnings on anticipated negative affect. Based on the results of Study 1 and Study 2, we predicted participants in the trigger warning conditions would report higher anticipated negative affect than those in the control condition. In line with these predictions, a one-way ANOVA revealed that participants who saw trigger warnings anticipated greater negative affect, $F(2, 943) = 47.46$, $p < .001$, $\eta^2 = 0.09$. Participants reported the least anticipated negative affect in the control condition followed by the Trigger Warning Only condition and the Trigger Warning with Content condition. Planned contrasts revealed that the difference between the control condition ($M_{control} = 1.75$, 95% CI [0.99, 2.51]) and the Trigger Warning Only condition ($M_{TW} = 4.79$, 95% CI [3.74, 5.84]) was significant, $t(943) = 4.06$, $p < .001$, $d = 0.13$. The difference between the Trigger Warning Only and the Trigger Warning with Content condition ($M_{TW+cont} = 9.00$; 95% CI [7.74, 10.26]) was also significant, $t(943) = 5.60$, $p < .001$, $d = 0.18$.

We next tested whether the effect of trigger warnings on anticipated negative affect was moderated by beliefs about trigger warnings as protective (versus coddling). Also replicating the previous studies, a Beliefs × Trigger Warning interaction revealed that the effect of trigger warnings on anticipated negative affect was stronger for who participants believed trigger warnings to be protective (versus coddling), $F(2, 763) = 3.70$, $p = .03$, partial $\eta^2 = 0.10$. See Fig. 2b.

9.2.2. Attention-regulation

To test whether trigger warnings would directly impact attention to information, we asked participants to rate their agreement with 10 statements about attention during the essay on a 9-point scale. A principal component analysis revealed three factors: feeling distracted (6 items; e.g., "I felt distracted as I read,"), monitoring attention (2 items; "I actively monitored the essay for distressing content,"), and avoidant attention (1 item; "I avoided reading the distressing content."). In this context, distracted attention is being conceptualized less as attention-regulation and more as a byproduct of reading a long essay in an uncontrolled environment. Thus, it was of less theoretical interest, and is presented in the supplementary materials. Participants who read faster than 700 words per minute were excluded (Just & Carpenter, 1987), resulting in 115 exclusions (no difference by essay or warning condition; all $ps > 0.12$, all Exp(B) between 0.58 and 1.54). Sixty-four additional participants dropped out before the essay appeared (no difference by condition, all $ps > 0.37$, all Exp(B) between 0.74 and 0.88).

9.2.2.1. Avoidant attention. One way of coping with distressing content is to avoid attending to it altogether. To test whether trigger warnings prompted avoidant attention, a 3 (warning) ×2 (essay-severity) ANOVA for avoidant attention revealed a main effect of trigger warning type ($F(2, 702) = 4.67$, $p = .01$, partial $\eta^2 = 0.01$), but no main effect of essay-severity ($F(1, 702) = 1.63$, $p = .20$, partial $\eta^2 = 0.002$) or interaction ($F(2, 702) = 0.56$, $p = .57$, partial $\eta^2 = 0.002$). Simple effects reveal no difference in avoidant attention between the control condition and the Trigger Warning Only condition ($M_{difference} = 0.06$, 95% CI [−0.18, 0.29], $p = .65$, $d = 0.04$); the Trigger Warning with Content condition elicited more avoidant attention than both the control condition ($M_{difference} = 0.33$, 95% CI [0.10, 0.56], $p = .01$, $d = 0.26$) and the Trigger Warning Only condition ($M_{difference} = 0.28$, 95% CI [0.05, 0.51], $p = .02$, $d = 0.22$).

9.2.2.2. Monitoring attention. Another way of coping with distressing content is to monitor information closely for any threatening content. To test whether trigger warnings elicited monitoring attention, a 3 (warning) ×2 (essay-severity) ANOVA for monitoring attention revealed a main effect of trigger warning type ($F(2, 702) = 15.81$, $p < .001$, partial $\eta^2 = 0.02$), no main effect of essay-severity ($p = .20$), and an interaction ($F(2, 702) = 4.16$, $p = .02$, partial $\eta^2 = 0.01$). Simple main effects of warning condition reveal a difference in monitoring attention between the control condition and the Trigger Warning Only condition ($M_{difference} = 0.88$, 95% CI [0.44, 1.32], $p < .001$, $d = 0.37$) as well as

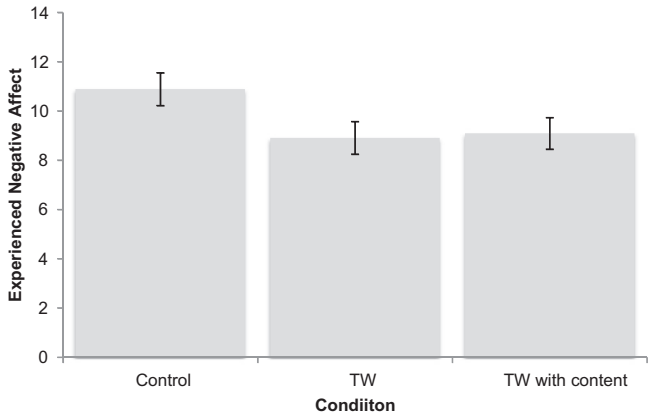

Fig. 4. Main effects of trigger warnings on experienced negative affect in Study 3. Error bars represent ± 1 Standard Error.

between the control condition and the Trigger Warning with Content condition ($M_{difference}$ = 1.20, 95% CI [0.77, 1.62], $p < .001$, $d = 0.50$); there was no difference in monitoring attention between the two warning conditions ($M_{difference}$ = 0.32, 95% CI [−0.11, 0.75], $p = .15$, $d = 0.13$). There were no a priori predictions for the Warning × Essay interaction, and thus, this interaction is only explored in the supplement.

9.2.3. Experienced Affect

Perhaps the primary reason people use trigger warnings is to reduce experiences of negative affect for those who engage with the warned-of content. To test whether trigger warnings were effective in reducing experiences of negative affect for participants, we ran a 3 (warning) ×2 (essay-severity) ANOVA for experienced negative affect (experienced negative affect calculated using the same composite of affective valence and intensity that was used to calculate anticipated negative affect). This test revealed a significant main effect of essay severity, (F(1, 793) = 108.17, $p < .001$, partial $\eta^2 = 0.12$); as predicted, the more severe essay elicited more negative affect (M = 13.58, 95% CI [12.50, 14.67]) than the less severe essay (M = 5.68, 95% CI [4.65, 6.70]). There was a only a marginally significant main effect of trigger warning type (F(2, 793) = 2.73, $p = .07$, partial $\eta^2 = 0.01$). Collapsed across essay type, simple effects reveal that participants in the control condition experienced significantly more negative affect as compared to those in either the Trigger Warning Only condition ($M_{difference}$ = 1.98, 95% CI [0.133, 3.83], $p = .04$, $d = 0.17$) or the Trigger Warning with Content condition ($M_{difference}$ = 1.80, 95% CI [−0.02, 3.62], $p = .05$, $d = 0.16$); there was no difference in experienced negative affect between the two trigger warning conditions ($M_{difference}$ = 0.18, 95% CI [−1.63, 2.00], $p = .85$, $d = 0.02$) (see Fig. 4). The two-way interaction between warning-type and essay severity was not significant (F(2, 793) = 1.19, $p = .30$, partial $\eta^2 = 0.00$).

9.2.3.1. Moderation by beliefs. Although there were no directional hypotheses, there were reasons to believe that the effect of trigger warnings on experienced negative affect could be moderated by beliefs about trigger warnings as protective (versus coddling). A 3 (warning) ×2 (essay-severity) x Continuous (beliefs) ANOVA for experienced negative affect again revealed a significant main effect of essay severity, (F(1, 702) = 72.08, $p < .001$, partial $\eta^2 = 0.10$). There was no main effect of trigger warning type, (F(2, 702) = 2.25, $p = .11$, partial $\eta^2 = 0.01$), warning x essay-severity interaction (F(2, 702) = 1.32, $p = .27$, partial $\eta^2 = 0.004$), or essay-severity x beliefs interaction (F(1, 702) = 0.08, $p = .78$, partial $\eta^2 = 0.00$). The warning x essay-severity x beliefs interaction was marginally significant, F(2, 702) = 2.62, $p = .07$, partial $\eta^2 = 0.01$ and is decomposed in the supplement for interested readers.

There was, however, a significant warning × beliefs interaction (F(2, 702) = 4.54, $p = .01$, partial $\eta^2 = 0.013$) (See Fig. 5). Simple effects reveal that those who believed trigger warnings to be coddling (one standard deviation below the mean) experienced more negative affect in the control condition compared to both the Trigger Warning Only condition ($M_{difference}$ = 4.69, 95% CI [1.74 7.65], $p = .002$, $d = 0.30$) and the Trigger Warning with Content condition ($M_{difference}$ = 4.15, 95% CI [1.45, 6.85], $p = .003$, $d = 0.28$). There was no difference in experienced negative affect between the two trigger warning conditions for these participants ($M_{difference}$ = 0.54, 95% CI [−2.28, 3.62], $p = .71$, $d = 0.03$). For those participants who believed trigger warnings to be of average protectiveness or relatively more protective (i.e., those at the mean and one standard deviation above the mean), there were no differences between trigger warning conditions and control conditions (for those at the mean, all $ps > 0.08$ for simple effects; for those one standard deviation above the mean, all $ps > 0.17$ for simple effects).

9.2.4. Attentional-regulation path analysis

We employed path analysis to test whether trigger warnings' effect on experienced negative affect was mediated by increases in anticipated negative affect and subsequent attention regulation strategies, such as

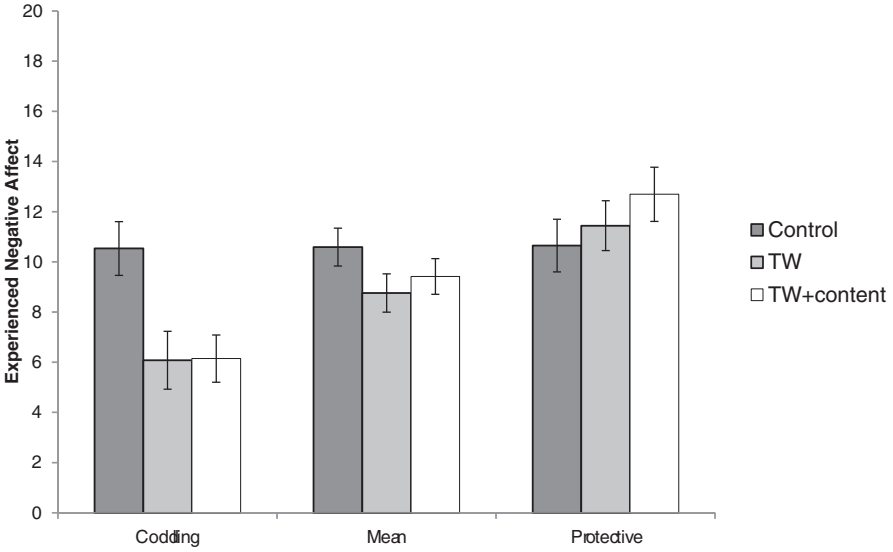

Fig. 5. Trigger warnings x Beliefs interaction on experienced negative affect in Study 3. Error bars represent ± 1 Standard Error.

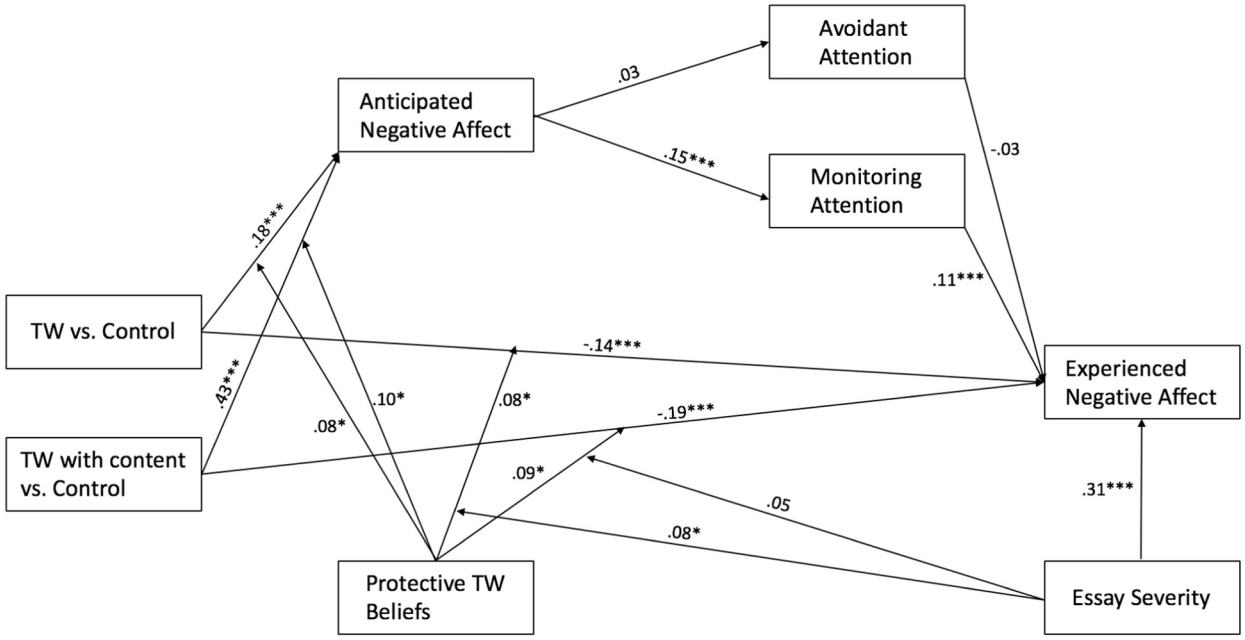

Fig. 6. Path analysis testing effects of warnings on experienced affect through anticipated affect and attention regulation. Only indirect and direct pathways of interest are drawn. See table for all modeled pathways. * indicates $p \leq .05$, ** indicates $p \leq .01$, *** indicates $p \leq .001$. $\chi^2 = 36.29$, $p = .07$; RMSEA = 0.03, 90% CI [0.00, 0.04]; CFI = 0.97; TLI = 0.93; SRMR = 0.02.

monitoring attention or avoidant attention (See Fig. 6 and Table 3). In specifying the model, two dummy variables were created and centered to test the effects of each trigger warning condition compared to the control condition. In addition, beliefs about trigger warnings were included as a moderator for the effect of condition on anticipated negative affect. Finally, beliefs about trigger warnings, essay severity, and their collective interaction with warning condition were included as factors predicting experienced affect.

9.2.4.1. Model fit. The model was tested using the lavaan package (Rosseel, 2012) for R statistical software. Model fit and parameter estimates between the variables were obtained after 1000 bootstraps. The model had a good fit by various fit indices (Barrett, 2007).

9.2.4.2. Indirect effects. Indirect effects were calculated according to the methods outlined in Hayes (2012). The model indicated indirect effects of both trigger warning conditions on experienced negative affect through anticipated negative affect and monitoring attention (Indirect effect for Dummy 1 through monitoring attention = 0.17, 95% CI [0.05, 0.36]; Indirect effect for Dummy 2 through monitoring attention = 0.07, 95% CI [0.02, 0.156]). Both of these indirect effects were positive, such that warnings elicited increases in anticipated negative affect, which were associated with subsequent increases in monitoring attention, which predicted subsequent increases in experienced negative affect. There were no indirect effects of either trigger warning condition on experienced negative affect through anticipated negative affect and avoidant attention, as both indirect effects include zero in their 95% confidence intervals (Indirect effect for Dummy 1 through avoidant attention = −0.004, 95% CI [−0.06, 0.01]; Indirect effect for Dummy 2 through avoidant attention = −0.01, 95% CI [−0.13, 0.02]).

9.2.4.3. Moderated mediation. To test whether the hypothesized mediated pathways through monitoring and avoidant attention were moderated by beliefs about trigger warnings, indexes of moderated mediation were calculated according to the methods outlined in Hayes (2012). This revealed that the indirect effects of both dummy variables on experienced negative affect through anticipated negative affect and monitoring attention were moderated by beliefs about trigger warnings (Index of Moderated Mediation for Dummy 1 through monitoring attention = 0.02, 95% CI [0.004, 0.0571]; Index of Moderated Mediation for Dummy 2 through monitoring attention = 0.003, 95% CI [0.005, 0.074]). Specifically, participants who believed trigger warnings to be protective (versus coddling) showed larger, positive indirect effects of trigger warnings on experienced negative affect through anticipated negative affect and monitoring attention. There was not significant moderated mediation of trigger warnings on experienced negative affect through anticipated negative affect and avoidant attention, as both indexes of moderated mediation include zero in their 95% confidence intervals.

9.3. Discussion

In Study 3, participants read essays that either had no warning, a trigger warning of distressing content, or a trigger warning of domestic violence; in addition, half of the essays were mildly distressing and half were moderately distressing. Participants in both warning conditions reported lower experiences of negative affect while reading compared to participants in the control condition, and this effect was not moderated by essay severity. There were also no differences between the two warning conditions, suggesting that general warnings of potential distress and content-specific warnings of distress can be equally effective in reducing experiences of negative affect. Critically, trigger warnings' effects on experienced negative affect were moderated by beliefs, such that decreases in negative affect due to trigger warnings were experienced most by those who believe trigger warnings to be coddling. Among participants who believed them to be protective there were no differences in experienced negative affect between trigger warning conditions and control conditions.

Conceptually replicating Study 1 and 2, both trigger warning conditions increased expectations of negative affect, although the Trigger Warning with Content had a stronger effect on increased expectations of negative affect than the Trigger Warning Only condition. One possible reason for this effect is that the specific content label that we used ("domestic violence") has the potential to be particularly severe, resulting in higher anticipated negative affect, as opposed to this effect being about the specificity of content itself. As was the case in the

Table 3
Path analysis testing effects of warnings on experienced affect through anticipated affect and attention regulation.

Lefthand side variable	Righthand side variable	Beta	b	95% CI (lower)	95% CI (upper)	*p*
Anticipated negative affect	TW vs. Control (D1)	0.18	3.56	2.24	4.96	< 0.001
	TW with Content vs. Control (D2)	0.43	8.49	6.98	10.02	< 0.001
	Beliefs	0.14	0.89	0.45	1.30	< 0.001
	D1 * Beliefs	0.08	1.05	0.09	2.02	0.03
	D2 * Beliefs	0.10	1.24	0.18	2.30	0.02
Avoidant attention	TW vs. Control (D1)	0.01	0.03	−0.15	0.24	0.75
	TW with Content vs. Control (D2)	0.11	0.28	0.04	0.55	0.03
	Anticipated Negative Affect	0.03	0.00	−0.01	0.02	0.57
Monitoring attention	TW vs. Control (D1)	0.14	0.76	0.33	1.27	< 0.01
	TW with Content vs. Control (D2)	0.18	0.89	0.43	1.35	< 0.001
	Anticipated Negative Affect	0.15	0.04	0.02	0.06	< 0.001
Experienced negative affect	TW vs. Control (D1)	−0.14	−3.56	−5.42	−1.31	< 0.001
	TW with Content vs. Control (D2)	−0.19	−4.44	−6.47	−2.17	< 0.001
	Essay Severity	0.31	7.25	5.50	8.59	< 0.001
	Beliefs	0.07	0.55	0.07	1.16	0.05
	D1 * Essay Severity	−0.01	−0.57	−5.00	3.59	0.80
	D2 * Essay Severity	0.05	2.42	−1.80	6.54	0.25
	D1 * Beliefs	0.08	1.30	−0.06	2.57	0.05
	D2 * Beliefs	0.09	1.44	0.14	2.89	0.03
	Essay Severity * Beliefs	0.00	−0.07	−1.08	1.06	0.91
	D1 * Essay Severity * Beliefs	0.08	2.66	−0.19	5.12	0.05
	D2 * Essay Severity * Beliefs	0.05	1.57	−1.28	4.09	0.25
	Avoidant attention	−0.03	−0.28	−0.91	0.40	0.39
	Monitoring attention	0.11	0.53	0.21	0.85	< 0.01
	Anticipated negative affect	0.24	0.29	0.19	0.41	< 0.001

previous studies, the effects of trigger warnings on anticipated negative affect were stronger for participants who believed trigger warnings to be protective (versus coddling). Monitoring attention was higher in both trigger warning conditions relative to the control condition, which was mediated by increased anticipated negative affect. Only the Trigger Warning with Content significantly increased avoidant attention, and this was not mediated by increased expectations of feeling negatively.

Path analysis revealed that decreases in negative affect were not mediated by avoidant or monitoring attention. Furthermore, path analyses indicated that there were indirect effects through anticipated negative affect and monitoring attention in the opposite direction of the direct effects, and that this indirect pathway was moderated by beliefs such that the indirect effect was stronger for those who believed trigger warnings to be protective (versus coddling). In other words, although trigger warnings had a total effect such that they decreased experienced negative affect, there was a simultaneous indirect effect that mitigated their wellbeing benefits, primarily due to expectations of negative affect and subsequent monitoring for distressing content.

10. General discussion

Across three studies, trigger warnings increased expectations of negative affective response to warned-of content, elicited avoidance of warned-of content, yet also decreased negative affect for warned-of content. These findings contribute to a growing literature on interpersonal emotion-regulation by demonstrating the effectiveness of a novel interpersonal emotion-regulation strategy that focuses on regulating others' future emotional experiences. These findings also contribute to the ongoing debate around the effects of trigger warnings on avoidance and experienced emotions.

In Study 1, a descriptive survey, participants indicated that trigger warnings increase anticipated anxiety for warned-of content; this anticipated anxiety was stronger for those who believed trigger warnings to be protective (versus coddling), and it mediated intentions to avoid warned-of material. In Study 2, when participants chose to watch one of two videos—one of which had a trigger warning-they reported higher anticipated negative affect for warned-of videos and were more likely to avoid them. Furthermore, participants who believed trigger warnings to be protective (versus coddling) were especially likely to anticipate negative affect, mediating their increased avoidance of warned-of videos. In Study 3, trigger warnings preceding an essay increased anticipated negative affect for the essay, increased monitoring and avoidant attention while reading, and decreased experiences of negative affect while reading. As in the previous studies, trigger warnings' effect on anticipated negative affect was stronger for participants who believed trigger warnings to be protective (versus coddling). However, such beliefs also moderated trigger warnings' effect on experienced negative affect, such that only those who believed trigger warnings to be coddling showed decreases in negative affect due to trigger warnings.

Even though the current paper addresses some of the most compelling issues in the debate around trigger warnings, important questions remain. Future research should test the generalizability of trigger warnings' effects on different populations (e.g., people with PTSD), with different materials (e.g., different subject matter), in non-laboratory contexts (e.g., in school), and with non-self-report measures (e.g., psychophysiology). In addition, more research is needed to understand the effects of general warnings and specific warnings-in the present research, the two warnings were equally effective in reducing experienced negative affect, but the Trigger Warning with Content condition produced more anticipated negative affect and avoidance than the Trigger Warning Only condition.

Future research should also explore why trigger warnings decrease experienced negative affect for warned-of content, and how trigger warnings affect specific emotion-regulation strategies. In the present research, although trigger warnings decreased negative affect, these decrements are not explained by the attention-regulation strategies measured. In contrast, monitoring attention predicted *increased* negative affect. More research is needed, however, to understand the dynamic and temporal nature of the attention-regulation strategies cued by trigger warnings, and their subsequent effects on experienced affect (Richards et al., 2014). Furthermore, future research should explore other emotion-regulation strategies (e.g., reappraisal), as well as other mechanisms such as mental contrasting and expectation violations, through which trigger warnings may decrease negative affect.

Finally, it is still unclear why trigger warnings did not help those who valued them most (i.e., those who believed them to be protective). It is possible that such participants reported feeling bad as a way of expressing their belief that the content warranted a trigger warning

(i.e., motivated response), or that their increased anticipated negative affect resulted in a expectancy effects that counteracted the warnings' benefits. Future research should uncover these mechanisms and test whether providing specific emotion-regulation strategies attached to trigger warnings allows such warnings to be more useful for those who value them most.

11. Conclusion

The present research is the first empirical research on the effects of trigger warnings on anticipated affect, emotion-regulation behavior, and experienced affect. By doing so, this research integrates literatures on interpersonal emotion-regulation, information avoidance, and warnings. Trigger warnings leverage a previously unmodeled, temporal dimension to interpersonal emotion-regulation to regulate future, anticipated emotions. In doing so, the present research also addresses a real-world debate by showing that trigger warnings can reduce negative emotions but can also increase avoidance. Thus, trigger warnings introduce difficult-to-weigh tradeoffs—avoidance of warned-of content might have short-term emotion-regulation benefits, but also could hypothetically result in decreased memory for important material or prevent people from learning to cope with distressing content. On the other hand, by reducing negative emotional experiences or signaling supportive environments, trigger warnings might promote engagement with otherwise distressing material in the long run. Making the tradeoffs that much more difficult to weigh is the fact that trigger warnings have heterogonous effects across populations–in the present research, beliefs about trigger warnings as protective (versus coddling) moderated their effect on important outcomes (e.g., experienced negative affect), and there are likely other important person-level variables (e.g., PTSD) that moderate the effect of trigger warnings, as well. Future research that explores the mechanisms through which trigger warnings work, the effects of trigger warnings across heterogenous populations, and the effects of trigger warnings across time and social context will help resolve these tradeoffs and allow people that use trigger warnings to best facilitate learning and well-being.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2018.08.006.

References

Aspinwall, L. G. (2011). Future-oriented thinking, proactive coping, and the management of potential threats to health and well-being. In S. Folkman (Ed.). *The Oxford handbook of stress, health, and coping* (pp. 1–35). . https://doi.org/10.1093/oxfordhb/9780195375343.013.0017.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences, 42*, 815–824.

Bass, S. A., & Clark, M. L. (2015). The gravest threat to colleges comes from within. *Chronicle of Higher Education, 62*(7), A26–A27.

Batson, C. D. (2017). Empathy and altruism. In K. W. Brown, M. R. Leary, K. W. Brown, & M. R. Leary (Eds.). *The Oxford handbook of hypo-egoic phenomena* (pp. 161–173). New York, NY, US: Oxford University Press.

Berkowitz, L., & Thome, P. R. (1987). Pain expectation, negative affect, and angry aggression. *Motivation and Emotion, 11*(2), 183–193.

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.

Bradley, M., & Lang, P. J. (1994). Measuring emotion: The self-assessment semantic differential manikin and the. *Journal of Behavior Therapy and Experimental Psychiatry, 25*(1), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9.

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences, 16*(2), 106–113. https://doi.org/10.1016/j.tics.2011.12.010.

Bushman, B. J. (1998). Effects of warning and information labels on consumption of full-fat, reduced-fat, and no-fat products. *The Journal of Applied Psychology, 83*(1), 97–101. https://doi.org/10.1037/0021-9010.83.1.97.

Bushman, B. J., & Stack, A. D. (1996). Forbidden fruit versus tainted fruit: Effects of warning labels on attraction to television violence. *Journal of Experimental Psychology: Applied, 2*(3), 207–226. https://doi.org/10.1037/1076-898X.2.3.207.

Butler, E. A., & Randall, A. K. (2013). Emotional coregulation in close relationships. *Emotion Review, 5*(2), 202–210. https://doi.org/10.1177/1754073912451630.

Cacioppo, J. T., Glass, C. R., & Merluzzi, T. V. (1979). Self-statements and self-evaluations: A cognitive-response analysis of heterosocial anxiety. *Cognitive Therapy and Research, 3*(3), 249–262.

Cohen, S., & Wills, T. A. (1985). Stress, social support, and the buffering hypothesis. *Psychological Bulletin, 98*(2), 310–357. https://doi.org/10.1037/0033-2909.98.2.310.

Cox, E. P., Wogalter, M. S., Stokes, S. L., Murff, E. J. T., Iii, E. P. C., Wogalter, M. S., ... Murff, E. J. T. (1997). Warnings. *Journal of Public Policy & Marketing, 16*(2), 195–204.

Crum, A. J., Salovey, P., & Achor, S. (2013). Rethinking stress: The role of mindsets in determining the stress response. *Journal of Personality and Social Psychology, 104*(4), 716.

Dweck, C. S. (2012). Mindsets and human nature: Promoting change in the Middle East, the schoolyard, the racial divide, and willpower. *American Psychologist, 67*(8), 614.

Earl, A., Albarracín, D., Durantini, M. R., Gunnoe, J. B., Leeper, J., & Levitt, J. H. (2009). Participation in counseling programs: High-risk participants are reluctant to accept HIV-prevention counseling. *Journal of Consulting and Clinical Psychology, 77*, 668–679.

Earl, A., Crause, C., Vaid, A., & Albarracín, D. (2016). Disparities in attention to HIV-prevention information. *AIDS Care, 28*(1), 79–86.

Earl, A., & Hall, M. P. (2018). Motivational influences on attitudes. In D. Albarracín, & B. T. Johnson (Eds.). *Handbook of attitudes*(2nd edition). (in press).

Earl, A., & Nisson, C. (2015). Applications of selective exposure and attention to information for understanding health and health disparities. *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource,* 1–14.

Earl, A., Nisson, C. A., & Albarracín, D. (2015). Stigma cues increase self-conscious emotions and decrease likelihood of attention to information about preventing stigmatized health issues. *Acta de Investigación Psicológica, 5*, 1860–1871.

Erceg-Hurn, D. M., & Steed, L. G. (2011). Does exposure to cigarette health warnings elicit psychological reactance in smokers? *Journal of Applied Social Psychology, 41*(1), 219–237.

Geers, A. L., & Lassiter, G. D. (1999). Affective expectations and information gain: Evidence for assimilation and contrast effects in affective experience. *Journal of Experimental Social Psychology, 35*(4), 394–413. https://doi.org/10.1006/jesp.1999.1377.

Gneezy, A., Gneezy, U., & Lauga, D. O. (2014). A reference-dependent model of the price-quality heuristic. *Journal of Marketing Research, 51*(2), 153–164. https://doi.org/10.1509/jmr.12.0407.

Gross, J. J. (1998). The emerging field of emotion-regulation: An integrative review. *Review of General Psychology, 2*(5), 271–299.

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.

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychological bulletin, 135*(4), 555.

Hayes, A. F. (2012). PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling [White paper]. Retrieved from http://www.afhayes.com/public/process2012.pdf.

Howell, J. L., & Shepperd, J. A. (2012). Reducing information avoidance through affirmation. *Psychological Science, 23*(2), 141–145. https://doi.org/10.1177/0956797611424164.

Just, M. A., & Carpenter, P. A. (1987). Speed reading. In M. A. Just, & P. A. Carpenter (Eds.). *The psychology of reading and language processing* (pp. 425–452). Newton, MA: Allyn and Bacon.

Kamenetz, A. (2016, September 07). Half of professors in NPR Ed survey have used 'trigger warnings'. Retrieved April 12, 2017, from http://www.npr.org/sections/ed/2016/09/07/492979242/half-of-professors-in-npr-ed-survey-have-used-trigger-warnings.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(5), 480–498.

Larsen, R. J., & Diener, E. (1992). Problems and promises with the circumplex model of emotion. *Review of Personality and Social Psychology, 13*, 25–59.

Lazarus, R. S. (1991). Progress on a cognitive-motivational-relational theory of emotion. *American Psychologist, 46*(8), 819–834.

Leventhal, H., Brown, D., Shacham, S., & Engquist, G. (1979). Effects of preparatory information about sensations, threat of pain, and attention on cold pressor distress. *Journal of Personality and Social Psychology, 37*(5), 688–714. https://doi.org/10.1037/0022-3514.37.5.688.

Lukianoff, G., & Haidt, J. (2015, September). *The coddling of the American mind.* The Atlantic. Retrieved from www.theatlantic.com/magazine/archive/2015/09/the-coddling-of-the-american-mind/399356/?utm_source_hpfb.

McNally, R. J. (2014, May 20). Hazards ahead: The problem with trigger warnings. Retrieved July 10, 2018, from https://psmag.com/education/hazards-ahead-problem-trigger-warnings-according-research-81946.

Mellers, B., Schwartz, A., & Ritov, I. (1999). Emotion-based choice. *Journal of Experimental Psychology: General, 128*(3), 332–345.

Miller, S. M. (1987). Monitoring and blunting: Validation of a questionnaire to assess styles of information seeking under threat. *Journal of Personality and Social Psychology, 52*(2), 345.

Oyserman, D. (2015). *Pathways to success through identity-based motivation.* USA: Oxford University Press.

Price, D. D., Finniss, D. G., & Benedetti, F. (2008). A comprehensive review of the placebo effect: Recent advances and current thought. *Annual Review of Psychology, 59*, 565–590. https://doi.org/10.1146/annurev.psych.59.113006.095941.

Richards, H. J., Benson, V., Donnelly, N., & Hadwin, J.a. (2014). Exploring the function of selective attention and hypervigilance for threat in anxiety. *Clinical Psychology Review, 34*(1), 1–13. https://doi.org/10.1016/j.cpr.2013.10.006.

Rimé, B. (2007). Interpersonal emotion-regulation. In J. J. Gross, & J. J. Gross (Eds.). *Handbook of emotion-regulation* (pp. 466–485). New York, NY, US: Guilford Press.

Rooke, S., Malouff, J., & Copeland, J. (2012). Effects of repeated exposure to a graphic smoking warning image. *Current Psychology, 31*(3), 282–290.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

Schmidt, P. (2015). Many instructors embrace trigger warnings, despite their peers' misgivings. *Chronicle of Higher Education, 61*(39), 9.

Stokes, M. (2014). In defense of trigger warnings. *The chronicle of higher education*. Retrieved from http://chronicle.com.proxy.lib.umich.edu/blogs/conversation/2014/05/29/in-defense-of-trigger-warnings/.

Sweeny, K., Melnyk, D., Miller, W., & Shepperd, J. A. (2010). Information avoidance: Who, what, when, and why. *Review of General Psychology, 14*(4), 340–353. https://doi.org/10.1037/a0021288.

Sweeny, K., & Shepperd, J. A. (2010). The costs of optimism and the benefits of pessimism. *Emotion, 10*(5), 750–753.

Trigger Warning. (n.d.). In Oxford Living Dictionary. Retrieved from https://en.oxforddictionaries.com/definition/trigger_warning

Trigger Warning. (n.d.). Retrieved from http://geekfeminism.wikia.com/

van Rooijen, R., Ploeger, A., & Kret, M. E. (2017). The dot-probe task to measure emotional attention: A suitable measure in comparative studies? *Psychonomic Bulletin & Review, 24*(6), 1686–1717.

Wadlinger, H. A., & Isaacowitz, D. M. (2011). Fixing our focus: Training attention to regulate emotion. *Personality and Social Psychology Review, 15*(1), 75–102. https://doi.org/10.1177/1088868310365565.

Wilson, T. D., Lisle, D. J., Kraft, D., & Wetzel, C. G. (1989). Preferences as expectation-driven inferences: Effects of affective expectations on affective experience. *Journal of Personality and Social Psychology, 56*(4), 519–530.

Wood, W., & Quinn, J. M. (2003). Forewarned and forearmed? Two meta-analytic syntheses of forewarnings of influence appeals. *Psychological Bulletin, 129*(1), 119–138. https://doi.org/10.1037/0033-2909.129.1.119.

Zaki, J., & Williams, W. C. (2013). Interpersonal emotion-regulation. *Emotion, 13*(5), 803–810. https://doi.org/10.1037/a0033839.