http://researchcommons.waikato.ac.nz/

Research Commons at the University of Waikato

Copyright Statement:

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

The thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of the thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from the thesis.

“Trigger Warnings” Are Trivially Effective at Reducing Distress

A thesis submitted in fulfilment of the requirements for the degree

of

Doctor of Philosophy in Psychology

at

The University of Waikato

by

MEVAGH SANSON

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2018

Abstract

University students are requesting, and many of their professors are issuing, "trigger warnings." Trigger warnings are a kind of content warning; they give a précis of the material to follow and caution that it may cause a reader or viewer to experience symptoms of distress. The use of these trigger warnings in higher education is controversial: Their purpose is to reduce or eliminate the distress students may otherwise experience in response to "triggering" course content, but no data exist regarding how effective they are at achieving this purpose. Moreover, there is reason to suspect these warnings may backfire and exacerbate the very symptoms they are intended to alleviate. We[1] conducted the first investigation into the effects of trigger warnings: We gave some subjects a trigger warning, but not others, then exposed all of them to negative material, and measured the analogue symptoms of posttraumatic stress disorder they experienced. Meta-analyses on data gathered across our six experiments revealed that trigger warnings had trivial effects—regardless of if subjects had seen a trigger warning beforehand, they judged the material to be similarly negative, and reported similar levels of negative affect, intrusions, and avoidance. These results suggest that trigger warnings are neither meaningfully helpful nor harmful, and that students and professors should not rely on them to mitigate students' distress.[2]

---

[1] Here, my use of the word "we" reflects that, although the research in this thesis is my own, I conducted it in a lab where I supervised a team comprised of undergraduate and Honours students. I also received advice and direction from my supervisors. For those reasons, I often use the word "we" in this thesis. Elsewhere in this thesis, I use the word "we" in a different sense; for example, to refer to what is or is not known in the wider scientific community.

[2] Portions of this thesis were adapted from:
Sanson, M., Strange, D., & Garry, M. (2018). *"Trigger warnings" are trivially helpful at reducing symptoms of distress.* Manuscript submitted for publication.
But I have expanded on the introduction, methods, results, and discussion.

Acknowledgements

I am grateful for funding from Victoria University of Wellington, the University of Waikato, a Fulbright New Zealand Graduate Award, and a New Zealand Federation of Graduate Women Fellowship, all of which helped support my PhD research.

I appreciate the many subjects who participated in these experiments. Thanks are also due to the undergraduate and honours students who helped with the data collection and materials creation for these experiments, and contributed to the discussions about this project.

To the Garry lab family: thank you all for being incredible sources of advice, friendship, inspiration, and mentorship through every twist and turn of this journey—and there have been a few! It has been an honour to walk this path alongside many of you, and to follow in the footsteps of those of you who graduated before me.

Special thanks to Deryn for welcoming me into her lab for a year and contributing a lot to this project, and to my adoptive labsters at John Jay for enriching my time in the US.

Thanks also to Gina Grimshaw for her mentorship throughout my time at Victoria University of Wellington.

To Maryanne: thank you for everything you have taught me, for the many opportunities you have given me, and for the innumerable ways you have helped me grow, as a researcher and as a person. I am profoundly grateful for the way in which you have shaped the course of my life.

Finally, to my friends and family: thank you for your love and support, expressed over and over in too many ways to list. Suffice to say, I couldn't have done it without you.

Contents

List of Figures and Tables

Chapter 1

The Origin and Spread of Trigger Warnings

Universities around the world are grappling with demands for *trigger warnings* (for example, in the UK, in the US, and in Australia: Manning & Wace, 2016; Medina, 2014; Palmer, 2017). These warnings caution students that the course content to follow may lead them to feel upset, or otherwise experience symptoms of distress—the implication being that the students may be harmed, and their learning may be negatively affected (Gust, 2016; Jarvie, 2014; Medina, 2014; National Coalition Against Censorship [NCAC], 2015; Palmer, 2017; University of California, Santa Barbara, Associated Students Senate [UCSBASS], 2014).

For example, in early 2014, a student newspaper at Rutgers University published an article suggesting that assigned readings of "works with grotesque, disturbing and gruesome imagery" should come with warnings alerting students to "the graphic content within these works [that] often serve as trauma triggers" (Wythe, 2014, para. 2, para. 3). The rationale for this suggestion was that these warnings would protect students, by allowing the students to "plan their reading schedule ahead of time for tackling triggering massages [sic] and/or discussing an alternate reading schedule with their professor" (Wythe, 2014, para. 9).

The following week, the student senate at the University of California, Santa Barbara attempted to make it mandatory for professors to include in their syllabi forewarnings about "[r]ape, Sexual Assault, Abuse, Self-Injurious Behavior [sic], Suicide, Graphic Violence, Pornography, Kidnapping, and Graphic

Depictions of Gore," because "[h]aving memories or flashbacks triggered [by those topics] can cause the person severe emotional, mental, and even physical distress. These reactions can affect a student's ability to perform academically" (UCSBASS, 2014, para. 7, para. 10). In fact, a recent survey of Arts and Humanities faculty in the US found that, of its respondents, "7.5% reported that students had initiated efforts to *require* trigger warnings on campus…and 12% reported that students had complained about the absence of trigger warnings" (NCAC, 2015, p. 3).

Taken together, these anecdotes and data provide evidence that at least some students believe being forewarned about the presence of negative material is a good idea. But where did those students get the idea that such forewarnings would be helpful?

The exact origin of trigger warnings is unclear—they seem to have evolved gradually into their current form. Journalists have examined this issue, and at least one has drawn links between reports of long-standing customs in group therapy sessions and the current use of trigger warnings (Vingiano, 2014). Specifically, she suggests that, with the advent of internet chat rooms, this group therapy community moved its discussions online and took with it the custom of prefacing material about some topics with a content warning. Wherever the practice of using trigger warnings originated, it is clear that niche online communities have long practiced labelling content that refers to traumatic experiences, or otherwise mentions subject matter that (in the judgment of the user posting the content) might elicit symptoms of disorders in other users (Jarvie, 2014; Marcotte, 2013; Vingiano, 2014; Waldman, 2016). The understanding in these communities is that

these warnings give users the choice to either avoid looking at or reading that content, or to proceed to do so with caution, having had the chance to mentally forearm themselves in whatever ways they think will help them cope with their reaction to the content (Jarvie, 2014; Marcotte, 2013; Vingiano, 2014; Waldman, 2016).

More recently, the practice of signposting material as containing possible "triggers" has become mainstream: Trigger warnings have become routine online, appearing on widely used sites including Tumblr, Twitter, and even the Government of Canada's website (Daro, 2016; Vingiano, 2014; Waldman, 2014; see also, Marcotte, 2013). Moreover, trigger warnings have migrated into the offline world, where they are increasingly applied to educational material on college campuses (American Association of University Professors [AAUP], 2014; Jarvie, 2014; Kamenetz, 2016; NCAC, 2015).

Trigger Warnings in Education: A Contentious Issue

How widespread is the use of trigger warnings in higher education? The aforementioned survey of Arts and Humanities faculty in the US further found that over half of respondents reported they had issued warnings to their students about course content (NCAC, 2015). Even more recently, when National Public Radio in the US surveyed a more diverse group of American faculty, they found very similar results: just over half of their respondents said they had issued warnings to their students about "potentially difficult" content (Kamenetz, 2016, para. 6). These data are not necessarily representative of all disciplines (cf. Boysen, Wells, & Dawson, 2016, for evidence these results may not generalise to

psychology instructors, for instance). But they nonetheless paint a picture in which trigger warnings now loom large on campuses.

What is more, there is evidence this rise to prominence has led academics—from disciplines as varied as communication and journalism, law, medicine, philosophy, politics and international relations, and queer studies—to consider the implications that trigger warnings have for their teaching and research (Bentley, 2016; Forstie, 2016; Kumagai, Jackson, & Razack, 2017; Manne, 2015; Suk Gersen, 2014; Wyatt, 2016). Perhaps unsurprisingly, views among the professoriate are divided (NCAC, 2015; Wyatt, 2016).

Some professors support the use of trigger warnings (Gust, 2016; Hanlon, 2015; Manne, 2015; NCAC, 2015). Many of their arguments describe trigger warnings as pedagogical courtesies that demonstrate educators' care for the mental wellbeing of their students (Wyatt, 2016). For example, one professor said she believes warnings allow students to "better manage their reactions" to possibly upsetting material (Manne, 2015, para. 4). Another said that giving students a trigger warning "reminds them to be particularly aware of the skills and coping strategies that they have developed and to switch them on" (Gust, 2016, para. 5).

In line with the views expressed by those professors, some universities have begun drawing up and implementing "pro-trigger warning" policies. For example, in 2013, a policy task force at Oberlin College wrote guidelines urging professors not to teach potentially "triggering" course material unless it was unavoidable, and to use trigger warnings before any such material (Oberlin College, Office of Equity Concerns, 2013). Closer to home, in 2017, Monash University began

piloting a policy that requires academics to place warnings about any "emotionally confronting" content in course outlines (Palmer, 2017, para. 4; see also, Lesh, 2016). But the Oberlin policy was subsequently withdrawn due to outcry from some of its faculty (Medina, 2014; Wilson, 2015).

Indeed, looking beyond Oberlin, many professors do not hold a positive view of trigger warnings (AAUP, 2014; Wyatt, 2016). Some of these professors are concerned about the effects trigger warnings may have on academic freedom of speech. For example, the American Association of University Professors issued a statement saying trigger warnings may have a chilling effect on the quality of education, both by narrowing the range of what is taught and by constraining students' interpretations of it. Moreover, the statement suggested it was unlikely that trigger warnings would help the students for whom they are intended, because educators cannot anticipate all possible "triggers," and simply pointing out upsetting topics does not constitute an effective treatment plan for students who become significantly distressed (AAUP, 2014). More recently, the AAUP's Canadian counterpart expressed similar sentiments (Canadian Association of University Teachers, 2015).

Individual universities have also taken public stands against trigger warnings, citing similar problems. For example, in 2016, just before the start of the US academic year, the University of Chicago sent a letter welcoming incoming first-year students. In this letter, students were told that "we do not support so-called 'trigger warnings,'" largely out of concern that the warnings could become a form of censorship (Grieve, 2016, para. 2).

But decline in the quality of education and restrictions on academic freedom are not the only potential problems that may stem from the use of trigger warnings. Still other professors—many of them professors of psychology—have weighed in on the "anti-trigger warning" side, out of concern for the mental wellbeing of students. These professors have suggested that—despite the intended effects of trigger warnings—flagging content as potentially "triggering" may backfire and actually be harmful to students. For example, Edna Foa, a clinical psychologist and expert in anxiety disorders and posttraumatic stress disorder (PTSD), has said issuing trigger warnings to students could harm them, because "[i]f we act as though they cannot handle distressing ideas, we communicate the unhelpful message that they are not strong" (quoted in Waldman, 2016, para. 25; for a review of her work, see Foa & McLean, 2016). Similarly, Jonathan Haidt, a social psychologist and expert in American "culture wars," and his lawyer colleague, Greg Lukianoff, wrote that

> expansive use of trigger warnings may also foster unhealthy mental habits in the vastly larger group of students who do not suffer from PTSD or other anxiety disorders. People acquire their fears not just from their own past experiences, but from social learning as well. (Lukianoff & Haidt, 2015, para. 32)

In other words, trigger warnings may lead many students—not only those at whom the warnings are typically aimed, but also those for whom they are not necessarily intended—to be more distressed by the content that follows than they otherwise would be (see also, McNally, 2014).

The professoriate is not the only group on campuses split by the issue of trigger warnings. One small survey of undergraduate students found that the same proportions of students (24%) agreed as disagreed that having read a trigger warning about topics such as genocide and acts of terrorism led them to feel better prepared to encounter that material in class (Bentley, 2017). This finding suggests that students, too, are divided on how useful trigger warnings are.

If we step back, then, and summarise the current state of affairs, we see that trigger warnings are widely used in higher education (as well as elsewhere). But their use is controversial. Indeed, at least one writer has dubbed trigger warnings a source of "never-ending…debate" (Flaherty, 2015). Yet something is missing that might help to settle the debate: data. There exists no empirical research on the effects of trigger warnings. Google Scholar and Web of Science searches (most recently conducted January 29, 2018) revealed no published experiments addressing the issue. Thus, as best as we have been able to ascertain, all the coverage by journalists, and discussion among faculty and students—not to mention all the policy-making both for and against trigger warnings at universities—has happened in the absence of data regarding what trigger warnings do or do not do.

Determining what trigger warnings do, or do not do, is of both practical and theoretical importance. On the practical side, surely it is important for universities, professors, and students alike to know the extent to which trigger warnings achieve their intended purpose—meaningfully reducing distress brought on by "triggering" material—as well as uncovering any unintended side-effects warnings may have on students' wellbeing or learning. On the theoretical side, it

is possible to draw on psychological research to make the case that trigger warnings should be helpful. But it is also possible to draw on psychological research to make the case that trigger warnings should be harmful. Such a discord means that it is important to determine if, when, and why trigger warnings increase or decrease distress.

Posttraumatic Stress Disorder and Trigger Warnings

Trigger warnings are commonly aimed at people who suffer from PTSD (for example, UCSBASS, 2014). That is, these warnings are often framed as a way to curb the distress people may feel in response to material that reminds them of a previous traumatic experience. But trigger warnings are also discussed more broadly, as a way to reduce the distress that many people may feel in response to potentially upsetting material per se, without reference to their having a diagnosed disorder—and sometimes even without reference to specific prior experiences (for example, Friedersdorf, 2016; Gust, 2016; Manning & Wace, 2016; NCAC, 2015). In other words, when it comes to students, trigger warnings are primarily intended to reduce distress following exposure to negative content encountered in course materials.

What are these symptoms of distress? Table 1 summarises a widely used list of symptoms, which are also diagnostic of PTSD (American Psychiatric Association [APA], 2013). As the table shows, to be diagnosed with PTSD, people must have experienced a traumatic event in one of several possible ways. This requirement comprises *criterion A*. Further, in the wake of that event, people must then report a debilitating combination of symptoms from each of four categories, which comprise *criteria B-E.* These symptoms include intrusive, mental

Table 1

The Diagnostic Criteria for Posttraumatic Stress Disorder (PTSD) in Adults

Criteria to be met (minimum to meet each)		Specific ways to meet each criterion	
A	Exposure to actual or threatened death, serious injury, or sexual violence (in at least one way)	1	Directly experiencing the event
		2	Witnessing the event occurring to others (in person)
		3	Learning the event occurred to a close family member/friend (where death-related event was violent or accidental)
		4	Repeated or extreme exposure to aversive details of the event (but not via seeing photos etc., unless work-related)
B	Intrusion symptoms (at least one)	1	Recurrent, involuntary, and intrusive distressing memories of the event
		2	Recurrent distressing dreams in which the content and/or affect of the dream are related to the event
		3	Dissociative reactions (e.g. flashbacks) in which the individual feels or acts as if the event were recurring
		4	Intense or prolonged psychological distress at exposure to internal or external cues that symbolise or resemble an aspect of the event
		5	Marked physiological reactions to internal or external cues that symbolise or resemble an aspect of the event
C	Persistent avoidance of stimuli (in at least one way)	1	Avoidance of or efforts to avoid distressing memories, thoughts, or feelings about or closely associated with the event

		2	Avoidance of or efforts to avoid external reminders that arouse distressing memories, thoughts, or feelings about or closely associated with the event
D	Negative alterations in cognitions and mood (in at least two ways)	1	Inability to remember an important aspect of the event
		2	Persistent and exaggerated negative belief or expectations about oneself, others, or the world
		3	Persistent, distorted cognitions about the cause or consequences of the event that lead to blaming oneself or others
		4	Persistent negative emotional state
		5	Markedly diminished interest or participation in significant activities
		6	Feelings of detachment or estrangement from others
		7	Persistent inability to experience positive emotions
E	Marked alterations in arousal and reactivity (in at least two ways)	1	Irritable behaviour and angry outbursts (with little or no provocation) typically verbal or physical aggression towards people or objects
		2	Reckless or self-destructive behaviour
		3	Hypervigilance
		4	Exaggerated startle response
		5	Problems with concentration
		6	Sleep disturbances
F	Duration of meeting Criteria B through E exceeds 1 month		

Note. These criteria were adapted from their description in the *Diagnostic and Statistical Manual,* 5th edition (American Psychiatric Association, 2013). Diagnosis also requires that the disturbance causes clinically significant distress or impairment, and that the disturbance is not due to the effects of a substance or other medical condition.

re-experiencing of the event, avoidance of things associated with the event, negative alternations in mood and cognition, and hyperarousal. Following a traumatic experience, many people experience some of these symptoms of distress for a relatively short time, but a few people experience these symptoms persistently, and develop a disorder (APA, 2013; Breslau, Kessler, Chilcoat, Schultz, Davis, & Andreski, 1998; Rothbaum, Foa, Riggs, Murdock, & Walsh, 1992).

Since 1980, when PTSD was first officially recognised as a disorder by the APA, the criteria for diagnosis, and the specific experiences and symptoms that count towards meeting those criteria, have been revised (Galatzer-Levy & Bryant, 2013). But some hallmark symptoms were recognised right from the start, such as intrusive memories of a negative experience, efforts to avoid reminders of the experience, and feelings of distress evoked by memories or reminders of it (Horowitz, Wilner, & Alvarez, 1979). Note that these are, of course, the very symptoms trigger warnings are said to alleviate or exacerbate (for example, Jarvie, 2014; Lukianoff & Haidt, 2015; Waldman, 2016). To understand how trigger warnings might alleviate or exacerbate them, we can turn to theories about why those symptoms arise.

Central to multiple theories of PTSD is the idea that how the person remembers their traumatic experience is crucial (Brewin, Dalgleish, & Joseph, 1996; Ehlers & Clark, 2000; Rubin, Berntsen, & Bohni, 2008; Rubin, Dennis, & Beckham, 2011; for a review, see Brewin & Holmes, 2003). That is, the symptoms of PTSD—such as intrusive memories—arise as a result of how people remember the experience.

Processing theories of traumatic experiences. Some of these theories account for symptoms by positing that while people are experiencing a traumatic event, the way they process incoming information changes (Brewin et al., 1996; Ehlers & Clark, 2000). For example, according to "dual representation theory," people form a verbal memory of the trauma, which they can voluntarily recall. But in addition, people form a sensory memory of the trauma, that they can only access involuntarily, and which remains perfectly preserved. This second type of memory is cued by overlap between the current situation and the original situation, causing people to experience upsetting intrusions of the trauma. Reducing intrusions is achieved by creating new, verbal memories of the trauma that interfere with the retrieval of the sensory memory (Brewin et al., 1996; Brewin & Holmes, 2003). But these theories rest on the underlying assumption that these PTSD symptom-causing memories are formed by processes that operate only in relation to traumatic experiences, and there is evidence that memory for traumatic experiences is not "special" (Geraerts, Kozarić-Kovačić, Merckelbach, Peraica, Jelicic, & Candel, 2007; Porter & Birt, 2001; Talarico & Rubin, 2003).

Autobiographical memory theory of traumatic experiences. By contrast, the "mnemonic model" takes a different view, positing that when people experience a traumatic event, no special processes operate (Berntsen, Rubin, & Bohni, 2008; Rubin, Berntsen et al., 2008; Rubin et al., 2011). Rather, people's memory for an experience is formed by the same processes—regardless of valence—and so the same factors affect the qualities of memories for all types of experiences. What is known about factors affecting autobiographical memory can

therefore explain why some people develop PTSD symptoms following some experiences. Figure 1 depicts this model.

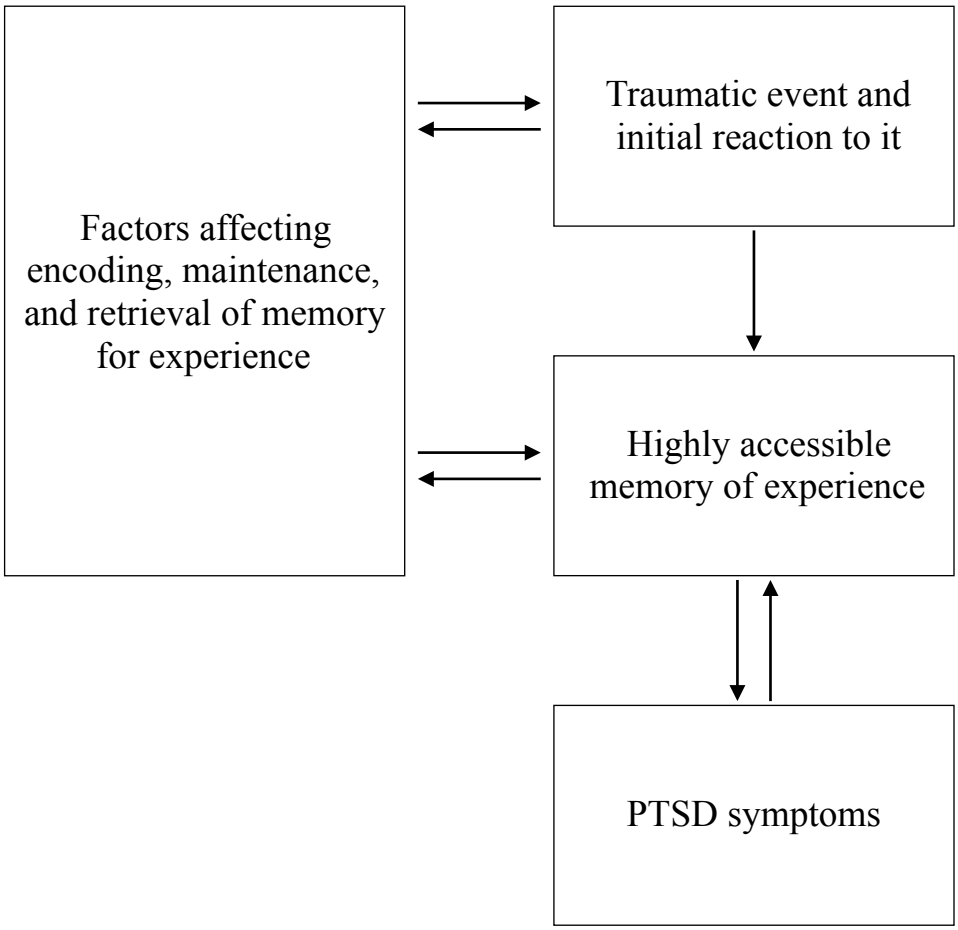

Figure 1

A mnemonic model of posttraumatic stress disorder (PTSD). A traumatic event happens; how people experience and encode that event is affected by a variety of factors. After a delay they have a memory of their experience—constructed anew each time it is retrieved; how people maintain and retrieve that memory is also affected by a variety of factors. These factors can in turn be influenced by the event and the memory of it. PTSD symptoms arise as a result of properties of this memory; these symptoms can in turn affect people's memory for their experience (adapted from Rubin, Berntsen, & Bohni, 2008).

What Figure 1 shows is that people experience intrusive memories of a trauma when their memory for it is "over-accessible," which means it is easily cued (Berntsen & Rubin, 2008; Berntsen et al., 2008; Rubin, Berntsen et al., 2008; Rubin et al., 2011). Therefore, any factor that increases the accessibility of people's memory for a negative experience will not simply make it easier to voluntarily retrieve that memory, but will increase their risk of developing PTSD symptoms. These factors include characteristics of the event, such as its emotional intensity (which should lead to better encoding), and individual differences, such as neuroticism (which should lead to greater intensity and more rehearsal). This model also suggests that "involuntarily" retrieved memories of the experience are more troubling than voluntarily retrieved ones, because people have less chance to regulate their emotional response. A growing body of evidence supports this model (Berntsen & Rubin, 2006; Berntsen & Rubin, 2008; Berntsen et al., 2008; Ogle, Siegler, Beckham, & Rubin, 2017; Rubin, Berntsen, et al., 2008; Rubin, Boals, & Berntsen, 2008; Rubin et al., 2011).

If the accessibility of a traumatic memory has a causal role in PTSD, it follows that resolving the symptoms of PTSD will mean reducing the accessibility of that traumatic memory, among other things. In line with that idea, two comprehensive reviews of treatments of PTSD have found evidence for the efficacy of exposure therapy and cognitive-behavioural therapy that target the traumatic memory (Bisson, Roberts, Andrew, Cooper, & Lewis, 2013; Committee on Treatment of Posttraumatic Stress Disorder, Institute of Medicine of the National Academies, 2008). It also follows that other interventions that change how accessible a negative memory is could change the rates of symptoms people

experience in relation to that memory. Further, if a particular cue becomes less able to prompt retrieval of that memory—as in the case of trigger warnings making negative material a less potent reminder—people should be less likely to experience symptoms. Likewise, if an intervention changes the properties of a negative memory when it comes to mind, people should find it less troubling.

Reasons to Expect Trigger Warnings Would Be Helpful

Several lines of research give us reason to expect that trigger warnings could ameliorate the symptoms of distress following exposure to negative material. For instance, trigger warnings could prompt people to regulate their emotions, or change how people believe they will respond to the material, thereby reducing people's unpleasant, intrusive thoughts related to negative material, their avoidance of thoughts and reminders of the material, and their ensuing feelings of distress (Gross, 2015; Kirsch, 1997). There is also reason to expect that trigger warnings could have further positive effects—on learning-related outcomes. Trigger warnings may change what people judge the warned-about material to be like, and how well people comprehend other material presented to them after the negative material (Leavitt & Christenfeld, 2011; Mooneyham & Schooler, 2013). In short, people who first see a trigger warning before negative material may subsequently experience fewer symptoms, think the material is less negative, and be better able to understand material they encounter afterwards.

Helping via emotion regulation. The first way in which trigger warnings might be helpful is if they prompt people to better regulate their emotions about the upcoming material. We know people use various strategies to regulate their emotions in an attempt to increase or decrease the degree or duration of negative

or positive affect they experience (Gross, 2015). The strategies are typically organised into five "families" primarily distinguished by when they are deployed in the time between encountering an emotion-provoking situation to experiencing an emotional response (for reviews, see Gross, 2015; Gross & Thompson, 2007; but see also, Aldao, Nolen-Hoeksema, & Schweizer, 2010).

According to this model, many of these strategies are proactive—that is, people can deploy them in anticipation of an emotional response, rather that once they are already having that emotional response (Gross, 2015; Gross & Thompson, 2007). For example, if people anticipate a situation is going to evoke emotions they do not want to experience, they can practice "situation selection" and take steps to avoid that situation, or vice versa (for example, Lang, Staudinger, & Carstensen, 1998). Similarly, people could choose to encounter the situation, but use "situation modification" to change its external aspects, in ways they think will make that situation less emotionally evocative (for example, McManus, Sacadura, & Clark, 2008).

Further along the chronology towards experiencing an emotional response, people could choose to experience the situation, but work on changing internal factors. That is, they could, via "attentional deployment," choose which aspects of the situation to pay attention to—focusing less on, say, the negative or positive aspects of the situation, and more on neutral aspects (depending on what sort of emotional response they are aiming to promote or avoid; for example, Bennett, Phelps, Brain, Hood, & Gray, 2007). Similarly, people could experience the situation but change internal aspects of the situation via "cognitive modification"—that is, reappraising their interpretation of the situation, which

could in turn change their emotional response to that situation (for example, Feinberg, Willer, Antonenko, & John, 2012).

Finally, once people are already experiencing an emotional response, they can attempt to shorten the time course or intensity of that response by engaging in "response modulation" of how they are feeling or behaving (for example, Gross & Levenson, 1997). People use this type of strategy once they have already begun to experience an emotional response to the situation.

One reason trigger warnings might be effective is they give people the opportunity to use any of these proactive strategies. For example, a trigger warning about negative course material might prompt a student to avoid the situation, and ask her professor for an alternative assignment. Second, the student might instead change external aspects of the situation, such as choosing to read (or watch) the negative material at home, rather than in the library. Third, while reading (or watching) the negative material, she could concentrate on the positive or neutral aspects of the material, distracting herself from the negative aspects. Fourth, she could reappraise the situation by, for instance, reminding herself that a particular assignment is a work of fiction. Effective use of any of these proactive strategies—helped along by a trigger warning—should mean people feel less negative affect, and judge the material to be less negative than they otherwise would (Gross, 2015); in fact, this outcome in itself would suggest trigger warnings were having one of their intended effects.

But a reduction in negative affect should be helpful in at least two more ways—both related to intrusive symptoms. First, if people find the negative material less intensely emotional, their memory of it afterwards should be less

accessible, because in general more emotional things are encoded better and rehearsed more often (Hall & Berntsen, 2008; Talarico, LaBar, & Rubin, 2004; for reviews, see also Christianson, 1992; Kensinger & Schacter, 2008). Then, if people's memory of the negative material is less accessible, they should experience fewer intrusions related to the material itself. In line with this idea, in one study, people saw a series of emotional photos and rated each for how emotionally arousing it was (Hall & Berntsen, 2008). The more arousing people initially found a photo, the more frequently they experienced involuntary memories of that photo in the days afterward.

Second, if people find negative material less negative, that material (and memories of it) should be less effective cues for people's other negative memories —namely, memories of previous traumatic experiences they have had—because involuntary memories are more likely to be retrieved when there is more overlap between the cue and the memory (Berntsen, Staugaard, & Sørensen, 2013). In line with this idea, in one study, people sometimes identified emotions as what had cued an involuntary memory they experienced (although other things, such as themes and objects, were more common cues; Berntsen & Hall, 2004).

Moreover, if the negative material is both less negative itself, and a less effective reminder of previous negative experiences, people should be less inclined to try not to think about it. In this way, trigger warnings could also reduce avoidance symptoms.

Helping via expectancies. A second way in which trigger warnings may be helpful is in relation to people's beliefs to do with the negative material—specifically, people's beliefs about how they will respond when exposed to

negative material, and their beliefs about how being forewarned that this negative material is coming will change the response they have to that material. These beliefs are examples of *response expectancies*.

Response expectancies are beliefs people hold about how they will respond to a given situation. People's expectancies about how they will respond to a situation causally contribute to how they actually respond, leading them to have the experience or behave in the way they expected they would (Kirsch, 1985, 1997). Yet people are not aware that their expectancies are what drive their responses—they attribute their response to the situation having set off automatic processes, out of their control (although people can access and report what they think their response would be in a given situation; Kirsch, 2004; Kirsch & Lynn, 1999). The classic example of an expectancy effect is the *placebo effect*, in which people respond to an inert substance, or other treatment, because they expect that substance or treatment will produce that response—for example, reducing their symptoms (for a review, see Price, Finniss, & Benedetti, 2008).

But expectancies have effects far broader than placebos (for reviews, see Michael, Garry, & Kirsch, 2012; Schwarz, Pfister, & Büchel, 2016). For example, people's expectations about how they will respond to a given situation are thought to arise from a number of different sources: prior experiences, associative learning, or social suggestion (Faasse & Petrie, 2016; Kirsch, 1985, 1997; Michael et al., 2012; Rief, Glombiewski, Gollwitzer, Schubö, Schwarting, & Thorwart, 2015). Once people encounter a situation about which they hold an expectancy, they respond in accordance with it—this response then brings about the expected outcome (Kirsch, 1997; Kirsch & Lynn, 1999). This expectancy-

induced response can take many forms, including observable physiological or behavioural changes—but can also include reported subjective changes, such as symptoms of mental disorders (Kirsch, 1997; Kirsch & Lynn, 1999; Rief et al., 2015).

Considered together, then, the literature suggests that if people hold the expectancy that trigger warnings make material less distressing, then seeing a trigger warning should lead people to experience fewer symptoms of distress (Kirsch, 1985, 1997). Thus, to the extent that people believe negative material will usually elicit involuntary memories, avoidance, and negative affect, and that prior warnings are helpful at reducing those symptoms, then people should experience decreases in these symptoms after seeing a trigger warning.

In line with this prediction, some research suggests people's beliefs about their ability to regulate their emotions matter (Catanzaro & Greenwood, 1994; Goldin et al., 2012). For example, in one study subjects reported how strongly they believed in their own ability to improve their negative moods, and the degree to which they had experienced symptoms of depression, at two time points (Catanzaro & Greenwood, 1994). At both times, the more subjects believed they could successfully regulate their negative moods, the fewer symptoms they reported experiencing, and what is more, change in the strength of their expectancy predicted change in their rate of symptoms. This research suggests that if trigger warnings lead people to expect they will be more successful at reducing the impact of subsequent negative material, we should see those people report feeling less negative.

What is more, some research suggests people's metacognitive and meta-memory beliefs (about how their own minds work, and what sorts of mental experiences are normal versus problematic) can lead them to interpret a given mental experience in one of several ways—and may lead them to be bothered by it, or not, depending on their beliefs. In one study, if people believed having involuntary memories was a worrying symptom, then a month later they were more likely to report having been bothered by unpleasant, intrusive memories following a negative experience (Takarangi, Smith, Strange, & Flowe, 2017; see also, Jamieson, Nock, & Mendes, 2013). Therefore, if trigger warnings lead people to interpret any symptoms they do experience as less bothersome, or simply makes those symptoms less noteworthy, people should end up reporting lower frequencies of these symptoms.

These studies point to an important role for people's beliefs in the creation of, noting of, and experience of their response to negative material. This literature give further purchase to the idea that trigger warnings should be helpful, to the extent these warnings fit with or alter people's beliefs about the material they precede, and people's response to that material, in way that reduces how distressing it is.

Helping via decreasing off-task thoughts. There is a third way that trigger warnings could be helpful. Research on mind-wandering shows that the more people's minds veer away from a reading task they are meant to be doing, the worse their comprehension of that reading is (Baird, Smallwood, Fishman, Mrazek, & Schooler, 2013; Schooler, Reichle, & Halpern, 2004; for a review, see Mooneyham & Schooler, 2013) The reason for this trade-off is most likely that

people have limited mental capacity to divide among concurrent tasks (Kane & McVay, 2012). If students who receive a trigger warning then experience less frequent distracting intrusions elicited by that material, then those students may be better able to comprehend other material they encounter soon afterwards, such as other readings. This helpful effect is one step removed from the primary purpose of trigger warnings, but nonetheless would have pedagogical benefit.

In fact, this negative relationship between off-task thoughts and reading comprehension has been observed specifically in the context of unpleasant, intrusive thoughts: In one study, people were instructed to not think about their former romantic partner and then asked to read an unrelated, non-fiction passage (Baird et al., 2013). The more times people were “caught" thinking about their ex-partner while reading, the worse their comprehension of that passage was. Similarly, in another study, people first viewed a film clip of a multi-fatality car crash, and then did an unrelated reading while they both self-reported and were intermittently asked about having off-task thoughts related to the clip (Takarangi, Strange, & Lindsay, 2014). The more of these intrusions subjects reported—specifically, the more times they reported, when asked, that they were thinking about the clip (but not the more times they reported intrusions, unprompted)—the worse they did on a later test about the reading.

Considered together, these studies suggest that if, while students are trying to read “ordinary,” more neutral class material, part of their mental resources are being used up by intrusions regarding traumatic experiences or previously encountered negative material, then their understanding of that class material is going to be hindered. Similarly, if they are distracted while reading (or watching)

negative material in class by intrusions about a traumatic experience, then their memory for that negative material might also be worse. But these studies further suggest that if trigger warnings reduce the incidence of these intrusive thoughts, then their comprehension of their primary reading task should be better.

Helping via fluency. Finally, at least one other line of research points to another possible positive effect of trigger warnings. This research suggests that knowing what to expect in a story may make that story more enjoyable. More specifically, two studies found that getting story "spoilers" that gave away the plot twist near the end of a story—despite people saying they did not like knowing the ending of a story before they had read it—increased people's reported enjoyment of that story, compared to people who read the story "unspoiled" (Leavitt & Christenfeld, 2011, 2013). One explanation for this effect is that knowing what is coming in a story leads reading it to feel more fluent, and fluency is generally associated with positive evaluations (Leavitt & Christenfeld, 2013; for a review, see Alter & Oppenheimer, 2009). It is possible that trigger warnings can act somewhat like these spoilers, creating a feeling of fluency by giving people more information about the content of the material ahead of time. The resulting positivity may counteract the negativity of the material, leading people who see a trigger warning to feel less negative, or to judge the material as less negative.

Reasons to Expect Trigger Warnings Would Be Harmful

Of course, we do not know if people use trigger warnings in any of these aforementioned helpful ways. In fact, there is also reason to expect trigger warnings could instead exacerbate symptoms of distress.

Harming via emotional (dys)regulation. First, people may not interpret trigger warnings as a prompt to down-regulate their negative emotions. Instead, they may interpret trigger warnings as an instruction that the negative aspects of the material are particularly noteworthy, and that a negative interpretation is the "correct" way to appraise the material. Put differently, trigger warnings may essentially lead people to adopt strategies that harmfully potentiate, or up-regulate, rather than diminish their negative emotional response to the material that follows (Gross, 2015; see also, Gross & Jazaieri, 2014).

If people interpret trigger warnings in this "negative focus" way, they may pay more attention to and better encode the negative aspects of the material. In line with this idea, some research has shown that knowing negative material is imminent can lead people to be more distracted by that material once they encounter it (Devue, Belopolsky, & Theeuwes, 2011; see also Kleinsorge, 2007). For example, in one study, people scared of spiders were repeatedly asked to identify whether a target line was present among an array of shapes (Devue et al., 2011). On some of these trials a task-irrelevant picture appeared on screen—and in some blocks these irrelevant pictures were spiders. In other blocks the distractors were other natural objects (such as butterflies, or leaves). When subjects knew a block of trials would contain spider pictures, when those pictures did appear the subjects were more distracted by them (that is, slower to respond) than when they knew the pictures would be of something else. This finding suggests that being able to anticipate a fearful situation led people to be more vigilant looking out for the feared object, so that when it did appear they devoted more attention to it than they would have.

If trigger warnings act in a similar way, then telling people to expect particular negative material may lead them to attend more to the negative aspects of the material when they come upon them. If warned people focus more on the negative aspects of the material they may find the material more upsetting, and better remember the negative aspects of the material, too (Devue et al., 2011; Gross, 2015; Hall & Berntsen, 2008). This increase in attention to the negative aspects, and increase in negative emotion felt may be particularly noticeable for people who are fearful about the negative content—such an effect would mean trigger warnings are particularly harmful to those who want them most.

In a related vein, telling people upcoming material will be negative may then change how they interpret (or appraise) that material for the worse (Cantor, Ziemke, & Sparks, 1984; de Wied, Hoffman, & Roskos-Ewoldsen, 1997). In each of two studies specifically looking at the effects of forewarning people about the content of horror films (similar to broadcasting standards-mandated ratings of television shows), some people were told they would see disturbing footage and some were not, and then they were shown footage, such excerpts from a film about vampires (Cantor et al., 1984; de Wied et al., 1997). Finally, subjects made a number of ratings about the footage they had seen, which revealed that forewarned subjects thought the clips were more upsetting than did unwarned subjects. What is more, this effect was greater for a more specific warning (detailing what the negative events that were going to be depicted were, such as a man driving a stake through the heart of his wife-turned-vampire) than a more general one (such as simply saying there would be a gruesome scene; Cantor et al., 1984). The researchers suggested these effects occurred because the

forewarnings created in people a vigilant anticipation, and constrained interpretation of the material.

Indeed, these findings fit more generally with classic work showing that prior information can affect which aspects of material people attend to and remember, and how they interpret it (Bransford & Johnson, 1972; Pichert & Anderson, 1977). For example, in one study, people were either told they should read a story about a house from the point of view of either a prospective home-buyer, or a prospective burglar, and later were asked to report everything they remembered about the house (Pichert & Anderson, 1977). Subjects in this study remembered more details relevant to the perspective they had adopted while they were reading about the event, suggesting the constraints on their perspective affected what information they took in (although a later study found that swapping perspectives led people to remember some additional details; Anderson & Pichert, 1978). If trigger warnings act similarly, implicitly constraining people to look out for negative aspects of the material, and focus on those negative aspects when they do arise, warned people would end up with a more negative impression of the material. Moreover, this bias would persist in how they remember the material, to the extent that they retain a "trigger warning perspective" when the memory comes to mind.

Other evidence also suggests that trigger warnings could continue to exert effects after the fact, by distorting or constraining people's memory for their interpretation of the material: In one study, people were asked to nominate and make ratings about a negative experience they had had (Takarangi & Strange, 2010). Then, regardless of how negative their experience actually was, everyone

received one of two forms of false feedback—either being told by experimenters that their event was far more negative than those other subjects had chosen to describe, or that it was far less negative than those other subjects had chosen to describe—or got no feedback. After a delay, when asked again about that same event, subjects who had been told their experience was overly negative rated it as more negative than other subjects rated theirs. This finding suggests the negative feedback led people to reappraise their memory of the event and come to remember it as worse than it was. Trigger warnings could have a similar effect: if they imply that the material to follow is more negative than what one would ordinarily encounter, then these warnings may not only lead people to better attend to negative aspects of material, and feel more negative emotions at the time, but also to remember it as even more negative afterward.

Taken together, these lines of research might lead us to expect that—although people could use trigger warnings as helpful prompts to reduce the negativity of their response—trigger warnings may instead lead people to feel worse after reading (or watching) the material, and to judge the material as more negative. These increases in negativity would then make it more likely for that material to crop up later, in the form of involuntary memories, and more likely to overlap with people's previous traumatic experiences and cue intrusive memories about those as well (Berntsen & Hall, 2004; Hall & Berntsen, 2008). What is more, when people's memories of traumatic events distort over time, those who come to remember things as worse (for example, a soldier who reported not having seen corpses immediately after returning from deployment, but several months later reports he did see corpses while deployed) also report an increased

rate of PTSD symptoms (Engelhard, van den Hout, & McNally, 2008; Southwick, Morgan, Nicolaou, & Charney, 1997).

Harming via expectancies. A second way in which trigger warnings may be harmful, and increase rates of symptoms, is via expectancies (Kirsch, 1985, 1997). Such an effect would be somewhat analogous to a *nocebo effect*—whereby when people expect that a substance or other treatment will have negative effects, that expectation leads people to generate responses to the treatment, producing those very effects (for a review, see Planès, Villier, & Malleret, 2016).

In the case of trigger warnings, telling people that the material to follow is distressing may lead them to expect they will feel more negative and experience more symptoms of distress, increasing the rate of those symptoms they experience after exposure to the negative material (Kirsch, 1985, 1997). Such a possibility fits with research showing that telling patients the side effects that a treatment may have leads to increased rates of those side effects (for a review, see Wells & Kaptchuk, 2012). In other words, instead of playing into an expectancy that warnings reduce symptoms, seeing a trigger warning may instead create in people the belief that the material to follow is dangerous, and that they will feel upset and experience other symptoms of distress in response to it. More specifically, trigger warnings might create—or play into—an expectancy that when people encounter particularly negative material they are likely to feel bad, and have intrusive thoughts related to it, and want to avoid thinking about it. As a result, when people go on to encounter that negative material having seen a trigger warning, they would have a stronger negative response to it than they otherwise would.

Harming via thought suppression. A third way in which trigger warnings may be harmful is if they encourage people to try and suppress their thoughts related to the negative material. Research shows that particularly trying to not think about something can backfire and lead people to think about it more (Harvey & Bryant, 1998; Wegner, Schneider, Carter, & White, 1987). One explanation for this effect is that in order to try not to think the unwanted thought, people have to keep in mind the very thought they do not want to be having (for a review, see Wenzlaff & Wegner, 2000).

If people take a trigger warning to mean that the material they are going to encounter will be unpleasant to think about, they may try to counter that negativity by increasing efforts to avoid thinking about it. This focus on trying not to think about the warned-about material may increase people's avoidance symptoms as well as increasing their intrusion symptoms.

Harming via increasing off-task thoughts. Finally, a fourth way in which trigger warnings may do harm is via their effects on outcomes of educational relevance. If trigger warnings increase the rate of intrusion symptoms, students may be distracted by those thoughts, which would worsen their comprehension of class material (Baird et al., 2013; Mooneyham & Schooler, 2013; Schooler et al., 2004; Takarangi, Strange et al., 2014). Of course, we do not know if people interpret trigger warnings in any of these harmful ways, any more than we know they will interpret them in helpful ways.

Reasons to Expect Trigger Warnings Would Be Unhelpful

Trigger warnings may work in a helpful direction, or they may work in a harmful direction. But a third possibility also exists: No matter what the direction

of trigger warnings' effects, the magnitude of those effects may mean that—in a practical sense—the warnings do nothing to people's responses.

Of course, not all small effects are meaningless (Abelson, 1985; Prentice & Miller, 1992; Rosenthal, 1990). It can be useful to know when the size of an effect is indistinguishable from zero—for example, in the case of ineffectual treatments. What is more, interventions that produce even a tiny effect may nevertheless be very important—if, for example, the effect influences mortality. If just one less person among thousands dies because of a given intervention, that effect is clearly meaningful (especially for that proverbial person). For other outcomes, the smallest effect size of clinical or practical significance can be harder to determine. For example, does a one-point reduction on a scale measuring degree of depression symptoms make a noticeable difference to a depressed person's life (Cuijpers, Turner, Koole, van Dijke, & Smit, 2014)?

In the absence of an answer to this question, with regard to outcomes that trigger warnings may affect, subjectivity remains regarding how big an effect trigger warnings would need to have for us to consider it meaningful. One suggested guideline for the minimum size an effect must be, to be considered practically meaningful, is a standardised mean difference of 0.41 (Ferguson, 2009). This criterion seems a reasonable place for us to start.

But there is reason to suspect that trigger warnings may yield effects too small to be considered meaningful. For example, people often fall prey to unwanted influences on their cognition and behaviour (such as the distress and intrusive thoughts that follow an encounter with negative material). But for a

variety of reasons those same people are sometimes unable to successfully correct for those influences (Wilson & Brekke, 1994).

Perhaps the difficulty arises because there are multiple ways in which the correction process can go awry (Wilson & Brekke, 1994). For one, sometimes people are unaware of the unwanted influence, in which case they cannot counter those influences. If people are aware of the influence but do not accurately know the direction or magnitude of its effect, then they cannot counter it. For another, if people do not have the desire to, or ability to do the countering, then they will fail to counter those influences. Trigger warnings may make people aware of the influence, but to the extent that trigger warnings are insufficiently informative, or the unwanted effects that trigger warnings are intended to counter are beyond people's ability to control, then trigger warnings will be ineffective.

Overview of Experiments

Trigger warnings are a contentious issue, yet no empirical research has investigated their effects. We aimed to address that gap in the literature. The literature gives us reason to expect that trigger warnings would reduce distress elicited by negative material, yet also gives us reason to expect trigger warnings may do very little in the way of meaningfully altering people's distress. Still other literature leads us to expect that trigger warnings could even increase the very distress they are meant to alleviate. What we do not know, then, is how previous research translates into this real-world situation.

Of course, one challenge with attempting to elucidate the "true" effects of trigger warnings is that many aspects of these warnings and their use are heterogeneous (Friedersdorf, 2016; Jarvie, 2014; Wilson, 2015): No widely agreed

upon set of defining criteria for trigger warnings exists, nor do instructions for when and how they should be used, nor is there consensus about trigger warnings' target population. Moreover, there is no constrained list of the warnings' intended effects, and concerns about the unintended consequences of trigger warnings are varied, too. In short, these variables mean there are many possible forms an initial empirical investigation could take.

We chose to begin the empirical enquiry by examining the direction and magnitude of trigger warnings' immediate effects on the wellbeing and learning of non-clinical populations. More specifically, our primary research question was: to what extent does a trigger warning affect the symptoms of distress people experience soon after exposure to "warned about" material? To address this question, we ran six experiments in which subjects did or did not see a trigger warning, were then exposed to negative material, and then reported their symptoms of distress. Table 2 outlines some of the key attributes of each of these experiments.

Subjects. As the second column of the table shows, we drew our samples from two populations. Each of these populations is comprised of people likely to often come across trigger warnings—for some experiments we recruited university students, and for others we recruited internet users. We did not recruit from populations of people known to be currently suffering from PTSD. Although trigger warnings are aimed by some at people who have diagnosed disorders, others use them or cite them as being used to minimise distress for everybody (AAUP, 2014; Friedersdorf, 2016; Gust, 2016; Manning & Wace, 2016; NCAC, 2015; UCSBASS, 2014). But even if trigger warnings were only for the benefit of

Table 2

Overview of Key Methodological Features of Each Experiment

Experiment	Subject pool	Independent variables	Material shown	Key dependent variables
1a	Students	Trigger warning	Story	Memory for material Negative affect Intrusions Avoidance
1b	MTurkers	Trigger warning	Story	Memory for material Negativity of material Negative affect Intrusions Avoidance
2a	Students	Trigger warning Negativity	Film	Negativity of material Negative affect Intrusions Avoidance
2b	MTurkers	Trigger warning Negativity	Film	Negativity of material Negative affect Intrusions Avoidance
3	MTurkers	Trigger warning Negativity	Film	Expected negativity of material Negativity of material Negative affect Intrusions Avoidance
4	MTurkers	Trigger warning Negativity	Film	History of trauma Negativity of material Negative affect Intrusions Avoidance

Note. Students were introductory psychology students at Victoria University of Wellington; MTurkers were members of Amazon's Mechanical Turk online crowdsourcing platform. "Trigger warning" means we manipulated the presence of a trigger warning; "Negativity" means we manipulated the negativity of the (film) material subjects were shown.

the minority of students with a disorder, all students still see them. For both of these reasons, it is therefore important to examine what effects trigger warnings

have on the majority of people. Moreover, to the extent that the same cognitive processes operate to produce symptoms of PTSD in people without and without the disorder—albeit producing these symptoms to different degrees—then we can make preliminary inferences from the effects trigger warnings have on these populations to the effects such warnings might have on clinical populations (Rubin, Boals et al., 2008; Rubin et al., 2011).

Trigger warnings and negative material. As the third column of the table shows, in each experiment we manipulated whether subjects received a trigger warning about the material all of them then saw. We used trigger warnings of the kind people might encounter online in everyday life, or during the course of their university studies. In the later experiments, as the third column of the table further shows, we also manipulated the negativity of the material we showed subjects after they were (or were not) warned about it.

As the fourth column of the table shows, the material that subjects in some experiments saw was a short story, whereas in other experiments it was a short film clip (of greater or lesser negativity). By using a range of materials, we can assess the effects of trigger warnings across a variety of circumstances.

The materials we warned subjects about were akin to materials for which people in real-world situations have requested and used trigger warnings. For example, humanities students have requested that warnings accompany some novels they are assigned to read (Wythe, 2014). Further, we chose these materials based on norming studies, in which we collected data about multiple stories and films.

Measures of distress and related outcomes. After subjects had read the story or watched the film, about which some of them had been warned, they then completed a variety of tasks, as indicated in the fifth column of the table. Broadly, the tasks that appeared across experiments measured how negative subjects were feeling in the wake of the material, how negative they judged that material to be, and the frequency with which they had experienced intrusion and avoidance symptoms since exposure to the material.

More specifically, we first asked subjects to rate the degree of negative affect they were currently feeling, after reading (or watching) the material. If trigger warnings are helpful at reducing how distressing people find negative material, then warned subjects should report feeling less negative than those who got no warning (or, conversely, if trigger warnings are harmful, "warning" subjects should report feeling more negative than "no warning" subjects).

Next, we asked subjects to read a short, non-fiction article—unrelated to the earlier material—and report intrusions related to the story or film that they experienced while reading this article. Because we used non-clinical samples, many of these intrusions are likely to be more related to the negative material itself than to our subjects' prior traumatic experiences. Still, the more negative the material seems, the more it might overlap with people's previous negative experiences, also cuing memories of those. Further, the more negative the material seems, the less people might try not to think about it. Thus, to the extent that trigger warnings affect how negative the material seems, or people's ability or desire to not think about it, this tally of intrusions should be informative.

We then asked subjects to rate the frequency of intrusion and avoidance symptoms they had experienced in the short time since exposure to the story or film. As for the intrusions-tally measure above, the better trigger warnings are at helping people reduce the degree of intrusions and avoidance symptoms they experience, the lower the ratings of the "warning" group should be relative to the "no warning" group (or, the more harmful trigger warnings are, the higher those ratings should be).

Next, we tested subjects' comprehension of the non-fiction article. Their performance on this test is relevant for two reasons: First, the more unnoticed intrusions subjects have—ones which, by definition, they cannot report in their intrusion-tally task, but would nonetheless hurt their ability to comprehend the content of the article—the worse their comprehension should be. Thus, this measure should give us some insight into non-self-reported intrusions, and how those are affected by trigger warnings. Second, if trigger warnings are somehow useful (or preoccupying) in a way that continues to affect students' learning after they have moved on from the warned-about material itself, we should see that "warning" subjects have different comprehension than "no warning" subjects—an effect that would be worthy of further investigation.

Finally, we asked subjects questions about the negative material itself, such as asking them to rate how negative they thought it was. If trigger warnings change what people think the material that follows is going to be like, trigger warnings may also change what people then judge that material was like. In the initial experiments, we also asked subjects questions about the negative stories, to see if trigger warnings affected how well people could remember the negative

material itself. In later experiments, we added additional measures to help gather evidence regarding possible explanations for the effects we observed.

Analyses. We adopted the recommended "New Statistics" approach as our primary form of analysis (American Psychological Association, 2009; Cumming, 2012; Eich, 2014). We present all our results for each experiment in terms of effect sizes and the confidence intervals around them. But we also carried out traditional Null-Hypothesis Significance Testing (NHST) analyses for some of our key comparisons, and report those as well. Finally, we carried out mini meta-analyses on our data, and present those results, too.

Chapter 2

Experiment 1a

Method.

Subjects. We aimed to recruit as many subjects as we could, within the constraints of departmental subject-hour allocations for the semester. We collected data from 254 introductory psychology students at Victoria University of Wellington, who participated in partial fulfilment of course requirements. Of the 240 whose data we retained for analysis (see below for details of the exclusion criteria), 75% were female and 25% male, and their ages ranged from 16-50, Median = 18, M = 18.68, 95% CI [18.36, 19.00].

Design. We used a two-group design, manipulating Presence of Trigger Warning (warning, no warning) between subjects.

Procedure. On consent forms, we told subjects we were interested in examining factors that affect the comprehension of different writing styles—we gave this cover story so as not to alert subjects that they may be exposed to negative material, prior to only some of them seeing a trigger warning. We gave subjects a full debriefing at the end of the semester. This methodology was approved by the School of Psychology Human Ethics Committee at Victoria University of Wellington.

Subjects participated in small groups, seated at individual computers in a lab. We used Qualtrics software (Qualtrics, 2018) to run the experiment, presenting all materials to subjects in a web browser window. The experiment had three phases. Figure 2 gives a general overview of these phases, which carried across each of our experiments.

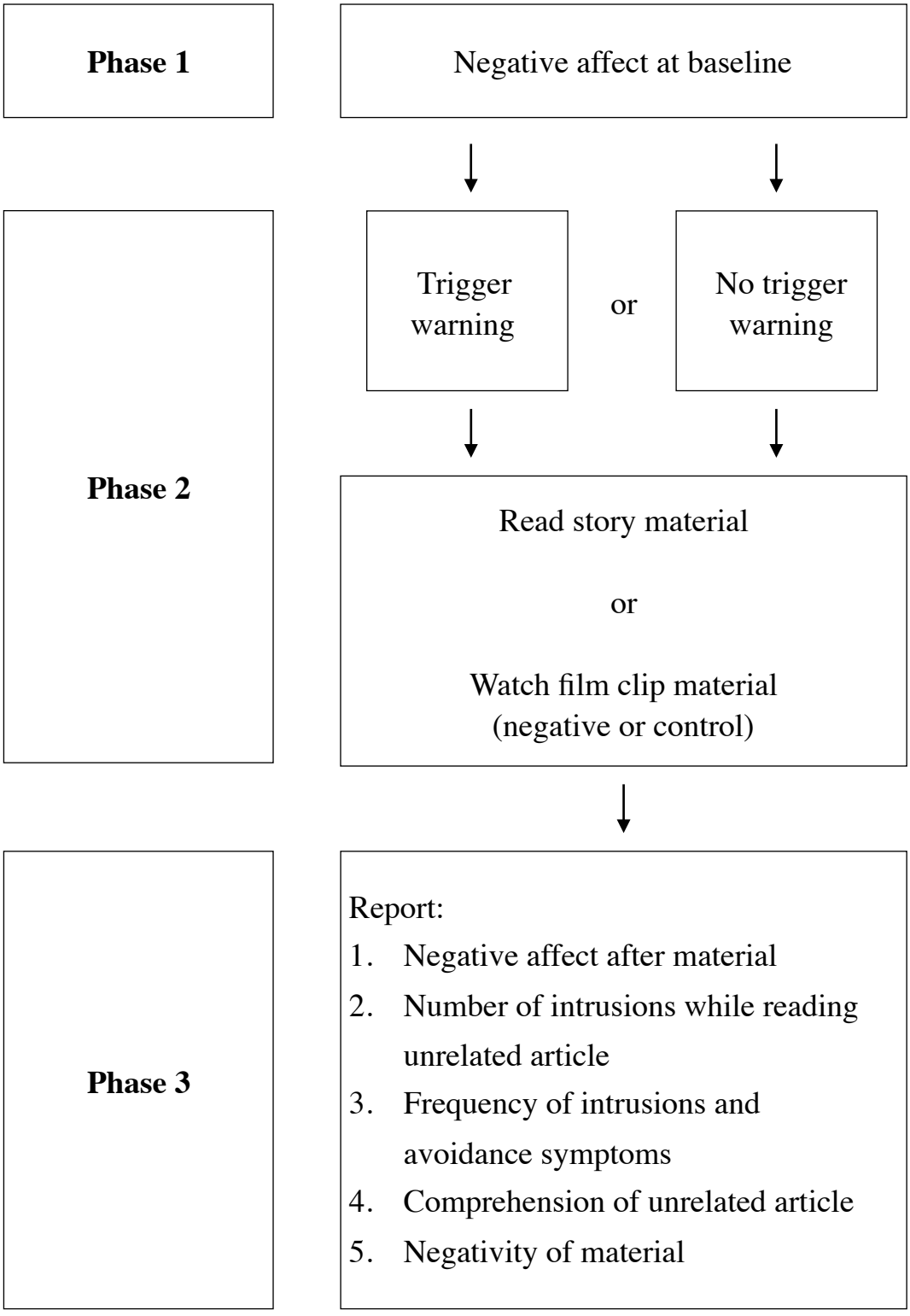

Figure 2.

Overview of the general method. These boxes summarise key components of the procedure, divided into three phases, for each of our experiments.

Phase one. To measure their baseline negative affect, subjects first completed the Positive and Negative Affect Schedule – Expanded form (PANAS-X; Watson & Clark, 1999). On this measure, subjects rate several affect-related words (such as "distressed") according to how much they feel that way "right now," from 1 (*very slightly or not at all*) to 5 (*extremely*). Ratings for the 10 items comprising the negative affect subscale are summed, yielding a total between 10-50. The negative subscale, rated regarding "the present moment," is internally consistent, with Cronbach's alpha of .85; has reasonable test-retest reliability across two months, with a correlation of .45; and has good external validity, correlating with the Hopkins Symptom Checklist, a measure of general distress over the past week, at 0.65 when rated regarding "today"; further, within-subject changes in ratings across a day regarding the present moment correlate with ratings of current stress (Watson, Clark, & Tellegen, 1988). Taken together, these psychometric properties suggest this scale is suitable for measuring how much negative affect subjects were feeling at various points during the experimental session.

Phase two. Next, subjects either saw a trigger warning, or did not see a trigger warning. We developed the warning subjects saw by looking at examples online, and at guidelines issued by student associations regarding the use of trigger warnings. Subjects in the "warning" condition read: "TRIGGER WARNING: The following story contains violence and death." Subjects in the "no warning" condition skipped this step.

Subjects were then randomly assigned to read one of two negative, fictional short stories. We chose to use stories because trigger warnings are often used or

requested prior to such material (for example, Wythe, 2014). The two stories we used were "A Dark Brown Dog" by Stephen Crane (of length 2356 words), and an abridged version of "The Veldt" by Ray Bradbury (which was originally of length 4603 words, but we edited it down to 3198 words in order to make reading time across story counterbalances more similar). These stories include animal and child abuse, and murder—topics of the kind for which warnings are issued or requested. We chose these particular stories based on norming data we gathered.

Norming of stories. To collect these norming data, we asked 120 introductory psychology students at Victoria University of Wellington to read one of eight possible stories (for a total of 15 responses about each story) and answer a variety of questions related to it. More specifically, we first asked subjects to complete the PANAS-X (Watson & Clark, 1999) with regard to how they were feeling "right now." Next, they read one short story; we timed how long they spent on this survey page. Afterwards, they completed the PANAS-X a second time—again, regarding how they were feeling "right now"; rated the story for how negative, positive, surprising, and interesting it was from 1 (*not at all*) to 7 (*extremely*); listed things in the story they or others might find disturbing or upsetting, and pleasant or amusing; indicated if they thought the story had a plot twist and, if so, what and where it was; and reported if they had read the story before, prior to the study. We then examined subjects' responses to each story, focusing on: change in the degree of negative affect that subjects reported feeling, and how negatively they rated the story, how long they took to read the story, and if they had previously read the story.

With regard to "A Dark Brown Dog," we found that, in norming, students rated this story as very negative (M = 5.60, 95% CI [5.10, 6.10], on the 1-7 scale), and likewise their degree of current negative affect increased from before to after reading this story (M_{diff} = 3.87, 95% CI_{diff} [0.81, 6.92]; although, given a maximum possible change of 40, this increase is slight). Further, students took approximately 10 minutes to read this story (M = 573 s, 95% CI [489, 657], Median = 620 s). Finally, none of the students had previously read this story.

With regard to "The Veldt," we found that, in norming, students rated the full-length version of this story as very negative (M = 5.27, 95% CI [4.62, 5.91]), and their degree of current negative affect very slightly increased from before to after reading this story (M_{diff} = 1.87, 95% CI_{diff} [-0.64, 4.37]). Further, students took approximately 15 minutes to read this story (M = 860 s, 95% CI [782, 938], Median = 857 s), and none of them had previously read it.

Phase three. After subjects in Experiment 1a had read one of those two stories, we measured three symptoms of distress—negative affect, intrusive cognitions, and avoidance—and subjects' memory of the story, using four tasks. First, to measure negative affect, subjects completed the PANAS-X a second time, again answering regarding how they were feeling "right now."

Second, to measure intrusions related to the negative story they had read, we next told subjects to focus on reading one of two randomly assigned non-fiction articles, unrelated to either story (specifically, these were 1022- and 1119-word excerpts from Bill Bryson's "A Short History of Nearly Everything," about atoms and cells, respectively—as used by Smallwood, Nind, & O'Connor, 2009), but to press the 'x' key on their keyboard each time "you notice that you are

experiencing an intrusive memory or thought about the story that you just read" (instructions adapted from Takarangi, Strange et al., 2014). These presses yielded a tally of intrusions. Subjects spent approximately 4 minutes reading their assigned non-fiction article and noting intrusions (*M* = 229.19 s, 95% CI [217.71, 240.66], Median = 213.13 s). Subjects then rated their adherence to the noting intrusions task, from 0 (*not at all well*) to 10 (*extremely well*; scale taken from Takarangi, Strange et al., 2014).

Third, to measure the frequency of their intrusions and avoidance symptoms, subjects completed the Impact of Event Scale (IES; Horowitz et al., 1979). On this measure, subjects rate how frequently items (such as "Pictures about it popped into my mind" and "I thought about it when I didn't mean to") have been true for them following a stressful event—here, reading the story—on a four-point scale from 0 (*not at all*) to 5 (*often*); ratings are then summed. We omitted two non-relevant items (relating to sleep disturbances), so the possible range for the intrusion subscale scores was 0-25, and for the avoidance subscale it was 0-40. Each subscale has good internal consistency, with Cronbach's alphas of 0.86 for intrusions and 0.82 for avoidance; and good external validity, each showing moderate to strong correlations with other measures of distress and PTSD (Sundin & Horowitz, 2002). These psychometric properties suggest this scale is suitable for measuring the frequency with which subjects have experienced intrusion and avoidance symptoms.

Fourth, to measure comprehension of the non-fiction article subjects read after they had read a negative story—and indirectly index intrusions they experienced while reading that article—subjects answered five four-alternative

forced-choice questions about it (these questions were those used by Smallwood et al., 2009). Subjects also answered five questions that tested their memory of the negative story they had read, also in four-alternative forced-choice format. We chose these questions about the stories based on norming data.

Norming of story questions. In order to arrive at the questions about the short stories, we first created a pool of 20 questions about each of the two stories we used. We then normed these 40 questions by giving 32 introductory psychology students one of the stories to read, and asking them all 20 questions about that story in a randomised order, such that 16 subjects saw each story and then answered questions about it. Next, for each question, we examined its difficulty (that is, the proportion of subjects who answered it correctly), and its reliability (that is, the strength of the correlation between whether subjects got that question correct and the overall number of questions they answered correctly). Based on those data we selected five questions for each story that, collectively, covered a range of aspects of the story and required a mixture of straightforward recall of details and a deeper understanding of the story. For both sets of five questions we chose, item difficulty ranged from 0.63 to 0.88. Item reliability for the chosen questions about “A Dark Brown Dog” ranged from 0.64 to 0.75; similarly, for the chosen questions about “The Veldt,” item reliability ranged from 0.62 to 0.81. These 10 chosen questions are listed in Appendix A.

At the end of Experiment 1a, subjects answered a few questions to establish whether their data should be excluded (such as asking if they had read either the story or the article before, and if they recalled seeing the trigger warning) and reported their demographics. The full list of exclusion criteria appears below.

Appendix B details the wording of all of the end-of-experiment questions we asked subjects drawn from a student population.

Results & Discussion.

Transformations. When we inspected the distributions of our measures, we saw that some of them had skewed distributions (mainly because many subjects reported relatively low absolute levels of symptoms). For each skewed measure, we investigated whether transformations helped to normalise the distribution; but they made little difference. As a result, unless otherwise noted, we analysed and reported untransformed data, here and for the experiments to follow.

Exclusions. We first excluded subjects who reported having previously read the story ($n = 4$), because they would have known what the story was about, regardless of if they saw a trigger warning. We also excluded subjects who had not correctly completed one or more critical tasks. Specifically, we excluded those who were not able to complete the experiment because they ran out of time, or who reported pressing the wrong key to note their intrusions, meaning their key presses were not able to be counted. In addition, we excluded four subjects because an earthquake disrupted their data collection. Altogether, we excluded 14 subjects (6%), leaving us with 240 subjects—121 in the warning condition, and 119 in the no warning condition.

Manipulation check. Before addressing our research question, we carried out a manipulation check. We calculated the percentage of "warning" subjects who, at the end of the experiment, reported that they remembered seeing the trigger warning before they read the negative story—95% of them reported they

did. This result suggests the vast majority of those subjects took note of our manipulation.

Memory of material. How well did subjects understand and remember the negative story they read, and to what extent did trigger warnings affect subjects' memory for that story? To answer that question, we calculated the proportion of questions subjects answered correctly about the story they read, classified by if they had seen a trigger warning beforehand or not. In both conditions the proportion of questions subjects got correct was high ($M_{\text{Warning}} = 0.79$, 95% CI [0.74, 0.83], $M_{\text{NoWarning}} = 0.81$, 95% CI [0.77, 0.85]), suggesting subjects attended to the stories and understood them well. Further, although "warning" subjects got a higher proportion of these questions right than did "no warning" subjects, the difference between the conditions was negligible—the maximum difference is 1, and so this difference represents a 2% movement on this measure. Moreover, the 95% CI around this difference is [-0.03, 0.08], which suggests that the true raw effect size of a trigger warning may plausibly be bigger (but only by a little), or it could be slightly in the opposite direction, or it could even be zero (Cumming, 2012). In NHST terms, this difference was not statistically significant, $t(238) = 0.93$, $p = .35$. These results suggest trigger warnings had little effect—and plausibly none at all—on how well subjects remembered the negative story they went on to read.

Negative affect. To what extent did trigger warnings influence how negative subjects felt after reading a negative story? We first checked that subjects' baseline negative affect levels (that is, their ratings on the first administration of the PANAS-X) were similar across those subsequently assigned to see a warning,

Table 3
Differences in Baseline Negative Affect Between Warning Conditions

Experiment	Difference before negative			Difference before control		
	M diff	95% CI LL	95% CI UL	*M* diff	95% CI LL	95% CI UL
1a	1.07	-0.52	2.66			
1b	1.06	-0.79	2.92			
2a	-1.42	-4.16	1.33	-0.50	-4.30	3.30
2b	-1.16	-3.06	0.75	-0.63	-2.20	0.94
3	0.50	-1.25	2.25	-1.11	-2.78	0.55
4	-0.54	-2.62	1.54	0.45	-1.48	2.37

Note. Negative affect measured using the Positive And Negative Affect Schedule (Watson & Clark, 1999; Watson, Clark, & Tellegen, 1988), prior to the introduction of manipulations. Differences calculated by subtracting the mean baseline negative affect of subjects who would not go on to see a warning from the mean baseline negative affect of those who would, separately for those who would go on to see a negative film and those who would go on to see a control film (where applicable).

and those not. As the mean difference (and 95% CI around it) displayed in Table 3 shows, these groups were indeed very similar, suggesting that at the start of the experiment, those subjects randomly assigned to see a warning, and those not, were comparable with regard to how negative they felt.

Then, to answer our research question, we turned to subjects' ratings of their negative affect after they had read their assigned story. We display these mean ratings of negative affect for "warning" and "no warning" subjects (and the 95% CIs around them) in Table 4, and the mean difference between those groups of subjects (and the 95% CI around that difference) in Table 5. As those tables show, after reading a negative story, "warning" subjects felt only very slightly more negative than did their "no warning" counterparts. The maximum difference on this scale is 40, and so this difference is very small—a 3% movement on this

Table 4

Descriptive Statistics for Key Measures of Distress Classified by Presence of Warning and Negativity of Material

Measure	Experiment	Warning and negative material			No warning and negative material			Warning and control material			No warning and control material		
		M	95% CI LL	95% CI UL	*M*	95% CI LL	95% CI UL	*M*	95% CI LL	95% CI UL	*M*	95% CI LL	95% CI UL
Negative rating	1a	–	–	–	–	–	–						
	1b	6.16	5.94	6.37	6.10	5.87	6.34						
	2a	6.52	6.23	6.81	6.30	5.81	6.78	2.96	2.26	3.67	3.14	2.54	3.75
	2b	6.46	6.21	6.71	6.61	6.43	6.79	2.78	2.34	3.22	3.03	2.58	3.48
	3	6.17	5.88	6.46	6.49	6.28	6.69	1.96	1.65	2.27	2.61	2.19	3.02
	4	6.35	6.09	6.61	6.47	6.20	6.74	2.65	2.24	3.05	2.87	2.48	3.27
PANAS negative[a]	1a	19.78	18.31	21.25	18.70	17.46	19.93						
	1b	17.45	15.73	19.18	16.15	14.47	17.83						
	2a	19.56	16.98	22.14	20.48	17.53	23.43	15.68	13.04	18.31	15.82	12.76	18.88
	2b	21.04	19.06	23.03	22.18	20.06	24.30	13.89	12.51	15.27	13.79	12.61	14.98
	3	20.83	18.90	22.76	21.69	19.82	23.56	14.64	13.29	15.99	15.17	13.51	16.83
	4	21.61	19.42	23.81	22.61	20.65	24.57	15.59	13.87	17.31	13.71	12.45	14.96
Intrusions tally	1a	6.86	5.93	7.80	7.06	6.08	8.03						
	1b	4.48	3.32	5.64	5.21	3.84	6.58						
	2a	10.83	5.76	15.90	8.16	5.76	10.57	7.08	4.95	9.21	8.71	5.44	11.99
	2b	7.02	5.42	8.62	7.50	6.04	8.96	4.97	3.72	6.21	4.79	3.53	6.06
	3	7.27	5.69	8.85	5.55	4.49	6.61	4.51	3.37	5.65	5.41	4.00	6.83
	4	6.64	5.22	8.05	7.27	5.84	8.70	4.78	3.66	5.90	5.74	4.33	7.16

IES intrusions[b]	1a	10.69	9.46	11.91	11.15	9.97	12.33						
	1b	9.61	7.94	11.28	10.33	8.50	12.15						
	2a	11.48	9.16	13.80	14.22	11.75	16.69	11.11	8.66	13.55	9.64	7.23	12.06
	2b	11.54	9.95	13.14	13.81	12.10	15.52	7.56	6.07	9.05	8.41	6.87	9.95
	3	11.55	9.95	13.16	13.05	11.55	14.56	7.95	6.34	9.55	9.29	7.67	10.91
	4	12.45	10.81	14.09	12.70	10.96	14.44	8.28	6.69	9.88	8.09	6.59	9.59
Compre-hension	1a	0.44	0.39	0.48	0.46	0.42	0.51						
	1b	0.65	0.59	0.71	0.59	0.53	0.66						
	2a	0.62	0.51	0.74	0.40	0.30	0.50	0.59	0.49	0.70	0.44	0.32	0.55
	2b	0.60	0.53	0.67	0.60	0.53	0.67	0.57	0.51	0.64	0.53	0.46	0.61
	3	0.55	0.50	0.61	0.52	0.47	0.58	0.56	0.50	0.63	0.57	0.50	0.63
	4	0.50	0.43	0.57	0.53	0.46	0.60	0.54	0.48	0.61	0.50	0.43	0.57
IES avoidance[b]	1a	14.60	13.06	16.13	14.96	13.42	16.49						
	1b	16.78	14.43	19.13	15.10	12.76	17.45						
	2a	13.84	10.37	17.31	16.30	13.51	19.08	13.64	9.75	17.54	12.89	10.07	15.72
	2b	17.66	15.80	19.51	19.19	17.14	21.25	13.80	11.31	16.29	14.95	12.69	17.20
	3	16.15	14.04	18.26	18.12	16.00	20.24	13.16	10.79	15.54	13.74	11.54	15.94
	4	17.35	15.24	19.46	17.73	15.89	19.56	13.34	11.10	15.58	12.36	10.42	14.29

Note. Dashes indicate subjects in Experiment 1a did not rate how negative the story was.

[a]Positive And Negative Affect Schedule negative subscale scores, made after exposure to the material (Watson & Clark, 1999; Watson, Clark, & Tellegen, 1988). [b]Impact of Event Scale intrusions and avoidance subscale scores (Horowitz, Wilner, & Alvarez, 1979).

Table 5

Raw Effect Sizes of Presence of Warning on Key Measures of Distress

Measure	Experiment	Effect of warning on negative material			Effect of warning on control material		
		M diff	95% CI LL	95% CI UL	*M* diff	95% CI LL	95% CI UL
Negative rating	1a	–	–	–			
	1b	0.05	-0.26	0.37			
	2a	0.22	-0.34	0.79	-0.18	-1.08	0.73
	2b	-0.15	-0.46	0.15	-0.25	-0.87	0.38
	3	-0.32	-0.69	0.06	-0.65	-1.16	-0.13
	4	-0.12	-0.49	0.25	-0.22	-0.79	0.34
PANAS negative[a]	1a	1.08	-0.83	2.99			
	1b	1.31	-1.10	3.71			
	2a	-0.92	-4.77	2.93	-0.14	-4.09	3.80
	2b	-1.14	-4.02	1.75	0.10	-1.69	1.88
	3	-0.86	-3.58	1.86	-0.53	-2.66	1.61
	4	-1.00	-3.92	1.93	1.89	-0.20	3.97
Intrusions tally	1a	-0.20	-1.54	1.15			
	1b	-0.73	-2.50	1.04			
	2a	2.67	-2.67	8.01	-1.63	-5.45	2.18
	2b	-0.48	-2.63	1.66	0.18	-1.59	1.94
	3	1.72	-0.28	3.72	-0.90	-2.71	0.91
	4	-0.64	-2.64	1.36	-0.96	-2.77	0.85
IES intrusions[b]	1a	-0.47	-2.16	1.23			
	1b	-0.72	-3.16	1.73			
	2a	-2.74	-6.06	0.58	1.46	-1.89	4.82
	2b	-2.26	-4.58	0.06	-0.85	-2.99	1.29
	3	-1.50	-3.73	0.73	-1.34	-3.61	0.92
	4	-0.25	-2.62	2.12	0.19	-1.98	2.36
Compre-hension	1a	-0.03	-0.09	0.04			
	1b	0.06	-0.03	0.15			
	2a	0.22	0.08	0.37	0.16	0.01	0.31
	2b	0.00	-0.09	0.10	0.04	-0.06	0.14
	3	0.03	-0.05	0.11	-0.01	-0.10	0.09
	4	-0.03	-0.13	0.07	0.04	-0.05	0.13
IES avoidance[b]	1a	-0.36	-2.52	1.80			
	1b	1.67	-1.63	4.98			
	2a	-2.46	-6.76	1.85	0.75	-3.95	5.45
	2b	-1.54	-4.29	1.21	-1.15	-4.47	2.17
	3	-1.97	-4.99	1.04	-0.57	-3.78	2.63

4	-0.38	-3.16	2.41	0.98	-1.94	3.90

Note. The effects of presence of a trigger warning (for each type of material) were calculated by subtracting "no warning" subjects' means from "warning" subjects' means; positive differences indicate higher scores for subjects who saw a trigger warning. Dashes indicate subjects in Experiment 1a did not rate how negative the story was.
[a]Positive And Negative Affect Schedule negative subscale scores, from ratings made after exposure to the material (Watson & Clark, 1999; Watson, Clark, & Tellegen, 1988). [b]Impact of Event Scale intrusions and avoidance subscale scores (Horowitz, Wilner, & Alvarez, 1979).

scale. Moreover, the confidence interval around the difference includes both 0 and values in the opposite direction among the plausible values for the true difference. In NHST terms, there was no significant effect of warning on how negative subjects felt after the story, $t(238) = 1.11$, $p = .27$. These results suggest that trigger warnings did very little either to exacerbate or mitigate how negative subjects felt immediately following a negative story.

We also analysed the degree of change in subjects' negative affect from before to after reading a negative story, by conducting a 2(warning, no warning) x 2(baseline rating, rating after) mixed ANOVA on subjects' negative affect. This analysis yielded no significant interaction between warning and time, $F(1, 238) < 0.01$, $p = .99$, and no main effect of warning, $F(1, 238) = 1.77$, $p = .18$, but there was a main effect of time, $F(1, 238) = 64.95$, $p < .001$. In short, this analysis suggests that subjects felt more negative after reading a negative story, regardless of if they had seen a trigger warning.

Intrusions. We now turn to a second symptom of distress: intrusive thoughts, which we measured in three ways. We first examined subjects' reported tally of intrusive thoughts. Subjects reported high adherence to noting intrusions

(M = 7.30, 95% CI [7.06, 7.55], on a 0-10 scale) indicating they took this task seriously. Nonetheless, there were a small number of very high tallies (maximum in this experiment = 59 intrusions) and so in this experiment, and each one to follow, we Winsorised the intrusions tally data such that the value of any tally exceeding the 95th percentile for that condition was replaced with the value of the 95th percentile of that condition (Sheskin, 2003; this technique reduces the skewing effect of extreme values in a distribution without having to remove those data points entirely). We then examined the mean number of intrusions subjects reported, according to whether they had seen a trigger warning or not, and the difference between those means—these means, and the mean difference are reported in Tables 4 and 5, respectively.

As those tables show, subjects in both conditions reported having some intrusive thoughts related to the story they had read, while they were reading the article, and "warning" subjects experienced slightly fewer intrusions than "no warning" subjects. But the difference between conditions is small—a difference of a fifth of a thought. Moreover, the confidence interval around the difference spans a range of small values, including 0. In NHST terms, there was no significant effect of warning, $t(238) = 0.29$, $p = .77$. These results suggest trigger warnings had little, if any, effect on the number of intrusions subjects experienced.

We next examined the second measure of intrusions: subjects' ratings on the IES subscale regarding the frequency of their intrusion symptoms. As for the measures above, we calculated the mean scores for "warning" and "no warning" subjects, and the mean difference between those conditions, and display those data in Tables 4 and 5. As the tables show, subjects in both conditions reported

experiencing moderately frequent intrusion symptoms since reading the story, with "warning" subjects reporting a lower frequency than "no warning" subjects. But again, the difference was small—the maximum difference on this scale is 25, and so this difference represents a 2% movement. Further, the 95% CI around this difference spans a range of small values, including 0. In NHST terms, there was no significant effect of warning, $t(238) = 0.54$, $p = .59$. These results suggest that trigger warnings had little, if any, effect on the frequency of subjects' intrusion symptoms.

We then examined the third, indirect, measure of subjects' intrusions: their performance on the comprehension test about the article they read. Recall that the more off-task, intrusive thoughts subjects have while they are attempting to read the article—particularly those they do not "catch" themselves having, and report via key-press—the worse their comprehension should be (Baird et al., 2013; Takarangi, Strange et al., 2014). We calculated the proportion of questions about the article subjects read that they answered correctly, classified by if they got a trigger warning or not, and display these means and the difference between them in Tables 4 and 5. As the tables show, subjects in both conditions did quite poorly on these questions, with "warning" subjects getting a smaller proportion of questions correct than "no warning" subjects. But the difference between conditions is very small—given the maximum difference on this measure is 1, this difference represents a 3% movement. Further, the 95% CI around this difference spans a range of small values, including 0. In NHST terms, there was no significant effect of warning, $t(238) = 0.79$, $p = .43$. These results suggest that

trigger warnings had little effect, and plausibly none, on how well subjects were able to comprehend an article they read, following a negative story.

Taken together, these three measures show that most subjects experienced some intrusion symptoms related to the negative story they read. But getting a trigger warning beforehand had only very slight effects on the frequency of those intrusions.

Avoidance. Finally, we turn to subjects' ratings on the IES subscale regarding the frequency of their avoidance symptoms. We once again examined the mean scores for "warning" and "no warning" subjects, and the difference between their scores, and display those data in Tables 4 and 5. As the tables show, subjects in both conditions reported experiencing avoidance symptoms with low to moderate frequency since reading the story, and "warning" subjects reported a lower frequency than did "no warning" subjects. But again, the effect was small—considering the maximum difference of 40 on this scale, this mean difference represents a 1% movement. Again, the 95% CI around this difference spanned a range of small values, including 0. In NHST terms, there was no significant effect of warning, $t(238) = 0.33$, $p = .74$. These results suggest that trigger warnings had little, if any, effect on subjects' avoidance symptoms.

Considering together the results for each of these measures, this experiment suggests that giving people a trigger warning prior to them reading a negative story has little effect. We found trigger warnings made little difference to how well subjects remembered the story, or to how frequently they experienced of symptoms of distress in relation to it afterward. Although many of the differences between subjects who did and did not see a warning were numerically in the

direction of trigger warnings being helpful, that is, reducing the frequency of symptoms, the absolute sizes of these effects were too small to be of practical significance. Moreover, the confidence intervals around the differences were relatively narrow, indicating the true population means were being estimated with reasonable precision, and yet these confidence intervals all included 0 within the range of values they spanned, indicating that trigger warnings may plausibly have no effect at all (Cumming, 2012).

If trigger warnings really have little or no effect on people's memory for the negative material, or on comprehension of material presented subsequently, that is good news for educators who want to use these warnings—these data suggest such warnings would not alter students' learning. It is possible that even if the true effect of trigger warnings on how well people remember details about the negative material (and on the symptoms they experience) is minuscule, warnings may nonetheless alter how negative subjects report the material to be. That is, seeing a trigger warning may lead people to reappraise the material and judge it to be more negative—such a finding would fit with research showing that false feedback given after an experience can lead people to report that that experience was more negative (Takarangi & Strange, 2010). To address this possibility, in our next experiment we made an addition to the end of the procedure: we asked subjects to rate how negative the story they read was.

Experiment 1b

The main purpose of this experiment was to replicate Experiment 1a. Therefore, the method of this experiment was the same, except for the changes noted below.

Method.

Subjects. We used Amazon's online crowdsourcing service, Mechanical Turk (MTurk; https://www.mturk.com/), to recruit 203 MTurk members, who received USD0.75 for completing the study remotely. We aimed to collect data from enough subjects that we could exclude up to 30% of our sample for failing our compliance checks (as per the highest rate of exclusions reported by Oppenheimer, Meyvis, & Davidenko, 2009) and still have 70 subjects per warning condition. The Exploratory Software for Confidence Intervals' (ESCI; Cumming, 2012) "precision for planning" feature shows that to compare two groups, with 99% assurance of achieving a margin of error of 0.35 around a standardised mean difference of 0.4—thereby allowing exclusion of 0 as a plausible value for the difference—the target *n* should be 69 subjects per group (recall that Ferguson, 2009, recommended a threshold of 0.41 for considering such an effect practically significant). We retained data from 144 subjects for analysis (our exclusion criteria are detailed below). Of those subjects, 56% were female and 44% male; their ages ranged from 19-71, Median = 35, $M = 38.33$, 95% CI [36.21, 40.46]; 99% of them reported that they were citizens of the US; and 99% of them reported that English was their first language.

Procedure. Prior to running this experiment, we collected more norming data about several negative stories. This time, we used a sample drawn from MTurk, in order to determine if the same stories would be suitable for use with this population—they were.

Norming of stories. To collect these norming data, we asked 150 MTurk members to read one of ten possible stories (two of these stories were abridged

versions of "The Veldt," the shortest of which we used in both Experiments 1a and 1b; see below). We then asked them the same questions as we asked when we collected norming data from a student sample.

Because MTurk subjects do not complete studies under controlled laboratory conditions, we included several attention checks throughout this norming study, and at the end we asked subjects questions about their compliance with instructions, and about the conditions under which they completed the study. We excluded and replaced responses from subjects who failed our attention checks, reported that they did not carefully read the entire story, or had a story-reading speed (calculated by dividing the story word count by the time they spent on the story survey page) of more than 600 words per minute. We then examined subjects' responses to each story, focusing—as we did for the student sample—on how negatively they rated the story, change in the degree of negative affect they reported feeling, how long they took to read the story, and if they had read the story before.

With regard to "A Dark Brown Dog," we found that, in norming, MTurk subjects rated this story as very negative, (M = 6.27, 95% CI [5.73, 6.80]), and likewise their degree of current negative affect increased from before to after reading this story (M_{diff} = 9.80, 95% CI_{diff} [5.06, 14.54]). Further, MTurk subjects took approximately 10 minutes to read this story (M = 610 s, 95% CI [402, 817], Median = 596 s). Finally, none of the MTurk subjects had previously read this story.

With regard to the shortest version of "The Veldt," we found that, in norming, MTurk subjects rated this story as very negative (M = 5.40, 95% CI

[4.52, 6.28]), and likewise their degree of current negative affect somewhat increased from before to after reading this story (M_{diff} = 2.93, 95% CI_{diff} [0.44, 5.43]). Further, MTurk subjects took approximately 12 minutes to read this story (*M* = 731 s, 95% CI [527, 934], Median = 661 s). Finally, only one of the MTurk subjects reported previously having read this story.

On the basis of these data, we asked subjects in Experiment 1b to read one or other of the same negative short stories as in Experiment 1a. After reading one of these stories, subjects completed the same four tasks as in Experiment 1a to measure their symptoms of distress, and their memory for the negative material. These subjects spent approximately 3 minutes reading the non-fiction article and noting their intrusions related to the negative story they had read (*M* = 206.26 s, 95% CI [184.78, 227.74], Median = 174.64 s).

But there were several changes to the procedure of this experiment from that of Experiment 1a. First, although subjects read the same negative stories, some of the four-alternative forced-choice questions we asked them about those stories were different. We used different questions because when we collected norming data about these questions using another MTurk sample, the item statistics were different than those from our student sample.

Norming of story questions. We followed a similar process to choose questions for use with an MTurk sample as we did for the student sample. We used data from 30 MTurk members (who completed the survey, passed attention checks, and reported they complied with our instructions) to examine item difficulty and reliability, and ultimately choose five questions about each story. For the chosen questions about "A Dark Brown Dog," item difficulty ranged from

0.67 to 0.93, and item reliability ranged from 0.41 to 0.57; for the chosen questions about the shortened version of "The Veldt," difficulty ranged from 0.73 to 0.93, and reliability ranged from 0.55 to 0.85. These 10 chosen questions appear in Appendix A.

The second change to the procedure of Experiment 1b was that subjects completed an additional task to measure how they experienced the story. Specifically, their final task in the experiment proper was to rate the story they had read for how negative, positive, surprising, and interesting it was, on 1 (*not at all*) to 7 (*extremely*) scales.

We also made some other changes to the procedure of Experiment 1b, due to these subjects participating remotely. These changes fell in three places: Before the experiment proper, we gave subjects instructions intended to ensure the quality of their data (for example, we told them they should not engage in other tasks during the experiment, and that they should complete the experiment in an environment free of noise and distraction). During the experiment, we incorporated several attention checks to encourage subjects to pay close attention (for example, in the PANAS-X there was an additional item that looked the same as the others but said "choose option 2"; Oppenheimer et al., 2009), although we paid subjects who reached the end of the experiment regardless of if they responded to those questions correctly. Finally, after the experiment, we included additional questions to check subjects' compliance with our instructions (for example, we asked subjects if they had read the entire story, and if they had completed the experiment in an environment free of noise and distractions); to encourage honesty we told subjects we would pay them no matter how they

responded to these questions. Appendix C details the complete wording of the instructions, attention checks, and compliance checks we used for samples recruited from MTurk.

Results & Discussion.

Exclusions. We first excluded subjects who reported having previously read the story (n = 12), or who had not correctly completed one or more critical tasks —specifically, those who had a story-reading speed greater than 1000 words per minute (calculated using the time they spent on that survey page) indicating they probably did not read the story closely, or who reported they did not read the entire story, or reported that they not did not read the entire article, or reported pressing the wrong key to note their intrusions. Altogether, we excluded 59 subjects (29%), leaving us with complete data from 144 subjects for analysis, who were somewhat unevenly distributed across conditions: n_{Warning} = 77, $n_{\text{NoWarning}}$ = 67.

Manipulation check. Next, we carried out a manipulation check. At the end of the experiment, 97% of "warning" subjects reported they remembered seeing the trigger warning. This finding suggests that, again, subjects in the "warning" condition took note of our manipulation. We then turned to our research question.

Memory and ratings of material. Recall that in this experiment, we added a task in which we asked subjects to make several ratings about the story they had read. To gauge the extent to which trigger warnings affected how negative subjects judged that material to be, we calculated their mean rating for how negative the story was, classified by if they saw a warning about it or not, and display those results in Table 4. We conducted the same calculations for the other

ratings, too; those results appear in Appendix D as Table S1. Next, we calculated the difference between these mean ratings of negativity, and display that result in Table 5. As Table 4 shows, subjects in both conditions rated the stories as very negative. But, as Table 5 shows, although “warning” subjects rated the stories as more negative than "no warning” subjects, the difference between conditions was very small—a movement of approximately 1% on this scale, with its maximum difference of 6. The confidence interval around the difference spans a range of small values, including 0, and there was no statistically significant effect of warning, $t(142) = 0.32$, $p = .75$. In an echo of the conclusions in Experiment 1a, these results suggest that trigger warnings had only a very small effect on how negative subjects judged the stories to be—suggesting, once again, that educators who use trigger warnings are doing little to change the negativity of their students’ experience.

Next, we examined the effect of trigger warnings on how well subjects remembered the story they read. We calculated the proportion of questions about the stories that subjects answered correctly, classified by if they had seen a warning beforehand or not. In both conditions the proportion of questions subjects got correct was high ($M_{\text{Warning}} = 0.75$, 95% CI [0.70, 0.80], $M_{\text{NoWarning}} = 0.78$, 95% CI [0.72, 0.83]). But the difference between conditions was negligible—a 3% movement. What is more, the 95% confidence interval around the difference is [-0.05, 0.10]; thus, it spans a range of small values, including 0. Put in NHST terms, the effect of warnings was not significant, $t(142) = 0.62$, $p = .54$. As in Experiment 1a, these results suggest that trigger warnings had little effect on subjects’ memory for the story.

Negative affect. We again found that baseline negative affect was similar for subjects later randomly assigned to see a warning as those who were not, as Table 3 shows. As for how negative subjects felt after reading the negative story, we display "warning" and "no warning" subjects' mean ratings in Table 4, and the mean difference between them in Table 5. Those tables show that "warning" subjects felt only very slightly more negative than their "no warning" counterparts did after reading a negative story—a movement of 3%—and that the confidence interval around the range of plausible differences spans 0, and values in the opposite direction. There was no significant effect of warning on how negative subjects felt after the story, $t(142) = 1.07$, $p = .29$.

Moreover, a 2(warning, no warning) x 2(baseline rating, rating after) mixed ANOVA on subjects' negative affect yielded no significant interaction between warning and time, $F(1, 142) = 0.06$, $p = .80$, and no main effect of warning, $F(1, 142) = 1.48$, $p = .23$, but there was a main effect of time, $F(1, 142) = 74.41$, $p < .001$. These results suggest that, as in Experiment 1a, trigger warnings had little effect on how much more negative subjects felt after reading a negative story.

Intrusions. Then, we examined the effect of trigger warnings on our three measures of intrusions by calculating, on each measure, "warning" and "no warning" subjects' means, and the mean difference between them (and the 95% CIs around these raw effect sizes). First, we compared the number of intrusions subjects reported while reading the article (maximum tally = 283, before we Winsorised them). Subjects reported high adherence to noting intrusions, $M = 8.75$, 95% CI [8.47, 9.03], indicating they took this task seriously. As Table 4 shows, both "warning" and "no warning" subjects experienced intrusions related

to the story. But, as Table 5 shows, the difference between their means is small—a difference of approximately three-quarters of a thought—and the true effect may plausibly be 0, or in the opposite direction. Further, there was no significant effect of warning, $t(142) = 0.82$, $p = .42$. Second, we made the same comparisons for subjects' ratings on the IES subscale regarding the frequency of their intrusion symptoms. As Tables 4 and 5 show, these data tell a similar story as the intrusion tally: Subjects in both conditions reported intrusive symptoms, but the difference between the conditions was small—a movement on this scale of only 3%—and the true effect may plausibly be 0, or in the opposite direction. There was no significant effect of warning on this measure, $t(142) = 0.58$, $p = .56$. Third, we made the same comparisons for subjects' performance on the comprehension test about the article they read, following the negative story. As Tables 4 and 5 show, subjects in both conditions did moderately well on these questions, but the difference between conditions was very small—a 6% movement—and the true effect may plausibly be 0, or slightly in the opposite direction; and there was no significant effect of warning, $t(142) = 1.35$, $p = .18$. Taken together, these results suggest trigger warnings had little effect on subjects' intrusion symptoms.

Avoidance. Finally, we examined the effect of trigger warnings on subjects' ratings on the IES subscale regarding the frequency of their avoidance symptoms, calculating the same statistics as before. Tables 4 and 5 show that—as for the measures of other symptoms—subjects in both conditions reported moderate levels of avoidance symptoms, but the difference between "warning" and "no warning" subjects was small—a 4% movement—and plausibly 0, or in the opposite direction. Further, there was no significant effect of warning, $t(142) =$

1.00, $p = .32$. That is, trigger warnings had little effect on the frequency of subjects' avoidance symptoms.

The results of Experiment 1b replicate the main findings of Experiment 1a. In Experiment 1b, trigger warnings had minimal effects on any of our measures, regarding how well subjects remembered the negative stories and the degree of distress those stories evoked in them, and warnings had similarly little effect on how negative subjects judged those stories to be.

It is possible that trigger warnings truly have a negligibly small effect on people's symptoms of distress, or on people's experience of and memory for the negative material and material presented afterward. But one alternative explanation for our observing very little effect of warnings is that our trigger warnings, or our negative materials, were not best suited for detecting these effects. There are several possible issues with the materials we used.

For a start, the wording of our trigger warnings may not have been specific enough for people to find them useful. It stands to reason that people should be better able to correct for unwanted mental influences—to the extent they are able—if they have more detailed information about what they are trying to correct for (Wilson & Brekke, 1994). In the rest of our experiments we therefore made our trigger warnings somewhat more explicit, while still retaining the form of those used in real-world situations. We changed our warnings so that they described the negative content in more detail, and also specified that people may have a negative response when exposed to that negative content.

What is more, the negative materials we used might not have been sufficient to evoke symptoms to a great enough degree for trigger warnings' effects to

become clear. That is, if trigger warnings reduce the rates of symptoms, but our story materials made "no warning" subjects only somewhat symptomatic, the "warning" subjects' symptoms could not get much lower, thereby masking any helpful effects—a problem sometimes referred to as a "floor effect." Conversely, if trigger warnings potentiate the negativity of material and the ensuing symptoms, but these stories were not negative enough to give trigger warnings enough negative content to potentiate, then any harmful effects of trigger warnings would similarly be masked.

In our subsequent experiments we therefore turned to using film clips as our warned-about material, instead of negative stories. Negative film clips are often used as trauma-analogues in research, and so we should be able to use such clips to generate sufficient symptoms of distress (for reviews of the trauma film paradigm, see Holmes & Bourne, 2008; James, Lau-Zhu, Clark, Visser, Hagenaars, & Holmes, 2016).

Moreover we used film clips with two versions, one very negative and the other—as a comparison, or control—far less negative, so that the effects of trigger warnings would become clearer. We expected that subjects in the negative condition would experience greater rates of symptoms than those in the control condition (Hall & Berntsen, 2008; Rubin, Boals, et al., 2008). Therefore, if trigger warnings need a certain amount of negativity and the ensuing high rates of symptoms to act on at a detectable level, then we should see bigger effects of trigger warnings when they precede the more negative film clips rather than the less negative, control film clips. Further, if subjects report lower levels of symptoms after seeing a control film clip than after seeing a negative clip, it

would suggest that "negative film" subjects' symptoms are not "at floor," making floor effects a less likely counterexplanation for why trigger warnings exerted little effect in our experiments.

Experiment 2a

Method.

Subjects. We aimed to recruit as many subjects as we could, within the constraints of departmental subject-hour allocations for the semester. We collected data from 130 introductory psychology students at Victoria University of Wellington, who participated in partial fulfilment of course requirements. Of the 108 whose data we retained for analysis (see below for the details of our exclusion criteria), 65% were female and 35% male, and their ages ranged from 17-37, Median = 19, M = 19.88, 95% CI [19.21, 20.55].

Design. We used a 2 x 2 factorial design, manipulating both Presence of Warning (warning, no warning) and Negativity of Material (negative film clip, control film clip) between subjects.

Procedure. The procedure of this experiment was the same as that of Experiment 1a, except for the following changes. On consent forms, we told subjects we were interested in examining visual and verbal learning, again using a cover story so as not to alert subjects that they may be exposed to negative material (prior to only some of them seeing a trigger warning regarding this material).

To measure their negative affect, subjects completed the Positive and Negative Affect Schedule (PANAS; Watson et al., 1988) as their first task, and a second time after exposure to the film clip. This measure is a shorter version of

the PANAS-X, and so quicker to complete; but the items that are used to calculate a score on the negative subscale are identical across these two versions of the scale.

Subjects who were assigned to get a trigger warning read the more specific (compared to the warnings in previous experiments) wording "TRIGGER WARNING: The following video may contain graphic footage of [a fatal car crash / violent domestic abuse]. You might find this content disturbing." The brackets indicate differences depending on which film clip counterbalance subjects were assigned to see, as described below. The "no warning" subjects skipped this step.

Next, all subjects were randomly assigned to watch one of four short film clips; two of these clips showed negative events, and the other two were the controls. The control clips each showed an event that was very similar to the one depicted in its negative clip counterpart, but which unfolded in a less negative way. All these film clips were taken from public service announcement campaigns.

More specifically, one pair of clips we used was from a 2010 TV advert campaign against speeding, by Australia's Queensland Government Department of Transport and Main Roads, called "Pram." We retrieved two clips, called "Pram1" and "Pram2," from their website (https://www.tmr.qld.gov.au/Safety/Safety-campaigns.aspx, but note that the clips are no longer hosted on that website). In Pram1, a man is speeding and can't stop in time to avoid hitting another vehicle, so he instead swerves, and hits a woman pushing a pram; we then see her lifeless body and her crying, bloodied baby. Pram2 depicts the man being

able to stop in time to avoid hitting the other vehicle; the woman walks by with her pram, unaffected.

The other pair of clips we used was from a 2013 cinema advert against domestic abuse, by the UK's Women's Aid charity, called "Blind Eye." Its two versions were designed for simultaneous projection to an audience wearing 3D glasses—delivering a different version to each eye. The versions respectively show a woman preparing dinner while being physically and verbally abused by her husband, and the woman carrying out a very similar sequence of actions, while alone. We obtained the footage directly from the charity, but their website contains an excerpt that alternates between the versions, giving a flavour of the contents of each (see https://www.womensaid.org.uk/what-we-do/blind-eye/).

We edited these clips to remove logos and make the content more similar across different clip versions. Our edited versions were all under a minute in length (Pram-negative = 52 s; Pram-control = 40 s; Blind Eye-negative = 43 s; Blind Eye-control = 45 s).

Thus, the (negative) films involved topics of the kind for which warnings are issued or requested. What is more, they depicted events which (if they really happened) would meet Criterion A for diagnosis with PTSD (see Table 1). Indeed, traffic accidents and domestic altercations are relatively common occurrences, and so should be likely to cue intrusive thoughts of any related personal experiences—as well as be generally unpleasant to watch. After all, public service campaigns are intended to reach and be relatable to wide swathes of the population. We chose to use these particular clips based on norming data.

Norming of film clips. We collected norming data about 10 negative and control film clips in a similar manner as we had about the short stories. We chose clips that came in pairs, taking them from public service announcement campaigns that used two different versions of the same event—one with a more negative or more graphic outcome than the other—to convey a common message. As described above, before norming the clips we lightly edited some of them, to remove logos and verbal campaign appeals, and in some cases we added or removed footage from within a pair to make the two clips more equivalent to one another in length. We showed a sample of introductory psychology students at Victoria University of Wellington one of the 10 clips, to obtain a total of 15 complete responses per clip. We excluded and replaced data from subjects who were observed replaying their assigned clip.

These norming subjects first completed the PANAS-X, then watched one film clip. Afterwards, they completed the PANAS-X a second time; rated how negative, positive, surprising, interesting, unpleasant, distressing, and disgusting the clip was; listed disturbing and pleasant details in the clip, summarised the storyline and message conveyed (if any); and indicated if they had seen the clip before. Given that we had edited the clips, we decided to also check subjects' feelings of transportation (that is, their feelings of absorption in a narrative) and of comprehension, so that we could pick pairs of negative and control clips that were at least moderately transporting and comprehensible. To assess those properties, subjects also rated nine items with regard to how transported they were while watching the clip (for example, "I was mentally involved in the video while watching it"), from 1 (*not at all*) to 7 (*very much*); and rated four items with

regard to how well they understood what happened in the clip, how comprehensible the order of events was, how realistic the storyline was, and how coherent a story it was, respectively (also on 1-7 scales); we then summed each of these two groups of items (which were variously adapted from: Glenberg, Wilkinson, & Epstein, 1982; Green & Brock, 2000; Johnson, Foley, Suengas, & Raye, 1988; Rubin, Schrauf, & Greenberg, 2003). For each pair of clips, we then examined subjects' responses. We primarily focused on change in the degree of negative affect that subjects reported feeling; how negative, unpleasant, disturbing, and disgusting they rated the clips as being; how transporting and comprehensible they rated the clips as being; and if subjects had seen the clips before.

The results of our norming for the "Pram" clips showed that, with regard to the negative version, subjects rated it as very negative on the four related items: $M_{negative}$ = 5.87, 95% CI [5.32, 6.42]; $M_{unpleasant}$ = 5.40, 95% CI [4.26, 6.54]; $M_{distressing}$ = 5.47, 95% CI [4.61, 6.33]; $M_{disgusting}$ = 4.47, 95% CI [3.38, 5.55]; whereas they rated the control version considerably lower on those items: $M_{negative}$ = 3.67, 95% CI [2.76, 4.57], M_{diff} = 2.20, 95% CI_{diff} [1.19, 3.21]; $M_{unpleasant}$ = 3.20, 95% CI [2.11, 4.29], M_{diff} = 2.20, 95% CI_{diff} [0.69, 3.71]; $M_{distressing}$ = 3.53, 95% CI [2.37, 4.70], M_{diff} = 1.93, 95% CI_{diff} [0.55, 3.31]; $M_{disgusting}$ = 1.93, 95% CI [1.06, 2.81], M_{diff} = 2.53, 95% CI_{diff} [1.20, 3.86]. Similarly, subjects' degree of current negative affect increased from before to after watching the negative version: M_{diff} = 4.53, 95% CI_{diff} [1.99, 7.08]. This increase is bigger than those observed in the norming of the negative stories but, considering the maximum possible change score is 40, it is still only a modest change. Nevertheless,

subjects' degree of current negative affect was only trivially different before versus after watching the control version: M_{diff} = 1.27, 95% CI_{diff} [-1.15, 3.68]. Further, subjects rated both the negative and control versions as moderately transporting, considering the possible range of summed scores was 9-63 ($M_{negative}$ = 39.60, 95% CI [34.41, 44.79], $M_{control}$ = 38.67, 95% CI [34.90, 42.43]), and as highly comprehensible, considering the possible range of summed scores was 4-28 ($M_{negative}$ = 25.53, 95% CI [24.32, 26.75], $M_{control}$ = 20.80, 95% CI [17.83, 23.77]). Finally, no subjects reported having previously seen either of these clips.

The results of our norming for the "Blind Eye" clips showed that, with regard to the negative version, subjects rated it as very negative on the four related items: $M_{negative}$ = 6.47, 95% CI [6.06, 6.88]; $M_{unpleasant}$ = 5.33, 95% CI [4.68, 5.98]; $M_{distressing}$ = 4.87, 95% CI [4.09, 5.65]; $M_{disgusting}$ = 5.47, 95% CI [4.58, 6.35]; whereas they rated the control version considerably lower: $M_{negative}$ = 2.07, 95% CI [1.49, 2.64], M_{diff} = 4.40, 95% CI_{diff} [3.73, 5.07]; $M_{unpleasant}$ = 1.93, 95% CI [1.17, 2.70], M_{diff} = 3.40, 95% CI_{diff} [2.44, 4.36]; $M_{distressing}$ = 1.87, 95% CI [1.06, 2.67], M_{diff} = 3.00, 95% CI_{diff} [1.93, 4.07]; $M_{disgusting}$ = 1.40, 95% CI [0.94, 1.86], M_{diff} = 4.07, 95% CI_{diff} [3.11, 5.02]. Similarly, subjects' degree of current negative affect increased from before to after watching the negative version, M_{diff} = 3.47, 95% CI_{diff} [-0.74, 7.68], but trivially decreased from before to after watching the control version: M_{diff} = -1.13, 95% CI_{diff} [-2.16, -0.11]. Further, subjects rated both versions as moderately transporting ($M_{negative}$ = 37.07, 95% CI [32.85, 41.29], $M_{control}$ = 36.13, 95% CI [33.02, 39.24]), and highly comprehensible ($M_{negative}$ = 21.60, 95% CI [18.95, 24.25], $M_{control}$ = 20.47, 95% CI [17.41, 23.52]). Finally, no subjects reported having previously seen either of these clips.

After subjects in Experiment 2a had watched one of these four film clips, we then measured their experience of the film material and their symptoms of distress, in a similar manner as in the previous experiments. Subjects spent approximately 4 minutes reading the unrelated, non-fiction article and noting intrusions related to the film clip they had seen (M = 234.17 s, 95% CI [217.82, 250.53], Median = 230.91 s). They later answered comprehension questions about that article, but we did not ask subjects additional questions testing how well they remembered the film clip they had seen. At the end of the experiment we did, however, ask subjects to make ratings about the film clip they had seen: subjects rated how negative, positive, surprising, interesting, unpleasant, distressing, and disgusting they found the clip, on 1 (*not at all*) to 7 (*extremely*) scales.

After the experiment proper, subjects answered a few questions to establish whether they should be excluded (such as if they had seen the film clip before) and reported their demographics. These questions were essentially the same as the ones we asked at the end of Experiment 1a, but their wording was adapted where needed in order to be about film clips. The full list of exclusion criteria appears below. Appendix B details the wording of all of the end-of-experiment questions we asked subjects drawn from a student population.

Results & Discussion.

Exclusions. We excluded subjects who reported having previously seen the film clip (n = 4), or who had not correctly completed one or more critical tasks. Specifically, we excluded data from subjects who were not able to complete the experiment because they ran out of time, or who reported or were observed playing the film clip more than once, or reported looking away from the film clip,

or pressing the wrong key to note their intrusions.We also excluded two subjects for, respectively, having a disability such that the experimenter needed to operate the computer on that subject's behalf, and overhearing this exchange. Altogether, we excluded 22 subjects (17%), leaving us with data for analysis from 108 subjects, distributed fairly evenly across conditions: $n_{\text{WarningNegative}} = 25$, $n_{\text{NoWarningNegative}} = 27$, $n_{\text{WarningControl}} = 28$, $n_{\text{NoWarningControl}} = 28$.

Manipulation check. Our manipulation check revealed that, at the end of the experiment, 81% of "warning" subjects reported they remembered seeing the trigger warning. This result suggests most subjects in the "warning" condition took note of our manipulation.

Ratings of material. How much did trigger warnings affect how negative subjects thought the material was? To address this question, we first calculated subjects' mean rating for how negative the film clip they saw was, classified by if they got a trigger warning about it beforehand or not, and by whether they had seen a negative film clip or a control film clip. We display those results in Table 4 (and the equivalent results for the other ratings can be found in Appendix D, in Table S1). We then calculated the mean difference between "warning" and "no warning" subjects' ratings of negativity for those who had seen a negative film, and did the same for those who had seen a control film. We display those results in Table 5. Similarly, we calculated the differences between "negative film" and "control film" subjects' ratings, for subjects who got a warning and for subjects who got no warning. We display those results in Table 6. Thus, Table 5 shows the effects of the trigger warning manipulation, and Table 6 shows the effects of the film negativity manipulation.

Table 6

Raw Effect Sizes of Negativity of Material on Key Measures of Distress

Measure	Experiment	Effect of negativity after warning			Effect of negativity after no warning		
		M diff	95% CI LL	95% CI UL	*M* diff	95% CI LL	95% CI UL
Negative rating	1a						
	1b						
	2a	3.56	2.78	4.33	3.15	2.39	3.92
	2b	3.68	3.19	4.16	3.58	3.10	4.07
	3	4.21	3.79	4.63	3.88	3.42	4.34
	4	3.70	3.23	4.17	3.60	3.12	4.07
PANAS negative[a]	1a						
	1b						
	2a	3.88	0.27	7.50	4.66	0.50	8.82
	2b	7.15	4.71	9.59	8.39	5.99	10.78
	3	6.19	3.71	8.67	6.52	4.04	9.00
	4	6.02	3.20	8.84	8.91	6.60	11.21
Intrusions tally	1a						
	1b						
	2a	3.75	-1.40	8.91	-0.55	-4.54	3.44
	2b	2.05	0.01	4.08	2.71	0.80	4.62
	3	2.76	0.72	4.80	0.13	-1.63	1.89
	4	1.85	0.03	3.68	1.53	-0.46	3.53
IES intrusions[b]	1a						
	1b						
	2a	0.37	-2.94	3.68	4.58	1.21	7.95
	2b	3.98	1.81	6.15	5.39	3.11	7.67
	3	3.61	1.32	5.90	3.76	1.57	5.96
	4	4.17	1.89	6.45	4.61	2.33	6.89
Compre-hension	1a						
	1b						
	2a	0.03	-0.12	0.18	-0.04	-0.18	0.11
	2b	0.03	-0.07	0.12	0.07	-0.03	0.16
	3	-0.01	-0.09	0.08	-0.04	-0.13	0.04
	4	-0.04	-0.14	0.05	0.03	-0.07	0.13
IES avoidance[b]	1a						
	1b						
	2a	0.20	-4.94	5.34	3.40	-0.47	7.28
	2b	3.86	0.82	6.90	4.25	1.23	7.27
	3	2.98	-0.17	6.14	4.38	1.35	7.42

4	4.01	0.96	7.07	5.37	2.72	8.02

Note. Negativity of material was only manipulated in experiments that used film materials. The effects of film negativity (in each warning condition) were calculated by subtracting "control" subjects' means from "negative" subjects' means; positive differences indicate higher scores for subjects who watched a negative film.
[a]Positive And Negative Affect Schedule negative subscale scores, from ratings made after exposure to the material (Watson & Clark, 1999; Watson, Clark, & Tellegen, 1988). [b]Impact of Event Scale intrusions and avoidance subscale scores (Horowitz, Wilner, & Alvarez, 1979).

As Table 5 shows, "warning" subjects who saw a negative film rated it as only very slightly more negative than their "no warning" counterparts. Similarly, "warning" subjects who saw a control film rated it as only very slightly less negative than their "no warning" counterparts. In other words, the differences between warning conditions were very small, for either type of film—movements of 4% for the "negative film" subjects, and 3% for the "control film" subjects. Further, the confidence intervals around those differences include a range of small values as plausible, spanning 0, and values in the opposite direction.

Although the effects of trigger warnings were small, the effect of film negativity was clear. As Table 6 shows, regardless of if they saw a warning first, subjects who saw a negative film rated it as much more negative than did subjects who saw a control film—movements of 59% for "warning" subjects, and 53% for "no warning" subjects.

Put in NHST terms, there was no significant interaction between presence of warning and negativity of material on ratings of negativity, $F(1, 104) = 0.55$, $p = .46$; and there was no significant main effect of warning, $F(1, 104) = 0.01$, $p = .93$; but there was a significant main effect of film negativity, $F(1, 104) = 152.87$, $p < .$

001, whereby subjects who saw a negative film clip rated it as more negative than those who saw a control one. Thus, these new, film materials performed as expected, even as trigger warnings again had little effect on how negative subjects judged material to be.

Negative affect. How much did trigger warnings affect how negative subjects felt? As for previous experiments, to answer this question we first established that baseline negative affect was similar across conditions; it was—see Table 3. Then, we calculated subjects' mean rating of negative affect after watching a film clip, classified by if they saw a trigger warning or not and whether they saw a negative or control film, and display those in Table 4, Further, we calculated the mean differences between the negative affect of "warning" and "no warning" subjects, and display those in Table 5, and the mean differences between "negative" and "control" subjects in Table 6.

As Table 5 shows, the answer to this question is: not much. "Warning" subjects felt only very slightly less negative than their "no warning" counterparts after watching a negative film clip—a 2% movement—and similarly so after watching a control film clip—a movement of less than 1%. Further, the confidence intervals around both effects are wide, spanning 0, and values in the opposite direction.

Looking at Table 6, we can also see that, regardless of whether they had seen a warning, "negative" subjects reported feeling more negative affect after watching their allotted film clip than did "control" subjects—movements of 10% for "warning" subjects, and 12% for "no warning" subjects.

Moreover, a 2(warning, no warning) x 2(negative film, control film) x 2(baseline rating, rating after) mixed ANOVA on subjects' negative affect yielded no significant interaction between warning, film negativity, and time, $F(1, 104) = 0.01$, $p = .94$, or between warning and film negativity, $F(1, 104) = 0.13$, $p = .72$, or between warning and time, $F(1, 104) = 0.22$, $p = .64$. But there was a significant interaction between film negativity and time, $F(1, 104) = 40.33$, $p < .001$, whereby subjects who saw a negative film clip reported a greater increase in negative affect than did subjects who watched a control film clip. There was also a significant main effect of time, $F(1, 104) = 43.34$, $p < .001$, such that subjects overall felt worse after watching a clip. But there were no significant main effects of warning, $F(1, 104) = 0.39$, $p = .54$, or of film negativity, $F(1, 104) = 1.27$, $p = .26$. Together, these findings suggest that manipulating the negativity of the film subjects saw had the expected effect on how negative they felt, whereas trigger warnings once again had little effect.

Intrusions. How much did trigger warnings affect how many intrusions subjects experienced? We answered this question in three ways. First, we calculated the mean number of intrusions subjects reported while reading the non-fiction article **(**maximum tally = 55, before we Winsorised them), classified by if they saw a trigger warning or not, and whether they saw a negative or control film, and display those means in Table 4. Subjects reported high adherence to noting these intrusions, $M = 7.81$, 95% CI [7.43, 8.20], indicating they took this task seriously. We then calculated the mean differences between the tallies of intrusions reported by "warning" and "no warning" subjects, and display those in Table 5, and between "negative" and "control" subjects, and display those in

Table 6. Second, we conducted these same calculations for subjects' rating of the frequency of their intrusion symptoms. Third, we conducted these same calculations for subjects' performance on the comprehension questions about the non-fiction article.

As Table 5 shows, "warning" subjects reported a higher number of intrusions than "no warning" subjects—approximately two and two-thirds of a thought more—after a negative film clip, and a lower number of intrusions—approximately one and two-thirds of a thought fewer—after a control film clip. Yet, importantly, the confidence intervals around both these differences are very wide, spanning a range of plausible values, including 0 and values in the opposite direction. Table 5 further shows that, conversely, "warning" subjects rated their intrusions as less frequent than "no warning" subjects after watching a negative film clip—an 11% movement—but as little more frequent after watching a control film clip—a 6% movement. But, again, the confidence intervals around these differences span a wide range of plausible values, including 0 and values in the opposite direction. As for comprehension, Table 5 shows that "warning" subjects correctly answered a similarly higher proportion of questions about the article they had read than did "no warning" subjects in both film conditions—a 22% movement for "negative" subjects, and a 16% movement for "control" subjects. Although the confidence intervals around those differences do not quite include 0 as a plausible value for the difference, they are nonetheless very wide, spanning a range of values from trivial to notable. Considered together, these results suggest trigger warnings affected subjects self-reported intrusions very little, but may

have reduced the incidence of intrusions of which subjects were not meta-aware. But, given the imprecision of these estimates, this result requires replication.

Turning next to Table 6, to examine the effect of film negativity, reveals that "negative" subjects reported a higher number of intrusive thoughts than "control" subjects after a trigger warning—three and three-quarters of a thought more—but a lower number without having seen a trigger warning—approximately half a thought less. Further, "negative" subjects reported a higher frequency of intrusive symptoms than "control" subjects—a 1% movement for "warning" subjects, and an 18% movement for "no warning" subjects. And finally, "negative" subjects performed slightly better than "control" subjects after a trigger warning—a 3% movement—but slightly worse without a trigger warning—a 4% movement. But, again, the confidence intervals around these differences are very wide, and most include 0, indicating the estimates of these effects are not very precise.

We also conducted NHST analyses on these three measures. With regard to the effect on the number of intrusions subjects reported, there was no significant interaction between presence of warning and negativity of material, $F(1, 104) = 1.77$, $p = .19$; there was no significant main effect of warning, $F(1, 104) = 0.10$, $p = .75$; and no significant main effect of film negativity, $F(1, 104) = 0.98$, $p = .32$. For subjects' ratings of the frequency of intrusions, there was no significant interaction between presence of warning and negativity of material, $F(1, 104) = 3.18$, $p = .08$; and there was no significant main effect of warning, $F(1, 104) = 0.29$, $p = .59$; but there was a significant main effect of film negativity, $F(1, 104) = 4.41$, $p = .04$, such that subjects who saw a negative film clip reported more frequent intrusions than those who saw a control clip. Finally, for subjects'

performance on the comprehension questions about the article, there was no significant interaction between presence of warning and negativity of material, $F(1, 104) = 0.40$, $p = .53$; there was a significant main effect of warning, $F(1, 104) = 13.00$, $p < .01$, whereby subjects who saw a trigger warning got more questions right than those who saw no warning; but there was no significant main effect of film negativity, $F(1, 104) < 0.01$, $p = .97$. The inconsistent results across these three measures, regarding the effects of our manipulations, are best understood in the context of how wide the confidence intervals around them are. Those intervals all suggest these results provide us with poor estimates of the true effect sizes.

Avoidance. How much did trigger warnings affect how avoidant people were? To answer this question, we calculated the mean frequency of subjects' avoidance symptoms after watching a film clip, classified by if they saw a trigger warning or not and whether they saw a negative or control film. We display those means in Table 4. We also calculated the mean differences between the avoidance symptom frequency of "warning" and "no warning" subjects, and display those in Table 5, and between "negative" and "control" subjects, and display those in Table 6.

Table 5 shows that the answer is not much: "warning" subjects were similarly avoidant as their "no warning" counterparts, a little less after watching a negative film clip—a 6% movement—and a very little more after watching a control film clip—a 2% movement. But, as for the other measures in this experiment, the confidence intervals around these differences are wide, spanning 0 and including values in the opposite direction.

Table 6 shows that subjects who saw a negative film reported slightly more frequent avoidance symptoms than subjects who saw a control film—movements of 1% for "warning" subjects, and 9% for "no warning" subjects. But, once again, the confidence intervals around these estimates are wide.

Put in NHST terms, there was no significant interaction between presence of warning and negativity of material on subjects' reported frequency of avoidance, $F(1, 104) = 1.01$, $p = .32$; nor was there a significant main effect of warning, $F(1, 104) = 0.29$, $p = .59$; nor was there a significant main effect of film negativity, $F(1, 104) = 1.27$, $p = .26$.

Looking across all our measures in this experiment tells a fairly consistent story: More negative materials mostly produced greater rates of symptoms of distress whereas trigger warnings had little effect on those rates, but the confidence intervals around many of those effect sizes were very wide. The wide confidence intervals suggest that we estimated the effects of both trigger warnings and negativity of material with very low precision, meaning we cannot be sure of their true sizes. What is more, many of the confidence intervals spanned positive values, zero, and negative values, meaning we cannot even be sure of the direction of those effects.

A key reason for this lack of precision is our low sample size. Although we collected as many subjects as we could, an administrative problem with the allocation of student subject pool hours that semester meant that we were able to recruit relatively few subjects. We therefore replicated this experiment, but collected data from a larger sample to increase the precision of our estimates.

Experiment 2b

The purpose of this experiment was to replicate Experiment 2a. It therefore followed the same method as Experiment 2a, apart from the changes as noted below.

Method.

Subjects. We used MTurk to collect data from 395 subjects, who received USD0.25 for completing the study online. As for Experiment 1b, we aimed to recruit enough subjects that we could exclude up to 30% of our sample yet still have 70 subjects in each of the four conditions—a number in line with the result from ESCI software's precision for planning feature, as described above (Cumming, 2012). We retained data from 279 subjects for analysis (our exclusion criteria are detailed below). Of those subjects, 56% were female and 44% male; their ages ranged from 19-76, Median = 33, M = 37.14, 95% CI [35.60, 38.68]; 94% of them reported that they were citizens of the US; and 97% of them reported that English was their first language.

Procedure. Prior to running this experiment, we collected norming data about the same film clips from a MTurk sample, in order to decide whether we would use the same clips as in Experiment 2a. Indeed, we did.

Norming of film clips. We showed 150 MTurk members one of 10 film clips, obtaining a total of 15 complete responses per clip. These subjects completed the same measures as the student norming sample. But these subjects also got several attention checks during the survey, and at the end answered several other questions about their compliance with instructions and the conditions under which they completed the study. We excluded and replaced data from MTurk subjects who failed our attention checks, reported that they did not carefully watch the

entire film clip, that they paused or replayed the clip, or that they had technical issues playing it. As before, in examining these norming data to choose our clips, we primarily focused on change in the degree of negative affect that subjects reported; how negative, unpleasant, disturbing, and disgusting they rated the clips as being; how transporting and comprehensible they rated the clips as being; and if they had previously seen the clips. Based on these data, we again chose the "Pram" and "Blind Eye" clips.

The results of this norming of the "Pram" clips showed that subjects rated the negative version as very negative on the four related items: $M_{negative}$ = 6.40, 95% CI [6.05, 6.75]; $M_{unpleasant}$ = 6.40, 95% CI [5.99, 6.81]; $M_{distressing}$ = 6.40, 95% CI [5.94, 6.86]; $M_{disgusting}$ = 4.73, 95% CI [3.72, 5.75]; whereas they rated the control version lower on those items: $M_{negative}$ = 3.73, 95% CI [2.97, 4.50], M_{diff} = 2.67, 95% CI_{diff} [1.86, 3.47]; $M_{unpleasant}$ = 3.80, 95% CI [2.86, 4.74], M_{diff} = 2.60, 95% CI_{diff} [1.62, 3.58]; $M_{distressing}$ = 4.87, 95% CI [4.01, 5.73], M_{diff} = 1.53, 95% CI_{diff} [0.60, 2.46]; $M_{disgusting}$ = 2.27, 95% CI [1.53, 3.01], M_{diff} = 2.47, 95% CI_{diff} [1.27, 3.66]. Similarly, subjects' degree of current negative affect increased from before to after watching the negative version: M_{diff} = 7.87, 95% CI_{diff} [4.02, 11.71]. This increase is bigger than that we observed in the norming of the negative stories but, considering the maximum possible change score is 40, it is still only a relatively modest change. Nevertheless, subjects' degree of current negative affect was only trivially different before versus after watching the control version: M_{diff} = 2.07, 95% CI_{diff} [-1.13, 5.26]. Further, subjects rated both the negative and control versions as moderately transporting, considering the possible range of sum scores was 9-63 ($M_{negative}$ = 46.87, 95% CI [42.97, 50.76], $M_{control}$ =

40.73, 95% CI [37.01, 44.46]) and as highly comprehensible, considering the possible range of sum scores was 4-28 ($M_{negative}$ = 24.87, 95% CI [23.39, 26.34], $M_{control}$ = 17.67, 95% CI [13.96, 21.37]). Finally, no subject reported having previously seen either of the clips.

The results of this norming of the "Blind Eye" clips showed that subjects who saw the negative version rated it as very negative: $M_{negative}$ = 6.80, 95% CI [6.49, 7.11]; $M_{unpleasant}$ = 6.33, 95% CI [5.79, 6.87]; $M_{distressing}$ = 5.80, 95% CI [5.04, 6.56]; $M_{disgusting}$ = 6.07, 95% CI [5.33, 6.81]; whereas those who saw the control version rated it as less so: $M_{negative}$ = 1.87, 95% CI [1.24, 2.49], M_{diff} = 4.93, 95% CI_{diff} [4.27, 5.60]; $M_{unpleasant}$ = 2.33, 95% CI [1.50, 3.16], M_{diff} = 4.00, 95% CI_{diff} [3.06, 4.94]; $M_{distressing}$ = 2.20, 95% CI [1.44, 2.96], M_{diff} = 3.60, 95% CI_{diff} [2.57, 4.63]; $M_{disgusting}$ = 1.73, 95% CI [1.06, 2.41], M_{diff} = 4.33, 95% CI_{diff} [3.38, 5.29]. Similarly, subjects' degree of current negative affect increased from before to after watching the negative version, M_{diff} = 5.60, 95% CI_{diff} [1.90, 9.30], but was only trivially different before versus after watching the control version: M_{diff} = 1.40, 95% CI_{diff} [-0.83, 3.63]. Further, subjects rated both versions as moderately transporting ($M_{negative}$ = 41.60, 95% CI [36.00, 47.20], $M_{control}$ = 36.73, 95% CI [31.27, 42.20]), and highly comprehensible ($M_{negative}$ = 24.87, 95% CI [22.68, 27.05], $M_{control}$ = 20.53, 95% CI [16.93, 24.14]). Finally, no subject reported having previously seen either of the clips.

Therefore, subjects in Experiment 2b watched one of these four clips. Afterward, they spent approximately 3 minutes reading the non-fiction article and noting intrusions related to the film clip they had seen (M = 187.70 s, 95% CI [173.44, 201.95], Median = 157.19 s).

The changes we made from Experiment 2a for this experiment were all to do with adapting it for administration to a remote sample (and are therefore very similar to the changes that happened from Experiment 1a to Experiment 1b). Before the experiment proper, we gave subjects instructions intended to ensure the quality of their data. During the experiment, we incorporated several attention checks to encourage subjects to pay close attention (although we paid subjects who reached the end of the experiment regardless). Finally, after the experiment, we included additional questions to check subjects' compliance with our instructions; to encourage honesty we told subjects we would pay them no matter how they responded to these questions. Appendix C details the complete wording of the instructions, attention checks, and compliance checks we used for samples recruited from MTurk.

Results & Discussion.

Exclusions. We excluded subjects who had not correctly completed one or more critical tasks (no subject reported having previously seen the film clip). Specifically, we excluded subjects if they stayed on the film clip survey page for a shorter duration than the length of the clip they were assigned to watch, or if they reported they did not watch the entire clip, looked away from the clip, paused the clip, or replayed the clip, or if they reported not reading the entire article, or they reported pressing the wrong key to note their intrusions. Altogether, we excluded 116 subjects (29%), leaving us with a distribution of subjects across conditions as follows: $n_{\text{WarningNegative}} = 70$, $n_{\text{NoWarningNegative}} = 72$, $n_{\text{WarningControl}} = 64$, $n_{\text{NoWarningControl}} = 73$.

Manipulation check. At the end of the experiment, 93% of "warning" subjects reported they remembered seeing the trigger warning, suggesting subjects in the warning conditions took note of our manipulation.

Ratings of material. How much did trigger warnings affect how negative subjects judged the material was? To address this question, as for Experiment 2a, we calculated subjects' mean rating of how negative the film clip was, classified by which conditions they were in, and display those means in Table 4 (these same results for the other ratings can be found in Appendix D, Table S1). As before, we then calculated the mean differences between conditions, and display those results in Tables 5 and 6.

As Table 5 shows, "warning" subjects gave only slightly lower ratings of negativity than their "no warning" counterparts, regardless of which film condition they were in—a movement of 3% for the "negative film" subjects, and 4% for the "control film" subjects. The confidence intervals around those differences are narrower than in Experiment 2a, yet the ranges of values they span still include 0 and values in the opposite direction, indicating that trigger warnings had little effect on how negative subjects thought the films were.

As Table 6 shows, although trigger warnings had little effect, subjects who saw a negative film rated it as much more negative than subjects who saw a control film rated it—movements of 61% for "warning" subjects, and 60% for "no warning" subjects.

Put in NHST terms, there was no significant interaction between presence of warning and negativity of material on ratings of negativity, $F(1, 275) = 0.07$, $p = .79$; and there was no significant main effect of warning, $F(1, 275) = 1.32$, $p = .25$;

but there was a significant main effect of film negativity, $F(1, 275) = 433.57$, $p < .001$, whereby subjects who saw a negative film clip rated it as more negative than those who saw a control clip.

Negative affect. How much did trigger warnings affect how negative subjects felt? As for Experiment 2a, having established that baseline negative affect was similar across conditions (it was; see Table 3), we calculated subjects' mean ratings of negative affect after watching a film clip, classified by which conditions they were in, and then calculated the mean differences between conditions, and display those results in Tables 4, 5, and 6.

As Table 5 shows, "warning" subjects felt similarly negative as their "no warning" counterparts, slightly less after watching a negative film clip—a 3% movement—and very slightly more after watching a control film clip—a movement of less than 1%. The confidence intervals around these effects are narrower than in Experiment 2a, but they both still span 0, and include values in the opposite direction. These results suggest trigger warnings had little effect on how negative subjects felt, regardless of the negativity of the film they watched.

But, as Table 6 shows, our film materials had the expected effect: "negative" subjects reported feeling more negative affect after watching the film clip than did "control" subjects—movements of 18% for "warning" subjects, and 21% for "no warning" subjects.

We also conducted a 2(warning, no warning) x 2(negative film, control film) x 2(baseline rating, rating after) mixed ANOVA on subjects' negative affect. This NHST analysis yielded no significant interaction between warning, film negativity, and time, $F(1, 275) = 0.21$, $p = .65$, or between warning and film

negativity, $F(1, 275) = 0.46$, $p = .50$, or between warning and time, $F(1, 275) = 0.23$, $p = .63$. But there was a significant interaction between film negativity and time, $F(1, 275) = 88.39$, $p < .001$, whereby subjects who saw a negative film reported a greater increase in negative affect over time than did subjects who watched a control film. There were also significant main effects of time, $F(1, 275) = 170.51$, $p < .001$, such that subjects overall felt worse after watching a film; and of film negativity, $F(1, 275) = 40.87$, $p < .001$, such that subjects who saw a negative film clip overall felt worse. But there was no significant main effect of warning, $F(1, 275) = 1.18$, $p = .28$.

Intrusions. How much did trigger warnings affect the degree of intrusions people experienced? As for Experiment 2a, we examined three measures to answer this question. For each one in turn, we calculated subjects' mean scores, classified by which conditions they were in, and display those in Table 4, as well as the mean differences between conditions, and we display those results in Tables 5 and 6.

First, we considered subjects' tally of intrusions they reported while reading the article **(**maximum tally = 156, before we Winsorised them); subjects reported high adherence to noting intrusions, $M = 8.61$, 95% CI [8.40, 8.83], indicating they took this task seriously. As Table 5 shows, "warning" subjects reported slightly fewer intrusions than their "no warning" counterparts after watching a negative film clip—a difference of approximately half a thought—and slightly more after watching a control film clip—a difference of approximately a fifth of a thought. Second, we considered subjects' ratings of their intrusions symptom frequency. Looking again at Table 5, we see that "warning" subjects reported

slightly less frequent intrusions than their "no warning" counterparts, after watching a negative film clip—a 9% movement—and after watching a control film clip—a 3% movement. Third, we considered subjects' performance on the comprehension questions about the article. As Table 5 shows, "warning" subjects performed similarly to their "no warning" counterparts, after watching a negative film clip—a movement of less than 1%—and after watching a control film clip—a 4% movement. Looking at Table 5 also reveals that, for each of these measures, the confidence intervals around the estimated effects of trigger warnings are narrower than in Experiment 2a. Nevertheless, they all still span a range of plausible values that includes 0 and values in the opposite direction. Together, then, these measures suggest trigger warnings did not affect subjects' intrusion symptoms in a meaningful way.

Turning now to Table 6, we can see that the negativity of the film subjects saw had bigger effects on these three measures, and the confidence intervals around these effects are narrower than in Experiment 2a. More specifically, "negative" subjects reported a greater number of intrusions than did "control" subjects—approximately two more thoughts for "warning" subjects, and approximately two and two-thirds of a thought more for "no warning" subjects. Similarly, "negative" subjects reported their intrusions were more frequent than "control" subjects—movements of 16% for "warning" subjects, and 22% for "no warning" subjects. "Negative" subjects performed similarly to "control" subjects on the comprehension test, however—movements of 3% for "warning" subjects, and 7% for "no warning" subjects. These results suggest that, for the most part,

changing the negativity of the film subjects saw had the expected effect on their rates of intrusions.

Likewise, NHST analyses of these measures revealed that, with regard to effects on the number of intrusions subjects reported, there was no significant interaction between presence of warning and negativity of material, $F(1, 275) = 0.22$, $p = .64$; and there was no significant main effect of warning, $F(1, 275) = 0.05$, $p = .83$; but there was a significant main effect of film negativity, $F(1, 275) = 11.36$, $p < .01$, such that subjects who saw a negative film clip reported a greater number of intrusions than those who saw a control one. For subjects' ratings of the frequency of intrusions, there was no significant interaction between presence of warning and negativity of material, $F(1, 275) = 0.78$, $p = .38$; and there was no significant main effect of warning, $F(1, 275) = 3.79$, $p = .05$; but there was a significant main effect of film negativity, $F(1, 275) = 34.38$, $p < .001$, such that subjects who saw a negative film clip reported more frequent intrusions than those who saw a control one. Finally, for subjects' performance on the comprehension questions about the article, there was no significant interaction between presence of warning and negativity of material, $F(1, 275) = 0.30$, $p = .58$; there was no significant main effect of warning, $F(1, 275) = 0.40$, $p = .53$; and there was no significant main effect of film negativity, $F(1, 275) = 1.84$, $p = .18$.

Avoidance. How much did trigger warnings affect how much avoidance people experienced? Once again, to address this question, we calculated subjects' mean frequency of avoidance symptoms after watching a film clip, classified by which conditions they were in, and display those in Table 4, as well as the mean differences between conditions, which we display in Tables 5 and 6.

As Table 5 shows, “warning” subjects reported only slightly lower levels of avoidance than their “no warning” counterparts, after either type of film clip—a 4% movement after a negative clip, and a 3% movement after a control clip. Further, the confidence intervals around these effects were narrower than in Experiment 2a, yet still span 0 and values in the opposite direction, indicating that trigger warnings had no meaningful effect on how much avoidance subjects experienced.

Table 6 shows that, as expected, subjects who saw a negative film reported more frequent avoidance symptoms than did subjects who saw a control film—movements of 10% for “warning” subjects, and 11% for “no warning” subjects.

In line with those results, NHST analyses revealed no significant interaction between presence of warning and negativity of material on subjects’ reported frequency of avoidance, $F(1, 275) = 0.03$, $p = .86$; and no significant main effect of warning, $F(1, 275) = 1.53$, $p = .22$; but there was a significant main effect of film negativity, $F(1, 275) = 13.94$, $p < .01$, such that subjects who saw a negative film clip reported more avoidance than those who saw a control clip.

Taken together, the results of this experiment lead to the conclusion that although we were successful in choosing negative films that generally elicited more distress than the control films, our more specific trigger warnings had very little effect on subjects’ distress in either case. As intended, the confidence intervals around the estimated effect sizes of trigger warnings and film negativity were narrower than in Experiment 2a, meaning we were estimating the size of these effects with greater precision in this experiment. But even with this

increased precision, the intervals spanning the plausible effect sizes of trigger warnings all still included zero.

One interpretation of these findings is that trigger warnings are ineffective. But counterexplanations for why trigger warnings had little—if any—effect remain. One such possibility is that people in our experiments simply did not notice our trigger warnings. But we have evidence that that supposition is not true: At the end of each experiment, the vast majority of subjects in the "warning" conditions reported that they remember seeing a trigger warning prior to the story or film.

Perhaps, then, the issue is that although subjects see the trigger warning, this warning does not change their expectations about what the material to follow will be like. As a result, there are no expectancy-produced changes in the rates of symptoms that "warning" subjects experience for us to detect. Relatedly, if people see the trigger warning, but it does not lead them to believe that the material to follow (in our case, a story or a film clip) will be negative to the point of being distressing, then they may decide not to expend the effort to proactively adopt strategies to reduce their negative emotions. Alternatively, they may not be concerned enough to become overly attentive to the negative aspects of the material. Either way, such trigger warnings would yield little effect.

Our next experiment addressed this possibility by asking subjects to rate what they thought the material to follow would be like, before they had actually been exposed to it—but after those in the "warning" condition had seen the trigger warning. If our trigger warnings do not change what subjects think the films will be like, then these pre-exposure ratings should be similar for "warning" and "no

warning" subjects. But if the warnings change subjects' expectations, then "warning" subjects should expect that the film they will see will be more negative, distressing, and so on than "no warning" subjects.

Experiment 3

Method.

Subjects. We used MTurk to recruit 460 subjects, who received USD0.25 for completing the study. As for Experiment 2b, we aimed to collect data from enough subjects that we could exclude up to 30% of our sample for failing our compliance checks and still have 70 subjects in each of the four cells. We retained data from 317 subjects for analysis (our exclusion criteria are noted below). Of those subjects, 62% were female and 38% male; their ages ranged from 18-72, Median = 33, M = 36.25, 95% CI [34.92, 37.59]; 93% of them reported that they were citizens of the US; and 98% of them reported that English was their first language.

Procedure. The design and procedure of this experiment were identical to that of Experiment 2b, except for the following addition: After all subjects had been told they were about to watch a film clip, and "warning" subjects had seen a trigger warning—but prior to any of them actually watching their assigned film clip—all subjects rated what they thought that film clip was going to be like. Specifically, they rated how negative, positive, surprising, interesting, unpleasant, distressing, and disgusting they expected it to be, from 1 (*not at all*) to 7 (*extremely*). In this experiment, subjects afterward spent approximately 3 minutes reading the non-fiction article and noting intrusions related to the film clip they had seen, M = 165.40 s, 95% CI [151.98, 178.82], Median = 142.22 s.

Results & Discussion.

Exclusions. We excluded subjects who reported having previously seen the film clip ($n = 4$), or who had not correctly completed one or more critical tasks—using the same criteria as in Experiment 2b. Altogether, we excluded 143 subjects (31%), leaving us with a distribution of subjects as follows: $n_{\text{WarningNegative}} = 94$, $n_{\text{NoWarningNegative}} = 74$, $n_{\text{WarningControl}} = 73$, $n_{\text{NoWarningControl}} = 76$).

Manipulation checks. At the end of this experiment, 90% of "warning" subjects reported they remembered seeing the trigger warning, suggesting subjects in the "warning" condition took note of our manipulation.

Moreover, recall that in this experiment, subjects twice rated the negativity of the film clip they saw—the first time they made this rating was prior to seeing a clip, but after some of them had seen a trigger warning about that clip. To what extent did a trigger warning change subjects' expectations about the negativity of the material to follow? To answer this question, we calculated subjects' pre-exposure mean rating of negativity, classified by if they had just read a trigger warning (but collapsing across the film negativity conditions, because that manipulation had yet to be introduced). We found that, initially, "warning" subjects thought the film would be far more negative than did their "no warning" counterparts: $M_{\text{Warning}} = 5.63$, 95% CI [5.45, 5.81], $M_{\text{NoWarning}} = 3.15$, 95% CI [2.94, 3.36], 95% CI_{diff} [2.20, 2.75]—a 41% movement. We further found that, compared to their "no warning" counterparts, "warning" subjects thought the film would be more distressing, $M_{\text{Warning}} = 5.16$, 95% CI [4.96, 5.37], $M_{\text{NoWarning}} = 2.66$, 95% CI [2.44, 2.88], 95% CI_{diff} [2.20, 2.81]; more disgusting, $M_{\text{Warning}} = 4.69$, 95% CI [4.47, 4.91], $M_{\text{NoWarning}} = 2.47$, 95% CI [2.23, 2,72], 95% CI_{diff} [1.89,

2.54]; and more unpleasant, $M_{Warning}$ = 5.38, 95% CI [5.16, 5.59], $M_{NoWarning}$ = 2.96, 95% CI [2.74, 3.18], 95% CI_{diff} [2.11, 2.73]—movements on these scales of 42%, 37%, and 40%, respectively (descriptive statistics for the other pre-exposure ratings can be found in Appendix D, in Table S1).

Critically, these findings suggest that not only did subjects in the warning conditions notice the presence of the trigger warning, the content of that warning altered what they thought the film to follow would be like. In other words, trigger warnings—as intended—negatively affected subjects' expectations about the material they preceded.

Ratings of material. Subjects rated the negativity of the film they saw a second time, too, after they had seen it. As for previous experiments, we calculated subjects' mean post-exposure rating of how negative they thought the film clip they saw was, classified by which conditions they were in, and display those results in Table 4 (those same results for the rest of the ratings appear in Appendix D, in Table S1). We then calculated the mean differences between conditions, and display those in Tables 5 and 6.

As Table 5 shows, once again "warning" subjects gave just slightly less negative ratings than their "no warning" counterparts—movements of 5% for the "negative film" subjects, and 11% for the "control film" subjects. The confidence intervals around those differences include a range of small values as plausible, although for "control" subjects only, the interval does not span 0—a result counter to what we would have expected, but which requires replication. Further, as Table 6 shows, we again found that subjects who saw a negative film judged it to be

much more negative than subjects who saw a control film judged it—movements of 70% for "warning" subjects, and 65% for "no warning" subjects.

NHST analyses on subjects' ratings of negativity yielded no significant interaction between presence of warning and negativity of material, $F(1, 313) = 1.09$, $p = .30$. But there was a significant main effect of warning, $F(1, 313) = 9.28$, $p < .01$, such that subjects who saw a trigger warning rated the film clip they saw as less negative than did those who saw no warning; and there was also a significant main effect of film negativity, $F(1, 313) = 655.78$, $p < .001$, such that subjects who saw a negative film clip rated it as more negative than those who saw a control clip.

Negative affect. As for previous experiments, we first established that baseline negative affect was similar across conditions (it was; see Table 3), and then calculated subjects' mean rating of negative affect felt after watching a film clip, classified by which conditions they were in, and the mean differences between conditions, and display those results in Tables 4, 5, and 6. As Table 5 shows, no matter which type of film clip they saw, "warning" subjects felt only very slightly less negative than their "no warning" counterparts—movements of 2% for "negative" subjects, and 1% for "control" subjects. Moreover, the confidence intervals around those differences span 0 and include values in the opposite direction. Once again, these results suggest trigger warnings had little effect on how negative subjects felt. Further, "negative" subjects reported feeling more negative affect after watching the film clip than "control" subjects—movements of 15% for "warning" subjects, and 16% for "no warning" subjects—replicating previous experiments.

We also conducted a 2(warning, no warning) x 2(negative film, control film) x 2(baseline rating, rating after) mixed ANOVA on subjects' negative affect. This NHST analysis yielded no significant interaction between warning, film negativity, and time, $F(1, 313) = 1.31$, $p = .25$, or between warning and film negativity, $F(1, 313) = 0.25$, $p = .62$, or between warning and time, $F(1, 313) = 0.21$, $p = .65$. But there was a significant interaction between film negativity and time, $F(1, 313) = 52.63$, $p < .001$, whereby subjects who saw a negative film reported a greater increase in negative affect over time than did subjects who watched a control film. There were also significant main effects of time, $F(1, 313) = 120.50$, $p < .001$, such that subjects overall felt worse after watching a film; and of film negativity, $F(1, 313) = 26.38$, $p < .001$, such that subjects who saw a negative film clip overall felt worse. But there was no significant main effect of warning, $F(1, 313) = 0.62$, $p = .43$.

Intrusions. As for previous experiments, we calculated subjects' mean number of intrusions after watching a film clip **(**maximum tally = 83, before we Winsorised them; subjects reported high adherence to noting intrusions, $M = 8.53$, 95% CI [8.32, 8.75], indicating they took this task seriously), subjects' mean frequency of intrusion symptoms, and subjects' mean performance on the comprehension questions about the article they read, each classified by which conditions subjects were in, and then calculated the mean differences between conditions, and display those results in Tables 4, 5, and 6. As the tables show, "warning" subjects reported slightly more intrusions than their "no warning" counterparts after watching a negative film clip—a difference of one and three-quarters of a thought—and slightly fewer after watching a control film clip—a

difference just shy of one thought; “warning” subjects reported slightly less frequent intrusions than their “no warning” counterparts, regardless of which type of film they saw—a 6% movement for “negative” subjects and a 5% movement for “control” subjects; and “warning” subjects performed similarly to their “no warning” counterparts on the comprehension questions, regardless of film type—a 3% movement for “negative” subjects, and a 1% movement for “control” subjects. Moreover, the confidence interval around each of these differences spans 0 and includes values in the opposite direction. Together, then, these measures suggest trigger warnings had little effect on subjects’ intrusions.

By contrast, changing the negativity of the film had the expected effect on rates of intrusions: “Negative” subjects reported more intrusions than did “control” subjects—approximately two and three-quarters of a thought more for “warning” subjects, and approximately an eighth of a thought more for “no warning” subjects; “negative” subjects rated their intrusions as more frequent than did “control” subjects—movements of 14% for “warning” subjects, and 15% for “no warning” subjects; and “negative” subjects had slightly worse comprehension than “control” subjects—movements of 1% for “warning” subjects, and 4% for “no warning” subjects.

NHST analyses likewise revealed that, with regard to effects on the number of intrusions subjects reported, there was no significant interaction between presence of warning and negativity of material, $F(1, 313) = 3.63$, $p = .06$; and there was no significant main effect of warning, $F(1, 313) = 0.35$, $p = .55$; but there was a significant main effect of film negativity, $F(1, 313) = 4.41$, $p = .04$, such that subjects who saw a negative film clip reported a greater number of

intrusions than those who saw a control clip. For subjects' ratings of the frequency of intrusions, there was no significant interaction between presence of warning and negativity of material, $F(1, 313) = 0.01$, $p = .92$; there was no significant main effect of warning, $F(1, 313) = 3.12$, $p = .08$; but there was a significant main effect of film negativity, $F(1, 313) = 20.93$, $p < .01$, such that subjects who saw a negative film clip reported more frequent intrusions than those who saw a control clip. Finally, for subjects' performance on the comprehension questions about the article, there was no significant interaction between presence of warning and negativity of material, $F(1, 313) = 0.33$, $p = .57$; there was no significant main effect of warning, $F(1, 313) = 0.13$, $p = .72$; and no significant main effect of film negativity, $F(1, 313) = 0.71$, $p = .40$.

Avoidance. As for previous experiments, we calculated subjects' rated mean frequency of avoidance symptoms after watching a film clip, classified by which conditions they were in, and the mean differences between conditions, and display those results in Tables 4, 5, and 6. As the tables show, "warning" subjects reported only slightly less avoidance than their "no warning" counterparts after either type of film—a 5% movement for "negative" subjects, and a 1% movement for "control" subjects. Moreover, the confidence intervals around those differences span 0 and include values in the opposite direction as plausible. This result suggests trigger warnings had little effect on the degree to which subjects experienced avoidance symptoms. But "negative" subjects reported more frequent avoidance symptoms than did "control" subjects—movements of 7% for "warning" subjects, and 11% for "no warning" subjects, indicating negativity of the material affected avoidance as expected.

Further, NHST analyses on subjects' reported frequency of avoidance yielded no significant interaction between presence of warning and negativity of material, $F(1, 313) = 0.40$, $p = .53$; and no significant main effect of warning, $F(1, 313) = 1.31$, $p = .25$. But there was a significant main effect of film negativity, $F(1, 313) = 10.96$, $p < .01$, such that subjects who saw a negative film clip reported a higher frequency of avoidance than those who saw a control clip.

The results of this experiment once again suggest the conclusion that trigger warnings had very little effect on the frequency of subjects' negative symptoms, even though more negative materials increased the rates of those symptoms. What is more, this experiment provides evidence against the counterexplanation that our trigger warnings have little effect because they do not change people's expectations about the material they precede. In this experiment, subjects anticipated that the film clip would be much worse if they had just seen a trigger warning about it, compared to if they had not. Considered as a whole, our findings converge on the idea that trigger warnings are unhelpful.

But there is another counterexplanation for our findings. Trigger warnings—at least, in their narrower conception—are intended to stave off symptoms elicited by negative material in people who have experienced a trauma, and should be particularly relevant when that material reminds people of their past trauma. But we have not examined the effects of trigger warnings using subjects with a known history of trauma, and we have not considered the effect of overlap between the content of the warned-about material and the potentially-traumatic previous experiences that our subjects have had.

Therefore, perhaps trigger warnings would have clearer, larger effects (helpful or harmful) in people who have had a traumatic experience, and especially in those for whom our negative materials are similar to a previous traumatic experience they have had. Based on previous epidemiological research we would expect that most of our subjects—like most of the general population—have experienced an event that could be considered traumatic (Breslau et al., 1998). We conducted Experiment 4 to collect evidence for that supposition, and to explore the effects of trigger warnings on people who have experienced a traumatic event similar to the one in the film clip they see.

Experiment 4

Method.

Subjects. We used MTurk to recruit 438 subjects, who received USD0.25 for completing the study. As for Experiments 2b and 3, we aimed to collect data from enough subjects that we could exclude up to 30% of our sample for failing our compliance checks and still have 70 subjects in each of the four conditions. We retained data from 306 subjects for analysis (our exclusion criteria are noted below). Of those subjects, 63% were female and 37% male; their ages ranged from 18-78, Median = 32, M = 35.64, 95% CI [34.24, 37.04]; 96% of them reported that they were citizens of the US; and 98% of them reported that English was their first language.

Procedure. The design and procedure of this experiment was identical to that of Experiment 2b, except for the addition of one scale, described below. In this experiment, subjects spent approximately 3 minutes reading the non-fiction article and noting intrusions related to the film clip they had seen, M = 169.13 s,

95% CI [156.15, 182.11], Median = 151.23 s. Then, after subjects had made the ratings about the film clip they had seen, they all completed the Trauma History Screen (THS; Carlson et al., 2011).

The THS is a two-part measure of prior traumatic experiences. In the first part of the measure, subjects see a list of objectively traumatic events (such as "a hurricane, flood, earthquake, tornado, or fire" and "attack with a gun, knife, or weapon"; events that would meet *criterion A* for diagnosis with PTSD, see Table 1; APA, 2013) and for each they indicate if that sort of event has happened to them, and (if it has) how many times. We made two slight alterations to this list of events: One, we separated out the first item, regarding accidents, into two items—the first asking specifically about car accidents (the content of one of our film clip pairs) and the second asking about boat, train, and airplane accidents. Two, we added a new item, "Domestic abuse – physical or psychological" (the content of the other of our film clip pairs). Subjects' final task in the first part of the measure is to indicate if any of those events that have happened to them had "really bothered [them] emotionally." If subjects answer "yes" to this question, they then see the second part of the measure.

In the second part of the THS, subjects answer a series of questions about each of up to five of the events they listed in part one that had really bothered them emotionally. Specifically, for each such event, subjects are asked to specify which type of event it was (with reference to the initial list of events), give a brief description of the event, state their age at the time of the event, and indicate whether or not, in the course of the event, "anyone [got] hurt or killed," if they were "afraid [they] or someone else might get hurt or killed," if they felt "very

afraid, hopeless, or horrified," and if they felt "unreal, spaced out, disorientated, or strange." Subjects also indicate how long they were bothered by the event afterwards, on a scale with the four options: "not at all," "1 week," "2-3 weeks," and "a month or more"; and how much it bothered them emotionally, on a scale with the five options: "not at all," "a little," "somewhat," "much," and "very much."

There are several ways to examine subjects' responses on the THS. One is to consider whether subjects have been exposed to objectively traumatic events, that is, events that really upset most people to whom they happen; these events are called *high magnitude stressors* (HMS), and this determination is made according to whether subjects indicated experiencing anything on the initial list of events. A second approach is to consider whether subjects have experienced a HMS that deeply upset them for a long time—that is, if they have experienced *persisting posttraumatic distress* (PPD) in response to at least one of these events. This determination is made according to whether subjects indicated having a HMS that really bothered them emotionally, much or very much, and for a month or more.

The THS was designed to be quick to administer and easily understood—features achieved by streamlining its structure and using straightforward language. In a variety of samples (including undergraduate students, and homeless veterans in rehabilitation), both the HMS and PPD scores had high test-retest reliability, with correlations ranging from .74-.93 for HMS and .73-.95 for PPD, across periods ranging from 1 week to 2 months. Both scores also had good convergent validity: HMS scores correlated highly with other measures of traumatic experiences, ranging from .73-.81, and both scores correlated at least moderately

with measures of PTSD symptoms, ranging from .22-.41 for HMS and .18-.38 for PPD. Further, those subjects who indicated at least one PPD event had higher rates of PTSD symptoms than those who indicated no PPD events (Carlson et al., 2011). Taken together, these psychometric properties suggest this scale is suitable for use screening our subjects for prior exposure to objectively and subjectively distressing events.

Results & Discussion.

Exclusions. We excluded subjects who reported having previously seen the film clip ($n = 9$), or who had not correctly completed one or more critical tasks—the same criteria as for Experiments 2b and 3. Altogether, we excluded 132 subjects (30%), leaving us with a distribution of subjects as follows: $n_{\text{WarningNegative}} = 80$, $n_{\text{NoWarningNegative}} = 77$, $n_{\text{WarningControl}} = 71$, $n_{\text{NoWarningControl}} = 78$.

Manipulation check. At the end of the experiment, 94% of "warning" subjects reported they remembered seeing the trigger warning, suggesting subjects in the warning condition took note of our manipulation.

History of trauma. We first examined the proportion of subjects who reported having experienced at least one HMS event, and found that the vast majority—89%, 95% CI [85, 92]—had, as we expected. This finding suggests that most of our sample have a history of trauma, and that whatever effects of trigger warnings we find can therefore be generalised to other people with a history of trauma. Next, we checked the proportions of subjects who indicated they had experienced car accidents or domestic abuse (the topics of our film clip materials). We found that 132 (43%) had experienced a really bad car accident, and 112 (37%) had experienced physical or psychological domestic abuse, suggesting our

film clip materials were about common traumatic experiences. Further, 47% of our sample, 95% CI [42, 53], had experienced PPD following at least one of the events they had reported.

But considering both whether subjects had experienced PPD and which type of event they had experienced it in relation to, and then adding the fact of random assignment to film-topic counterbalance, ultimately only 34 subjects (11%) had experienced PPD due to an event that overlapped with the topic of the film clip they were assigned to see. This number of subjects is too small to be useful in making meaningful comparisons between conditions. We therefore conducted the rest of our analyses without consideration of these subgroups.

Ratings of material. We calculated subjects' mean rating for how negative the film clip was, classified by which conditions they were in, and the mean differences between conditions, and display those results in Tables 4, 5, and 6 (and descriptive statistics for the other ratings in Appendix D, in Table S1). Echoing previous experiments, "warning" subjects gave just slightly less negative ratings than their "no warning" counterparts, no matter which type of film they saw—movements of 2% for the "negative film" subjects, and 4% for the "control film" subjects—and the confidence intervals around those differences include a range of small values as plausible effect sizes, including 0. But subjects who saw a negative film rated it as much more negative than subjects who saw a control film rated it—movements of 62% for "warning" subjects, and 60% for "no warning" subjects.

NHST analyses on subjects' ratings of negativity yielded no significant interaction between presence of warning and negativity of material, $F(1, 302) =$

0.10, $p = .75$; and no significant main effect of warning, $F(1, 302) = 1.02$, $p = .31$. But there was a significant main effect of film negativity, $F(1, 302) = 467.78$, $p < .001$, such that subjects who saw a negative film clip rated it as more negative than those who saw a control clip.

Negative affect. We established that baseline negative affect was similar across conditions (it was; see Table 3), and then calculated subjects' mean ratings of negative affect after watching a film clip, classified by which conditions they were in, and the mean differences between conditions, and display those results in Tables 4, 5, and 6. As in previous experiments, "warning" subjects felt similarly negative as their "no warning" counterparts. "Negative" subjects felt slightly less negative, having been warned—a 3% movement—whereas "control" subjects felt slightly more so—a 5% movement. But the confidence intervals around both those differences include a range of small values as plausible effect sizes, including 0. By contrast, "negative" subjects reported feeling more negative affect after watching their assigned film clip than "control" subjects felt after theirs—movements of 15% for "warning" subjects, and 22% for "no warning" subjects.

We also conducted a 2(warning, no warning) x 2(negative film, control film) x 2(baseline rating, rating after) mixed ANOVA on subjects' ratings of negative affect. This NHST analysis yielded no significant interaction between warning, film negativity, and time, $F(1, 302) = 1.58$, $p = .21$, or between warning and film negativity, $F(1, 302) = 1.74$, $p = .19$, or between warning and time, $F(1, 302) = 0.42$, $p = .52$. But there was a significant interaction between film negativity and time, $F(1, 302) = 63.38$, $p < .001$, whereby subjects who saw a negative film reported a greater increase in negative affect over time than did subjects who

watched a control film. There were also significant main effects of time, $F(1, 302) = 130.64$, $p < .001$, such that subjects overall felt worse after watching a film clip; and of film negativity, $F(1, 302) = 36.93$, $p < .001$, such that subjects who saw a negative film clip overall felt worse. But there was no significant main effect of warning, $F(1, 302) = 0.07$, $p = .79$.

Intrusions. We calculated subjects' mean number of intrusions after watching a film clip **(**maximum tally = 138, before we Winsorised them; again, subjects reported high adherence to this "noting intrusions" task, $M = 8.51$, 95% CI [8.27, 8.74]), subjects' mean frequency of intrusion symptoms, and subjects' mean performance on the comprehension questions about the article they read, each classified by which conditions subjects were in, and then calculated the mean differences between conditions—these results appear in Tables 4, 5, and 6. We found that "warning" subjects reported slightly fewer intrusions than their "no warning" counterparts, following either type of film (a difference of approximately two-thirds of a thought for "negative" subjects, and approximately one thought for "control" subjects), very similarly frequent intrusions ("negative" subjects slightly less—a 1% movement—and "control" subjects slightly more—a 1% movement), and performed very similarly on the comprehension questions ("negative" subjects slightly worse—a 3% movement—and "control" subjects slightly better—a 4% movement). In each case, the confidence interval indicating the range of plausible effect sizes for trigger warnings spans 0 to include values in the opposite direction. But the effects of film negativity were larger: compared to "control" subjects, "negative" subjects reported more intrusions (a difference of approximately one and six-sevenths of a thought for "warning" subjects, and

approximately one and a half thoughts for "no warning" subjects), and rated their intrusions as more frequent (movements of 17% for "warning" subjects, and 18% for "no warning" subjects), although they had very similar comprehension ("warning" subjects slightly worse—a 4% movement—and "no warning" subjects slightly better—a 3% movement).

NHST analyses of these measures likewise revealed that, with regard to effects on the number of intrusions subjects reported, there was no significant interaction between presence of warning and negativity of material, $F(1, 302) = 0.05$, $p = .82$; and no significant main effect of warning, $F(1, 302) = 1.36$, $p = .25$; but there was a significant main effect of film negativity, $F(1, 302) = 6.10$, $p = .01$, such that subjects who saw a negative film clip reported a greater number of intrusions than those who saw a control clip. For subjects' ratings of the frequency of intrusions, there was no significant interaction between presence of warning and negativity of material, $F(1, 302) = 0.07$, $p = .79$; and no significant main effect of warning, $F(1, 302) < 0.01$, $p = .97$; but there was a significant main effect of film negativity, $F(1, 302) = 28.93$, $p < .001$, such that subjects who saw a negative film clip reported more frequent intrusions than those who saw a control clip. Finally, for subjects' performance on the comprehension questions about the article, there was no significant interaction between presence of warning and negativity of material, $F(1, 302) = 1.02$, $p = .31$; no significant main effect of warning, $F(1, 302) = 0.03$, $p = .87$; and no significant main effect of film negativity, $F(1, 302) = 0.05$, $p = .82$.

Avoidance. We calculated subjects' mean frequency of avoidance symptoms after watching a film clip, classified by which conditions they were in, and the

mean differences between conditions, and display those results in Tables 4, 5, and 6. As before, “warning” subjects were very similarly avoidant as their “no warning” counterparts (“negative” subjects slightly less—a 1% movement—and “control” subjects slightly more—a 2% movement), and the confidence intervals around the range of plausible differences span 0, and values in the opposite direction. But “negative” subjects reported more frequent avoidance symptoms than “control” subjects (movements of 10% for “warning” subjects, and 13% for “no warning” subjects).

NHST analysis of subjects’ reported frequency of avoidance yielded no significant interaction between presence of warning and negativity of material, $F(1, 302) = 0.44$, $p = .51$; and no significant main effect of warning, $F(1, 302) = 0.09$, $p = .77$. But there was a significant main effect of film negativity, $F(1, 302) = 21.10$, $p < .001$, such that subjects who saw a negative film clip reported a higher frequency of avoidance than those who saw a control clip.

Our findings in this experiment are in essence the same as those in our previous experiments, in that we found trigger warnings had no meaningful effects on subjects’ distress, even as more negative materials increased subjects’ distress. In this experiment, we also found that the majority of our sample had experienced an event with the potential to be extremely stressful, in line with previous research (Breslau et al., 1998; Carlson et al., 2011). Further, almost half of our sample had indeed experienced ongoing distress due to one such event. In other words, much of our sample had a history of trauma—yet, once again, we found little effect of trigger warnings. Taken together, these findings suggest trigger warnings are unlikely to be uniquely helpful for people who have a history

of trauma. Even so, based on this experiment, we cannot rule out that trigger warnings may have effects of a meaningful size in people for whom the warned-about content is closely related to a previous experience of theirs, which caused them great distress.

To more precisely estimate the direction and size of the effects trigger warnings have—that is, to get a clearer picture of what these warnings do, based on the data we do have—we next conducted a series of meta-analyses using the 1394 subjects whose data we retained in each of our experiments.

Meta-analyses of Key Measures

Ratings of material. To more precisely estimate trigger warnings' effect on how negative people judge the subsequent material to be, we used the data from all our experiments (except Experiment 1a, in which we did not ask subject to make these ratings) to conduct a random effects model mini meta-analysis in ESCI software (Cumming, 2012). As Figure 3 shows, "warning" subjects rated the materials just 0.15 less negatively than did "no warning" subjects (recall the maximum possible difference was 6, making this a 2.5% movement). Importantly, the 95% confidence interval around this estimated effect size, [-0.29, -0.01], is narrow—more so than in any of the individual experiments. This increased precision means the range of plausible values of the true effect size that this interval spans no longer quite includes 0, but still comes very close to it. Meaningful variance across effect size estimates was low, $I^2 = 0\%$, and the estimated standard deviation of the distribution of true effect sizes is small, $T = 0$, 95% CI [0, 0.30], indicating little heterogeneity.

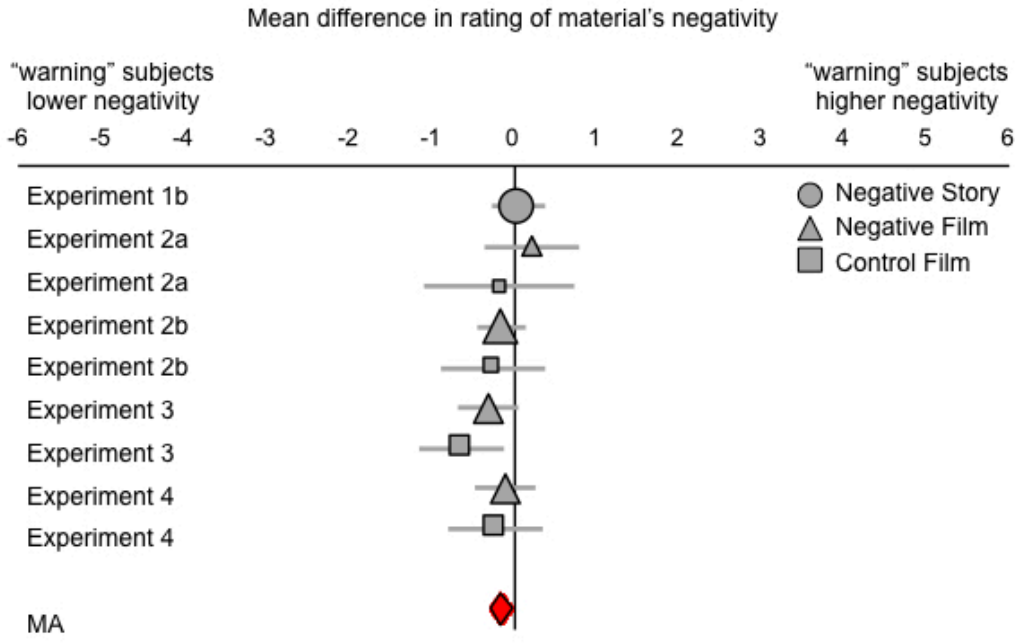

Figure 3.

Forest plot of trigger warnings' effect on rating of the material's negativity. The width of the horizontal axis represents the maximum possible raw mean difference between subjects who did and subjects who did not get a trigger warning. The points plotted vertically represent the mean difference for each experiment, or for each type of material within each experiment. The vertical line indicates the point of no mean difference between "warning" and "no warning" subjects; points to its left indicate a lower mean score for "warning" subjects, whereas points to its right indicate a higher mean score for "warning" subjects. The black lines extending from each point show the 95% confidence interval around that mean difference. Larger points indicate samples given greater weighting in the meta-analysis. The diamond (labelled "MA") represents the result of the meta-analysis; its centre represents the estimated raw mean effect size of trigger warnings, and its width shows the 95% confidence interval around that estimate.

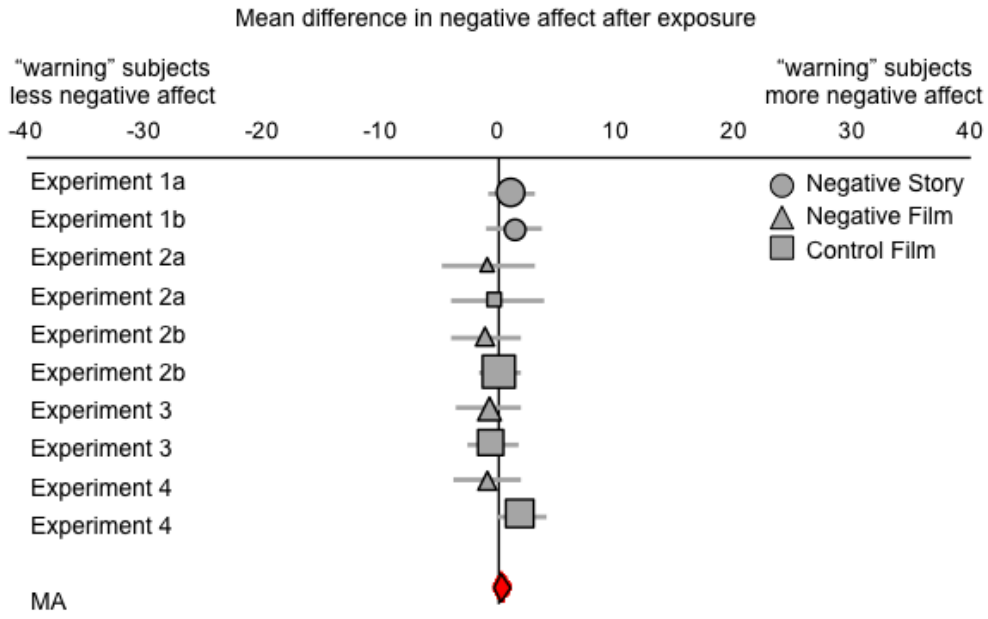

Figure 4.

Forest plot of trigger warnings' effect on negative affect felt after exposure to the material. The width of the horizontal axis represents the maximum possible raw mean difference between subjects who did and subjects who did not get a trigger warning. The points plotted vertically represent the mean difference for each experiment, or for each type of material within each experiment. The vertical line indicates the point of no mean difference between "warning" and "no warning" subjects; points to its left indicate a lower mean score for "warning" subjects, whereas points to its right indicate a higher mean score for "warning" subjects. The black lines extending from each point show the 95% confidence interval around that mean difference. Larger points indicate samples given greater weighting in the meta-analysis. The diamond (labelled "MA") represents the result of the meta-analysis; its centre represents the estimated raw mean effect size of trigger warnings, and its width shows the 95% confidence interval around that estimate.

This reduction is very small in standardised terms, too: $d_{unbiased}$ = -0.14, 95% CI [-0.26, -0.03] (this effect size is also known as Hedge's *g,* and is calculated by dividing the mean difference by the pooled standard deviation and then applying a bias correction, but, as recommended, the CI given is around the uncorrected effect; Cumming, 2012). In summary, the best estimate suggests that providing a trigger warning about material—written or visual, of greater or lesser negativity—only very slightly decreased how negative subjects judged that material to be.

Negative affect. To more precisely estimate the size of trigger warnings' effect on how much negative affect people feel after exposure to the material, we conducted another mini meta-analysis, using data from all our experiments. As Figure 4 shows, "warning" subjects felt 0.25 more negative affect after exposure to the material than "no warning" subjects (the maximum possible difference was 40, making this a less than 1% movement). Again, the 95% CI around this effect size is far narrower than those in any one experiment, but in spite of this increased precision it still includes 0 as a plausible value [-0.51, 1.00]; or, in standardised terms, $d_{unbiased}$ = 0.02, [-0.08, 0.13]. Further, heterogeneity was low; I^2 = 0%, T = 0, [0, 1.30]. This analysis leads us to the conclusion that seeing a trigger warning essentially had no effect on how negative subjects felt, following exposure to a variety of materials.

Intrusions. To more precisely estimate the effects of trigger warnings on people's intrusion symptoms, we conducted three more meta-analyses. As Figure 5 shows, "warning" subjects reported 0.27 intrusions fewer than "no warning" subjects—a difference of approximately quarter of a thought—while reading the non-fiction article. What is more, the 95% CI around this difference is narrow,

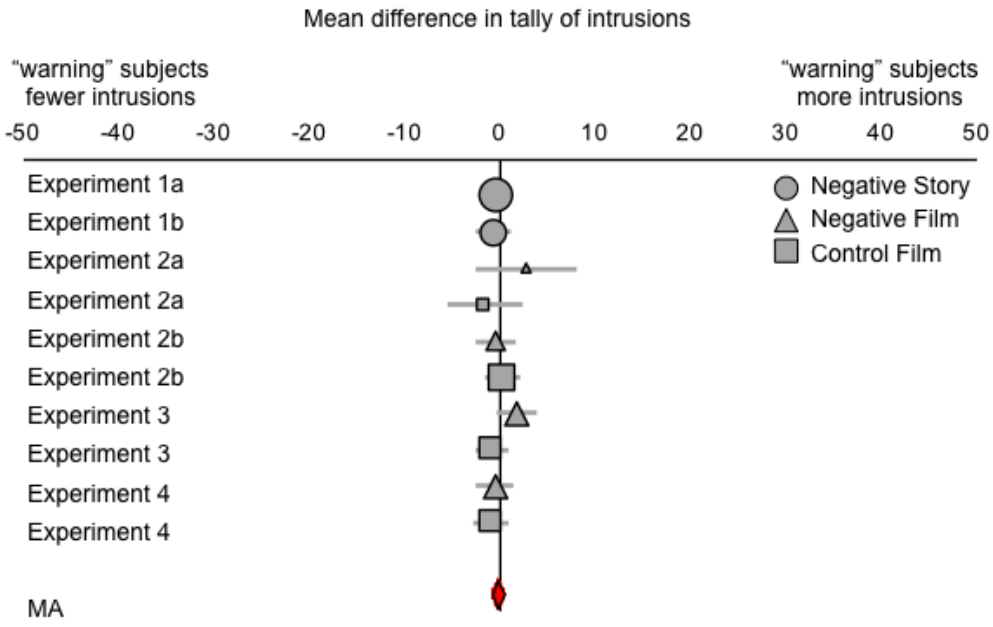

Figure 5.

Forest plot of trigger warnings' effect on tally of intrusions. The width of the horizontal axis represents the maximum possible raw mean difference between subjects who did and subjects who did not get a trigger warning, given the maximum (after Winsorisation) number of intrusions reported by subjects. The points plotted vertically represent the mean difference for each experiment, or for each type of material within each experiment. The vertical line indicates the point of no mean difference between "warning" and "no warning" subjects; points to its left indicate a lower mean score for "warning" subjects, whereas points to its right indicate a higher mean score for "warning" subjects. The black lines extending from each point show the 95% confidence interval around that mean difference. Larger points indicate samples given greater weighting in the meta-analysis. The diamond (labelled "MA") represents the result of the meta-analysis; its centre represents the estimated raw mean effect size of trigger warnings, and its width shows the 95% confidence interval around that estimate.

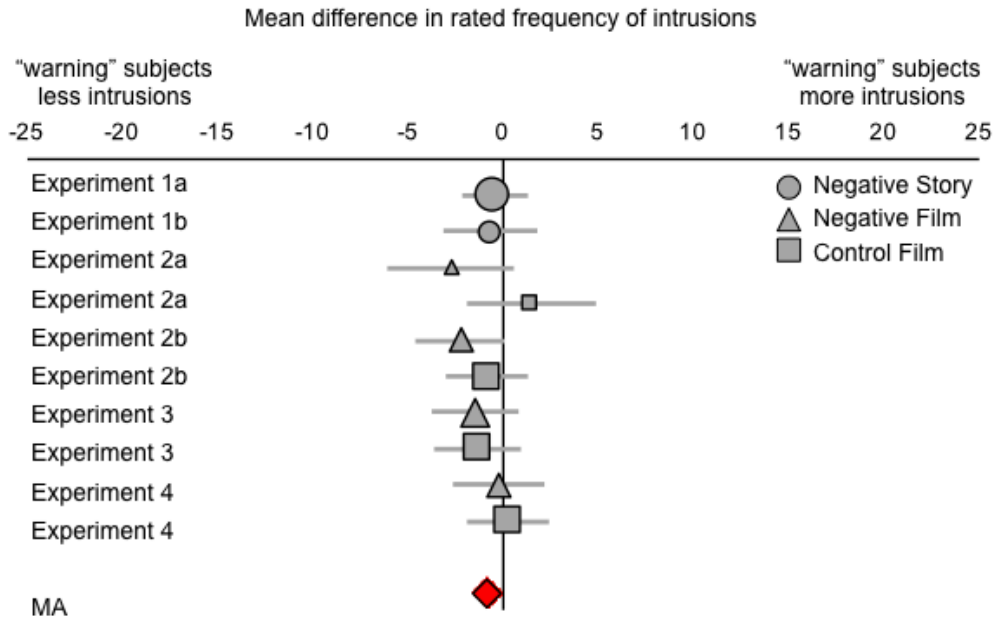

Figure 6.

Forest plot of trigger warnings' effect on rated frequency of intrusions. The width of the horizontal axis represents the maximum possible raw mean difference between subjects who did and subjects who did not get a trigger warning. The points plotted vertically represent the mean difference for each experiment, or for each type of material within each experiment. The vertical line indicates the point of no mean difference between "warning" and "no warning" subjects; points to its left indicate a lower mean score for "warning" subjects, whereas points to its right indicate a higher mean score for "warning" subjects. The black lines extending from each point show the 95% confidence interval around that mean difference. Larger points indicate samples given greater weighting in the meta-analysis. The diamond (labelled "MA") represents the result of the meta-analysis; its centre represents the estimated raw mean effect size of trigger warnings, and its width shows the 95% confidence interval around that estimate.

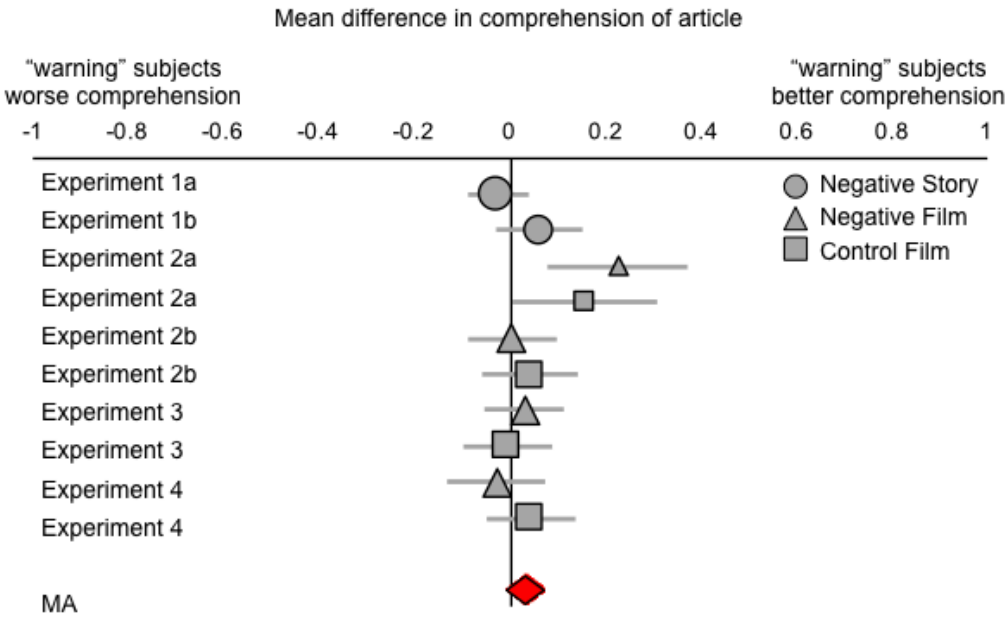

Figure 7.

Forest plot of trigger warnings' effect on comprehension of the article. The width of the horizontal axis represents the maximum possible raw mean difference between subjects who did and subjects who did not get a trigger warning. The points plotted vertically represent the mean difference for each experiment, or for each type of material within each experiment. The vertical line indicates the point of no mean difference between "warning" and "no warning" subjects; points to its left indicate a lower mean score for "warning" subjects, whereas points to its right indicate a higher mean score for "warning" subjects. The black lines extending from each point show the 95% confidence interval around that mean difference. Larger points indicate samples given greater weighting in the meta-analysis. The diamond (labelled "MA") represents the result of the meta-analysis; its centre represents the estimated raw mean effect size of trigger warnings, and its width shows the 95% confidence interval around that estimate.

[-0.88, 0.34], but it still includes 0 among the plausible sizes for the effect; and there was low heterogeneity, I^2 = 0%, T = 0, [0, 1.09]. Figure 6 shows "warning" subjects rated their intrusions on the IES as 0.84 less frequent than did "no warning" subjects (the maximum possible difference was 25, making this a 3% movement). The 95% CI around this difference no longer quite includes 0 among the most plausible values for the effect, [-1.56, -0.11], but still spans a narrow range of small values. Again, heterogeneity was low, I^2 = 0%, T = 0, [0, 1.14]. Figure 7 shows the proportion of comprehension questions "warning" subjects got correct was 0.03 greater than the proportion "no warning" subjects got correct (the maximum possible difference was 1, making this a 3% movement), the 95% CI around this difference is narrow yet still includes 0 as a plausible value, [-0.01, 0.07], and although a moderate proportion of the variance between experiments was meaningful, I^2 = 41.52%, its absolute magnitude is low, T = 0.04, [0, 0.08].

What is more, each of these effects is very small in standardised terms, too: Intrusions tally, d_{unbiased} = -0.04, [-0.14, 0.07]; IES intrusions, d_{unbiased} = -0.12, [-0.23, -0.02]; and comprehension, d_{unbiased} = 0.11, [-0.03, 0.25]. Put differently, although most subjects experienced intrusive thoughts related to the material they saw, the best estimates suggest that seeing a trigger warning beforehand only slightly decreased the degree to which they experienced these intrusions.

Avoidance. To more precisely estimate the size of the effect of trigger warnings on people's avoidance symptoms, we once again meta-analysed our data. As Figure 8 shows, "warning" subjects rated their avoidance on the IES as 0.50 less frequent than "no warning" subjects (the maximum possible difference was 40, making this a 1% movement). The 95% CI around this difference is

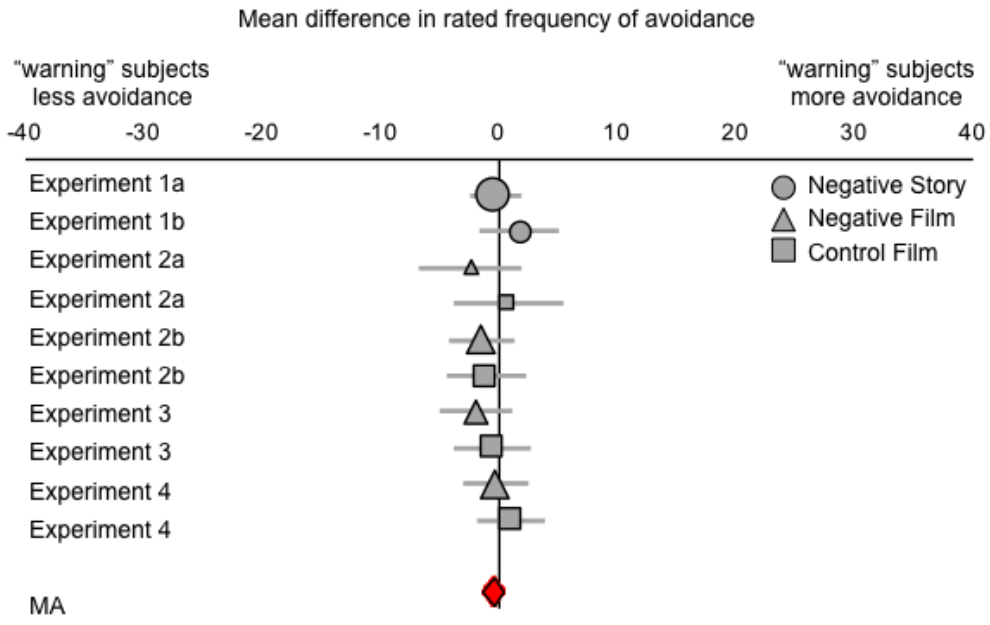

Figure 8.

Forest plot of trigger warnings' effect on rated frequency of avoidance. The width of the horizontal axis represents the maximum possible raw mean difference between subjects who did and subjects who did not get a trigger warning. The points plotted vertically represent the mean difference for each experiment, or for each type of material within each experiment. The vertical line indicates the point of no mean difference between "warning" and "no warning" subjects; points to its left indicate a lower mean score for "warning" subjects, whereas points to its right indicate a higher mean score for "warning" subjects. The black lines extending from each point show the 95% confidence interval around that mean difference. Larger points indicate samples given greater weighting in the meta-analysis. The diamond (labelled "MA") represents the result of the meta-analysis; its centre represents the estimated raw mean effect size of trigger warnings, and its width shows the 95% confidence interval around that estimate.

narrow, [-1.45, 0.45], indicating high precision. But this interval still includes 0 among the plausible true values for the effect. Heterogeneity was low, I^2 = 0%, T = 0, [0, 1.25]. This effect was also small in standardised terms: d_{unbiased} = -0.05, [-0.16, 0.05]. That is, the best estimate suggests that seeing a trigger warning only slightly decreased subjects' avoidance symptoms in relation to the various materials we used.

Of course, one criticism of these meta-analyses is that including the control material conditions as we did could dilute the effect of warnings on more negative material—conditions under which trigger warnings would have greater opportunity to be helpful. Indeed, one reason we switched from using story negative materials to film materials was examine the effects of trigger warnings given about more versus less negative material. To address this issue, we therefore conducted these meta-analyses again, but used only data from subjects exposed to negative material. That is, we included data from all subjects in Experiments 1a and 1b, who all read a negative short story, but only data from subjects who saw a negative film clip in Experiments 2a, 2b, 3. and 4, resulting in a total N = 903. The outcomes of these meta-analyses were very similar to those reported above.

Ratings of material. More specifically, the first in this additional round of meta-analyses revealed that "warning" subjects who had seen negative material rated that material as just 0.09 less negative than did "no warning" subjects who had also seen negative material, 95% CI [-0.25, 0.07]; there was little heterogeneity, I^2 = 0%, T = 0, 95% CI [0, 0.35]; and the effect was small in standardised terms, too, d_{unbiased} = -0.10, 95% CI [-0.25, 0.06].

Negative affect. Similarly, "warning" subjects who had seen negative material reported feeling just 0.10 more negative affect after exposure to that material than did "no warning" subjects who had also seen negative material, 95% CI [-0.96, 1.15]; there was little heterogeneity, $I^2 = 0\%$, $T = 0$, 95% CI [0, 2.02]; and the effect was small in standardised terms, too, $d_{\text{unbiased}} < 0.01$, 95% CI [-0.13, 0.13].

Intrusions. Further, "warning" subjects who had seen negative material reported 0.04 fewer intrusions than did "no warning" subjects who had also seen negative material, 95% CI [-0.85, 0.76]; there was little heterogeneity, $I^2 = 4.28\%$, $T = 0.21$, 95% CI [0, 1.81]. "Warning" subjects who had seen negative material also rated their intrusions as 1.11 less frequent on the IES than did "no warning" subjects who had seen negative material, 95% CI [-2.03, -0.20]; there was little heterogeneity, $I^2 = 0\%$, $T = 0$, 95% CI [0, 1.48]. The proportion of comprehension questions "warning" subjects who had seen negative material got correct was 0.03 greater than the proportion "no warning" subjects who had seen negative material got correct, 95% CI [-0.03, 0.09]; there was a moderate proportion of meaningful heterogeneity, $I^2 = 56.46\%$, but its absolute magnitude was small, $T = 0.05$, 95% CI [0, 0.01]. Each of these effects was small in standardised terms, too: Intrusions tally, $d_{\text{unbiased}} < 0.01$, 95% CI [-0.13, 0.14]; IES intrusions, $d_{\text{unbiased}} = -0.15$, 95% CI [-0.28, -0.02]; comprehension, $d_{\text{unbiased}} = 0.10$, 95% CI [-0.10, 0.30].

Avoidance. Finally, "warning" subjects who had seen negative material rated their avoidance as 0.72 less frequent on the IES than did "no warning" subjects who had also seen negative material, 95% CI [-1.88, 0.44]; there was

little heterogeneity, $I^2 = 0\%$, $T = 0$, 95% CI [0, 2.16]; and the effect was small in standardised terms, too, $d_{\text{unbiased}} = -0.08$, 95% CI [-0.21, 0.05].

For each of these additional meta-analyses, the estimated effect of trigger warnings and the 95% CI around it are substantially similar to those yielded by the previous meta-analyses using all our data. The confidence interval around each of them spans a still-narrow range of small effect sizes, all but one of which includes zero as a plausible value for the size of the effect trigger warnings have. Taken together, these analyses show that the effects of trigger warnings that appear prior to negative material are of similarly trivial size as their effects when they have appeared prior to a mixture of more and less negative material.

Chapter 3

Summary of Findings

Recall that our primary research question was: to what extent does a trigger warning affect the symptoms of distress people experience soon after exposure to "warned about" material? We conducted six experiments addressing this question. These experiments yielded effects of trigger warnings that were inconsistent in direction and, at best, small. Subjects who saw trigger warnings judged material to be similarly negative, felt similarly negative, experienced similarly frequent intrusive thoughts and avoidance, and comprehended subsequent material similarly well as their unwarned counterparts.

In Experiments 1a and 1b, the material to which we exposed subjects was a negative story. Finding little effect of administering a trigger warnings prior to those materials, in Experiments 2a, 2b, 3, and 4, we instead turned to film clips. Subjects saw either a more negative, or a less negative film clip. Those experiments revealed that more negative materials led subjects to feel more negative, and experience higher rates of symptoms than did less negative materials. Yet trigger warnings continued to have little influence, regardless.

In Experiment 3, we addressed the possibility that we had found little effect of trigger warnings because our warnings were not changing subjects' expectations about the material to follow. We asked warned and unwarned subjects to rate what they thought the film clip would be like, before they had seen it. We found that subjects who had seen a trigger warning thought the film to follow would be more negative, suggesting the trigger warnings changed subjects' expectations.

In Experiment 4, we addressed the possibility that the effects of trigger warnings would be greater among subjects who had experienced trauma—those for whom trigger warnings are often intended. We asked all subjects about highly stressful events they had previously experienced. We found that almost all of our subjects had experienced events with the potential to be highly distressing, and that a little less than half of our subjects had indeed experienced persistent distress following one (or more) of those events, suggesting that trigger warnings have little effect on populations in which a history of trauma is common.

Finally, we conducted meta-analyses on data from each of our key measures across experiments. These analyses increased the precision with which we could estimate the true effect sizes of trigger warnings and allowed us to better answer our research question. They revealed that the answer to our research question is: trigger warnings have trivial effects.

Interpretation of Effect Sizes

As noted earlier, not all small effects inherently lack significance (Abelson, 1985; Prentice & Miller, 1992; Rosenthal, 1990). But the confidence intervals around our meta-analytically estimated effect sizes are narrow, yet most of these intervals still include zero as a plausible true value of that effect size. Further, putting aside the confidence intervals and taking the meta-analytically estimated mean differences as our best estimate of these effects—which they are—we can see these, too, are small (Cumming, 2012). We can put "small" into context in a number of ways.

For a start, the standardised mean differences we found in our meta-analyses, looking at the effect of providing a trigger warning, ranged from 0.02 to

0.11 (in absolute terms). Each of these differences is much smaller than 0.41—recall that this value is suggested as the minimum size an effect must be for interpretation as a practically meaningful difference in social science data (Ferguson, 2009). But of course, such cut-offs are merely recommendations; we must consider these effect sizes alongside other relevant reference points, too.

As one such reference point, a Cochrane review found the standardised mean difference in self-reported symptoms between those who underwent therapy for PTSD and controls was -1.60, 95% CI [-2.02, -1.18] (Bisson et al., 2013). Although trigger warnings are supposed to reduce symptoms of distress to a practically helpful degree, they are not intended to substitute for a course of therapy, and so we might expect their effects to be somewhat smaller than that of therapy. In fact, the symptom reductions we observed are minuscule in comparison to the reductions achieved by therapy.

As a second reference point, a recent paper meta-analytically examined evidence for the controversial claims about the benefits of "brain training" activities on underlying mental abilities, such as intelligence—that is, the extent to which these activities showed "far transfer" effects (Melby-Lervåg, Redick, & Hulme, 2016). These meta-analyses yielded effects on various measures of "far transfer," reported as standardised mean differences, that ranged from 0.01 to 0.20—notice that the biggest of these effects is larger than the largest meta-analytic effect of trigger warning we found. The authors concluded the effects of brain training were not sufficiently large to be considered practically meaningful.

As a third reference point, a large review considered effect sizes resulting from over 300 previous meta-analyses, with the aim of examining a broad range

of treatment effects (Lipsey & Wilson, 1993). Together, these meta-analyses had examined the effects of a large variety of psychology-based interventions on a wide range of outcomes related to mental health, workplace performance, and educational success. The review revealed that, among a representative sample of methodologically strong meta-analyses of treatment effects, the mean standardised difference observed was 0.47. Further, 83% of those effect sizes were greater than 0.20. Thus, relative to this wide range of other effects—comparable in terms of manipulations and outcomes—the effects of trigger warnings here was undoubtedly very small. Therefore, each of these comparisons points to the same conclusion that, in practical terms, trigger warnings are neither helpful, nor are they harmful.

Relationships with Previous Findings

This conclusion fits with previous work showing that people are often unsuccessful at avoiding unwanted mental influences (Wilson & Brekke, 1994). In our case, trigger warnings did not help people reduce the effects of the negative materials to which we exposed them. This finding further fits with previous work showing that, although people can regulate their emotions to some degree, they are not always successful at doing so (Gross, 2015; Gross & Jazaieri, 2014).

Given that trigger warnings did not help people reduce the negativity associated with the material we showed them, it makes sense that trigger warnings did not help reduce their other symptoms of distress, either. That result fits with previous work showing that the negativity of material (or an event) is linked to the degree of related intrusions people experience about it, because greater intensity

leads to more accessible memory (Hall & Berntsen, 2008; Rubin, Boals et al., 2008; Rubin et al., 2011).

Our secondary findings are also consistent with this previous work on the relationship between negativity and symptoms: We found that people who saw the more negative clips both reported feeling more negative, and experienced greater rates of symptoms than did people who saw the less negative, control clips (Hall & Berntsen, 2008; Rubin, Boals et al., 2008; Rubin et al., 2011). Of course, these results do not provide direct evidence in line with this memory accessibility explanation, because in these latter experiments we did not test subjects' memory for the two types of film clips.

Our finding that trigger warnings had little effect on how negative subjects felt diverges from previous research on the effects of warnings administered prior to viewing excerpts of negative films (Cantor et al., 1984; de Wied et al., 1997). Two previous studies found that subjects who were forewarned that they would see graphic, negative footage afterward reported feeling more upset and distressed than subjects told nothing or that they would not see footage of that nature. There are multiple methodological differences between those experiments and ours that may be responsible for this discrepancy. For example, their clips were of much long duration than ours, meaning their subjects had a longer period over which to anticipate the warned-about, negative aspects of the material. Future work could attempt to address this issue; but another possible explanation is that the small effects they found, using small samples and complex designs, do not replicate.

In addition, at first glance, our findings do not seem to fit with response expectancy theory (Kirsch, 1985, 1997). We found that trigger warnings changed

how negative people expected the material to follow would be, yet we did not find that trigger warnings changed how negative people felt, or the frequency with which they experienced symptoms of distress afterwards. One possible explanation for this discord is that, although we changed people's belief about what the material would be like, we did not change their expectancies about how they would respond to that material. Future research could examine this possibility by first developing trigger warnings that explicitly describe the symptoms that negative material usually elicits, and which alter people's expectancies about their likelihood of experiencing those symptoms, and then testing the effects such warnings have on the rates of symptoms people report.

Limitations and Future Directions

Of course, it is possible that trigger warnings do make a meaningful difference to people's symptoms and related outcomes, but for a variety of reasons we failed to observe their effects. One obvious reason trigger warnings may have exerted little influence is that our subjects did not notice the warnings. But, as outlined in Chapter 2, we have evidence to suggest that supposition is not true: A large majority of subjects in the "warning" conditions across experiments said, at the end of the experiment, that they remembered earlier seeing the trigger warning (and informal inspection of their written explanations describing the warning they saw suggested that most subjects accurately recalled the gist of the warning). What is more, in Experiment 3 we found that subjects who saw a trigger warning expected the material to follow would be more negative—demonstrating that they did notice and take on board that warning. Together, these findings suggest we

cannot explain the consistently small differences between warning conditions as being due to subjects simply not noticing the trigger warning manipulation.

Another counterexplanation for our finding that trigger warnings are unhelpful is that our materials were not sufficiently negative or symptom-evoking, meaning that there was little room for warnings to reduce how negative subjects felt or the frequency with which they experienced symptoms, and so the helpful effects of warnings were obscured. There are two reasons that is not a satisfying explanation for our results.

First, if we consider the responses of unwarned subjects who saw negative material in Table 4, we can see that the lower limits of the 95% CIs around their reported levels of negative affect, or other symptoms, did not overlap with the bottom of the scale in any of the experiments. Put another way, people who did not get a trigger warning (and were therefore our baseline condition) reported having high enough degrees of negative affect and symptoms in response to our negative materials that those rates could have been reduced by trigger warnings—but were not (it is also worth noting that neither did the upper limits of the confidence intervals approach the scale maximums, and so there was also room to observe effects in a harmful direction).

Second, we found that subjects who instead saw a control film clip reported even lower rates of these symptoms, again demonstrating that the effects in the negative condition were not at floor. That said, it is true that the negative films did not produce dramatically higher levels of negative affect and symptoms than the negative stories had. To more definitely rule out this explanation, future research could examine the effects of trigger warnings placed prior to extremely negative

materials—although whether it would be ethical to show people such material without acquiring their informed consent beforehand would require careful consideration.

A further counterexplanation for our failure to observe any notable effects of trigger warnings is that we neglected to measure symptoms or behaviours that trigger warnings really did influence. For example, some subjects may have withdrawn from our experiments upon getting the trigger warning, thereby avoiding the negative material and any ensuing negative emotions. We might, of course, think of such behaviour as an example of successful emotion regulation, via a "situation selection" strategy (Gross, 2015). But, equally, we might see the trigger warning as doing its job, encouraging subjects to withdraw from the study to avoid the consequences they fear. In which case, these "drop out" subjects would in fact be exhibiting avoidance—itself another symptom of PTSD (see *criterion C* in Table 1; APA, 2013). Therefore, if trigger warnings led subjects to withdraw for that reason, it would suggest trigger warnings were acting harmfully, rather than helpfully.

No matter the reason subjects might have withdrawn upon seeing a trigger warning, we have evidence that very few subjects in fact did so. To gather this evidence, we examined responses from subjects who quit our Qualtrics surveys before they had completed every task (and who were therefore excluded from our other analyses), classifying those subjects according to which warning condition they had been assigned to, and what the last survey page they saw was, before they stopped responding and withdrew from the experiment. If many "warning"

subjects quit upon seeing the trigger warning, that would suggest those subjects had used the trigger warning as an opportunity to avoid the negative material.

We instead found that the number of "warning" subjects who quit specifically upon seeing the warning were: none in Experiment 1a, none in Experiment 1b, none in Experiment 2a, nine in Experiment 2b, six in Experiment 3, and one in Experiment 4 (Appendix E gives more information about how many and when subjects quit each experiment). Compared to the nearly 1900 subjects who did complete these experiments (some of whom were then excluded for reasons other than non-completion), and the hundreds more who started the experiments but quit them at some point other than upon seeing a trigger warning, these numbers are vanishingly small. These data suggest that few, if any, subjects used the trigger warning to avoid negative material—a finding that fits with previous work suggesting that people sometimes choose to have a negative experience, out of curiosity (Hsee & Ruan, 2016).

Inspecting these data further revealed that similar numbers of subjects withdrew from the "warning" condition as from the the "no warning" condition overall. Assuming that subjects were more likely to drop out the more they were being bothered by symptoms of distress following the negative material, if trigger warnings reduced symptoms then subjects would have been more likely to drop out of the "no warning" condition, whereas if trigger warnings worsened symptoms then subjects would have been more likely to drop out of the "no warning" condition. But that similar numbers of subjects withdrew from each condition further fits with our conclusion that trigger warnings instead had trivial effects.

It is still true that there are yet other behaviours or symptoms that we did not measure, but which trigger warnings could have affected to a non-trivial degree. For one, we did not ask subjects anything about the content or phenomenology of their intrusions. Yet it is possible that warnings may have altered both of those—for instance, previous research found that telling people an experience they had was particularly negative increased the sense of reliving and vividness with which they remembered that event, and how stressed they felt about that event (Takarangi & Strange, 2010). Therefore, although trigger warnings do not change how often people have thoughts related to the negative material, they may make those thoughts more vivid and distressing.

Moreover, we asked people only about their intrusion and avoidance symptoms, and did not measure their hyperarousal symptoms (see *criterion E* in Table 1; APA, 2013). It is therefore possible we would have observed effects of trigger warnings on physiological measures, such as people's heart-rate, or the sensitivity of their startle reflex, had we measured them (see, for example, Cantor et al., 1984; Lepore, Fernandez-Berrocal, Ragan, & Ramos, 2004).

Relatedly, we relied on people to self-report their symptoms. We did not, for example, periodically ask people "just now, were you thinking of something related to the film clip?" Yet previous research shows that people are not always meta-aware that their minds have turned to negative memories (Baird et al., 2013; Takarangi, Strange et al., 2014). We did measure subjects' comprehension of the article they read, but this task only indirectly measures the incidence of such off-task thoughts. Therefore, it is still possible that trigger warnings affect rates of symptoms of which people are not meta-aware.

Future research could examine the effects of trigger warnings on these other symptoms, by including additional measures. But as things stand, one important caveat to our conclusion—that trigger warnings are trivially effective—is that we cannot speak to trigger warnings' effects on variables we did not measure.

A second important caveat to our conclusion is that we do not know how well our findings generalise to clinical populations of people diagnosed with PTSD, or people with a history of trauma, related to the negative material. In Experiment 4 we found that much of our sample had experienced one or more traumatic events, reflective of the population at large (Breslau et al., 1998). But we were not able to look at the effect of trigger warnings on people who had experienced a traumatic event similar to the one depicted in the negative material we showed them.

Of course, given that these warnings are being rolled out in a widespread way—for example, appearing in syllabi that all students receive (Palmer, 2017)—it is important to know what effects such warnings have on most people. But future research should next address what effects trigger warnings have on the minority of people who may be particularly susceptible to being distressed by a given piece of negative material, due to having experienced a similar traumatic event that is already causing them persistent distress, or even PTSD. Trigger warnings may have even less chance of helping those people. For instance, a meta-analysis found that people with more severe PTSD symptoms are worse at emotional regulation (Seligowski, Lee, Bardeen, & Orcutt, 2015). This finding suggests that a non-specific prompt to regulate one's emotions—in the form of a trigger warning—would be less helpful for people who already experience

significant symptoms of distress than it would for people with greater ability to regulate their emotions.

Finally, a third important caveat to our conclusion is that we did not address the effects of trigger warnings over time—either the effect of one trigger warning after a longer delay, or the cumulative effects of seeing multiple trigger warnings, spread across multiple materials, over multiple occasions. But in each of these scenarios, there is reason to suspect that trigger warnings could have harmful effects.

In the case of a single trigger warning, warned people may come to remember the warned-about material as more negative over time, and experience more symptoms related to it, compared to unwarned people. This change could happen if knowing that the material was deemed "warning-worthy" leads warned people's memory to become distorted in line with that knowledge. Such an effect would fit with previous experiments showing that social feedback about a negative experience, given immediately afterward, can change how people later feel about that experience after a delay, how accurately they remember it, and the degree of PTSD symptoms they experience following it (Lepore et al., 2004; Takarangi, Segovia, Dawson, & Strange, 2014; Takarangi & Strange, 2010). This effect would further fit with work showing an association between increases in how negatively people remember a traumatic experience and increases in their rate of PTSD symptoms (Engelhard et al., 2008; Southwick et al., 1997).

Repeated and widespread use of trigger warnings—a scenario that a student could very possibly encounter, in a semester filled with trigger warnings—could have the cumulative effect of teaching her to look for the negative side of all

materials and situations she encounters. Typically, people more prone to anxiety are more likely to pay attention to the negative aspects of situations, and interpret emotionally ambiguous situations in a negative way (for a review, see Mathews & MacLeod, 2005). But repeated practice at not processing stimuli in these negative ways can reduce these tendencies in people, and this reduction is associated with reductions in people's symptoms of anxiety (for reviews, see Hertel & Mathews, 2011; Koster, Fox, & MacLeod, 2009). If trigger warnings do the opposite, essentially leading people to practice preferentially attending to negative information and resolving ambiguity in a negative way, then trigger warnings may promote symptoms of anxiety. This prospect is troubling, given that anxiety among college students already appears to be on the rise (Center for Collegiate Mental Health, 2016).

Implications and New Questions

Despite these caveats, our conclusion—that trigger warnings did not reduce people's distress—has implications for theory. Our findings fit within a framework explaining why people's thinking and behaviour are often influenced in unwanted ways: People may be unaware of the influence, they may not know the size or direction of the influence, or they may not want or be unable to counter the influence (Wilson & Brekke, 1994). Failing at any one of these points stymies people's ability to counter the influence. An important question our finding raises is: where along this process do trigger warnings stumble?

Our results suggest trigger warnings make people broadly aware of the influence that the negative material to follow will have—after all, they judge that the material to follow will be very negative. But perhaps trigger warnings are not

sufficiently informative regarding the specific effects negative material tends to have, or regarding the magnitude of its influence. Alternatively, trigger warnings may not be worded in a way that makes people want to expend the effort to counter the influence. Finally, it may be that people are not able to exert control over the processes that give rise to these unwanted influences.

This latter possibility seems unlikely, though, given that people are able to regulate their emotions to some degree, and (at least in the short term) reduce the frequency with which a particular thought comes to mind (Gross, 2015; Wenzlaff & Wegner, 2000). In any case, addressing this question, regarding why trigger warnings are ineffective, would be a useful step towards finding an alternative intervention that does reduce people's distress.

One such intervention could be a warning that instead downplays how negative the material to follow is, and positively changes people's expectancies about how they will react to the negative material. For example, the warning could say that the story to follow describes child abuse in detail, but most people do not find it overly upsetting, and find it easy to stop thinking about it afterward. A successful outcome using this kind of warning would with fit with previous research showing that social feedback after a negative experience, which conveys that the experience was not so bad, tends to reduce how upset people report feeling, and the symptoms they go on to experience (Lepore et al., 2004; Takarangi, Segovia et al., 2014).

Another such intervention could be a warning that includes an invitation for people to regulate their negative emotions while reading (or watching) the material, and gives detailed instructions about an emotion regulation strategy they

could use to do so. For example, people could be specifically told to reappraise the negative aspects of the story (Feinberg et al, 2012; Gross, 2015). Of course, a warning worded along these lines has ceased to resemble the prototypical trigger warning. Further, suggesting that educators should attempt to write and administer new warnings such as these is approaching the suggestion that they engage in therapeutic practices for which they are not trained.

What is more, it may be that these "positive warnings" still would not help those for whom they are particularly intended—those who have previously experienced a related negative experience, which causes them ongoing distress. That is, people who experience ongoing posttraumatic distress may have the most maladaptive, and difficult-to-shift expectancies about how they will react. These people may also be particularly poor at effectively regulating their emotions. Both these factors would make it harder for these interventions to help these people (Rief et al., 2015; Rubin, Berntsen et al., 2008; Rubin et al., 2011; Seligowski et al., 2015; Takarangi et al., 2017). Future work could examine which individual differences predict how much people are affected by warnings prior to negative material.

We attempted to investigate one individual difference. In Experiment 4, we asked about people's history of trauma. We found that almost everybody had experienced a potentially traumatic event, but fewer had experienced ongoing distress as a result. Although we were not able to go on and make the comparisons we had intended, this finding alone accords with the model of PTSD proposing that events themselves are less important in producing the disorder than are people's reactions to and memories for their experiences (Berntsen et al., 2008;

Rubin, Berntsen et al., 2008; Rubin et al., 2011). Future work could investigate the differential rates of symptoms reported—and the effect of trigger warnings on them—by people who have and have not experienced ongoing distress in response to a previous negative experience.

That some students want (and professors use) trigger warnings despite these warnings being ineffective—at least, as far as our data show—raises the question: if these warnings do not work, why do some people still think they are useful? This apparent mismatch fits with other work showing that people are bad at predicting the effects interventions will have on them, and at accurately recognising the factors that determined their behaviour in a given situation (Michael et al., 2012; Nisbett & Wilson, 1977).

In particular, the trigger warnings scenario is reminiscent of work on "learning styles." Many students believe that their learning is optimised when material is taught in their preferred style—such as visually, rather than verbally (and, conversely, that their learning is impaired when it is not; for a review, see Pashler, McDaniel, Rohrer, & Bjork, 2008). In line with this belief, people judge that they have better learned material when there is a match between their preferred style and the delivery style (Knoll, Otani, Skeel, & Van Horn, 2017). But, in actual fact, how well people learn material is not related to the degree of this match (Knoll et al., 2017; Pashler et al., 2008).

Our findings also have practical implications. Our experiments bring the first experimental data to the "never ending debate" about the use of trigger warnings in higher education (Flaherty, 2015). Our results are bad news for educators who want to help ease distress their students may feel, because our

results suggest trigger warnings do not have that effect. At the same time, our data suggest that neither do trigger warnings harm students, as others feared, and so in that way our result constitute good news.

More broadly, our findings also have implications for other situations in which people have negative experiences about which they could be forewarned. For instance, reporters usually tell viewers (or readers) if their news story contains graphic details, and other television shows are typically preceded by information about their rating and brief explanation of their content. Our findings suggest that media outlets are not inadvertently raising their chances of upsetting their audience by providing them with such warning information.

Relatedly, institutional ethics committees ordinarily require that subjects give informed consent to participate in studies. Therefore, if a study involves exposure to, say, very negative photographs then subjects must be informed beforehand that they will see photographs like that and the experience might upset them. Our findings suggest that reading such a warning on a consent form would not intensify the negativity of subjects' experience.

Finally, people are sometimes traumatised by their experiences as jurors, finding experiences such as seeing gruesome evidence very upsetting (Lonergan, Leclerc, Descamps, Pigeon, & Brunet, 2016). Unlike television audiences, or experimental subjects, jurors cannot easily opt out of exposure to negative material. Some jurisdictions forewarn jurors that they are about to hear or see some distressing evidence (Bright & Goodman-Delahunty, 2004). Our results suggest that such an instruction would be unlikely to help jurors, but nor would it worsen their experience.

Where does this outcome leave universities, grappling with the issue of making policies around the use of trigger warnings? If trigger warnings are merely unhelpful then, on the one hand, perhaps professors should use them. Professors could treat trigger warnings as a benign concession to students who still feel as though trigger warnings are helpful, and are pleased to have them. On the other hand, perhaps professors should not use them. If professors and students continue to believe and behave as though trigger warnings are helpful—by using them—then they may be less open to looking beyond the use of trigger warnings. This set of experiments paves the way toward looking for such alternative solutions that will effectively reduce unnecessary distress.

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97,* 129-133. doi:10.1037/0033-2909.97.1.129

Aldao, A., Nolen-Hoeksema, S., & Schweizer, S. (2010). Emotion-regulation strategies across psychopathology: A meta-analytic review. *Clinical Psychology Review, 30,* 217-237. doi:10.1016/j.cpr.2009.11.004

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13,* 219-235. doi:10.1177/1088868309341564

American Association of University Professors. (2014). *On Trigger Warnings.* Retrieved July 30, 2015, from https://www.aaup.org/report/trigger-warnings

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

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Anderson, R. C., & Pichert J. W. (1978). Recall of previously unrecallable information following a shift in perspective. *Journal of Verbal Learning and Verbal Behavior, 17,* 1-12. doi:10.1016/S0022-5371(78)90485-1

Baird, B., Smallwood, J., Fishman, D. J. F., Mrazek, M. D., & Schooler, J. W. (2013). Unnoticed intrusions: Dissociations of meta-consciousness in thought suppression. *Consciousness and Cognition, 22,* 1003-1012. doi: 10.1016/j.concog.2013.06.009

Bennett, P., Phelps, C., Brain, K., Hood, K., & Gray, J. (2007). A randomized controlled trial of a brief self-help coping intervention designed to reduce

distress when awaiting genetic risk information. *Journal of Psychosomatic Research, 63,* 59-64. doi:10.1016/j.jpsychores.2007.01.016

Bentley, M. (2016). Scenes of a disturbing nature: trigger warnings. *Critical Studies on Security, 4,* 114-117. doi:10.1080/21624887.2016.1163932

Bentley, M. (2017). Trigger warnings and the student experience. *Politics, 37,* 470-485. doi:10.1177/0263395716684526

Berntsen, D., & Hall, N. M. (2004). The episodic nature of involuntary autobiographical memories. *Memory & Cognition, 32,* 789-803. doi: 10.3758/BF03195869

Berntsen, D., & Rubin, D. C. (2006). The centrality of event scale: A measure of integrating a trauma into one's identity and its relation to post-traumatic stress disorder symptoms. *Behaviour Research and Therapy, 44,* 219-231. doi:10.1016/j.brat.2005.01.009

Berntsen, D., & Rubin, D. C. (2008). The reappearance hypothesis revisited: Recurrent involuntary memories after traumatic events and in everyday life. *Memory & Cognition, 36,* 449-460. doi:10.3758/MC.36.2.449

Berntsen, D., Rubin, D. C., & Bohni, M. K. (2008). Contrasting models of posttraumatic stress disorder: Reply to Monroe and Mineka (2008). *Psychological Review, 115,* 1099-1107. doi:10.1037/a0013730

Berntsen, D., Staugaard, S. R., & Sørensen, L. M. T. (2013). Why am I remembering this now? Predicting the occurrence of involuntary (spontaneous) episodic memories. *Journal of Experimental Psychology: General, 142,* 426-444. doi:10.1037/a0029128

Bisson, J. I., Roberts, N. P., Andrew, M., Cooper, R., & Lewis, C. (2013). Psychological therapies for chronic post-traumatic stress disorder (PTSD) in adults. *Cochrane Database of Systematic Reviews, 12*, CD003388. doi: 10.1002/14651858.CD003388.pub4

Boysen, G. A., Wells, A. M., & Dawson, K. J. (2016). Instructors' use of trigger warnings and behavior warnings in abnormal psychology. *Teaching of Psychology, 43,* 334-339. doi:10.1177/0098628316662766

Bradbury, R. (1951). The Veldt. In *The Illustrated Man.* New York: Doubleday & Company.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11,* 717-726. doi:10.1016/ S0022-5371(72)80006-9

Breslau, N., Kessler, R. C., Chilcoat, H. D., Schultz, L. R., Davis, G. C., & Andreski, P. (1998). Trauma and posttraumatic stress disorder in the community: The 1996 Detroit area survey of trauma. *Archives of General Psychiatry, 55,* 626-632. doi:10.1001/archpsyc.55.7.626

Brewin, C. R., Dalgleish, T., & Joseph, S. (1996). A dual representation theory of posttraumatic stress disorder. *Psychological Review, 103,* 670-686. doi: 10.1037/0033-295X.103.4.670

Brewin, C. R., & Holmes, E. A. (2003). Psychological theories of posttraumatic stress disorder. *Clinical Psychology Review, 23*, 339-376. doi:10.1016/ S0272-7358(03)00033-3

Bright, D. A., & Goodman-Delahunty, J. (2004). The influence of gruesome verbal evidence on mock juror verdicts. *Psychiatry, Psychology and Law, 11,* 154-166. doi:10.1375/pplt.2004.11.1.154

Bryson, B. (2005). *A short history of nearly everything.* London: Doubleday.

Canadian Association of University Teachers. (2015, May). *Trigger warnings.* Retrieved September 3, 2016, from https://www.caut.ca/about-us/caut-policy/lists/caut-policy-statements/trigger-warnings

Cantor, J., Ziemke, D., & Sparks, G. C. (1984). Effect of forewarning on emotional responses to a horror film. *Journal of Broadcasting, 28,* 21-31 doi:10.1080/08838158409386512

Carlson, E. B., Smith, S. R., Palmieri, P. A., Dalenberg, C., Ruzek, J. I., Kimerling, R., Burling, T. A., & Spain, D. A. (2011). Development and validation of a brief self-report measure of trauma exposure: The Trauma History Screen. *Psychological Assessment, 23,* 463-477. doi:10.1037/a0022294

Catanzaro, S. J., & Greenwood, G. (1994). Expectancies for negative mood regulation, coping, and dysphoria among college students. *Journal of Counseling Psychology, 41,* 34-44. doi:10.1037/0022-0167.41.1.34

Center for Collegiate Mental Health. (2016, January). *2015 annual report* (Publication No. STA 15-108). Retrieved December 14, 2017, from the Penn State University's Counseling and Psychological Services website https://sites.psu.edu/ccmh/files/2017/10/2015_CCMH_Report_1-18-2015-yq3vik.pdf

Christianson, S.-Å. (1992). Emotional stress and eyewitness memory: A critical review. *Psychological Bulletin, 112,* 284-309. doi: 10.1037/0033-2909.112.2.284

Committee on Treatment of Posttraumatic Stress Disorder, Institute of Medicine of the National Academies (2008). *Treatment of posttraumatic stress disorder: An assessment of the evidence.* Washington, D.C.: The National Academies Press.

Crane, S. (1901, March). A dark-brown dog. *Cosmopolitan, 30,* 481-486.

Cuijpers, P., Turner, E. H., Koole, S. L., van Dijke, A., & Smit, F. (2014). What is the threshold for a clinically relevant effect? The case of major depressive disorders. *Depression and Anxiety, 31,* 374-378. doi:10.1002/da.22249

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York: Routledge.

Daro, I. N. (2016, January 6). The government of Canada added a trigger warning to its website. *BuzzFeed.* Retrieved September 8, 2016, from https:// www.buzzfeed.com/ishmaeldaro/government-of-canada-added-trigger-warning-to-mmiw-page

Devue, C., Belopolsky, A. V., & Theeuwes, J. (2011). The role of fear and expectancies in capture of covert attention by spiders. *Emotion, 11,* 768-775. doi:10.1037/a0023418

de Wied, M., Hoffman, K., & Roskos-Ewoldsen, D. R. (1997). Forewarning of graphic portrayal of violence and the experience of suspenseful drama. *Cognition and Emotion, 11,* 481-494. doi:10.1080/026999397379890

Ehlers, A., & Clark, D. M. (2000). A cognitive model of posttraumatic stress disorder. *Behaviour Research and Therapy, 38,* 319-345. doi:10.1016/S0005-7967(99)00123-0

Eich, E. (2014). Business not as usual. *Psychological Science, 25,* 3-6. doi: 10.1177/0956797613512465

Engelhard, I. M., van den Hout, M. A., & McNally, R. J. (2008). Memory consistency for traumatic events in Dutch soldiers deployed to Iraq. *Memory, 16,* 3-9, doi:10.1080/09658210701334022

Faasse, K. & Petrie, K. J. (2016). From me to you: The effect of social modeling on treatment outcomes. *Current Directions in Psychological Science, 25,* 438-443. doi:10.1177/0963721416657316

Feinberg, M., Willer R., Antonenko, O., & John, O. P. (2012). Liberating reason from the passions: Overriding intuitionist moral judgments through emotion reappraisal. *Psychological Science, 23,* 788-795. doi: 10.1177/0956797611434747

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*, 532-538. doi:10.1037/a0015808

Flaherty, C. (2015, December 3). The never-ending trigger-warning debate. *Slate.* Retrieved December 7, 2015, from http://www.slate.com/articles/life/inside_higher_ed/2015/12/trigger_warning_debate_some_professors_say_they_build_trust_others_say_they.html

Foa, E. B., & McLean, C. P. (2016). The efficacy of exposure therapy for anxiety-related disorders and its underlying mechanisms: The case of OCD and PTSD. *Annual Review of Clinical Psychology, 12,* 1-28. doi:10.1146/annurev-clinpsy-021815-093533

Forstie, C. (2016). Trigger warnings. In N. M. Rodriguez, W. J. Martino, J. C. Ingrey, & E. Brockenbrough (Eds.) *Critical Concepts in Queer Studies and Education* (pp. 421-433). New York: Palgrave Macmillan. doi: 10.1057/978-1-137-55425-3_40

Friedersdorf, C. (2016, August 31). Grading the University of Chicago's Letter on Academic Freedom. *The Atlantic.* Retrieved September 2, 2016, from http://www.theatlantic.com/politics/archive/2016/08/grading-the-university-of-chicagos-letter-on-academic-freedom/497804/

Galatzer-Levy, I. R., & Bryant, R. A. (2013). 636,120 ways to have posttraumatic stress disorder. *Perspectives on Psychological Science, 8,* 651-662. doi: 10.1177/1745691613504115

Geraerts, E., Kozarić-Kovačić, D., Merckelbach, H., Peraica, T., Jelicic, M., & Candel, I. (2007). Traumatic memories of war veterans: Not so special after all. *Consciousness and Cognition, 16,* 170-177. doi:10.1016/j.concog.2006.02.005

Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition, 10,* 597-602. doi:10.3758/BF03202442

Goldin, P. R., Ziv, M., Jazaieri, H., Werner, K., Kraemer, H., Heimberg, R. G., & Gross, J. J. (2012). Cognitive reappraisal self-efficacy mediates the effects

of individual cognitive-behavioral therapy for social anxiety disorder. *Journal of Consulting and Clinical Psychology, 80,* 1034-1040. doi: 10.1037/a0028555

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology, 79,* 701-721. doi:10.1037/0022-3514.79.5.701

Grieve, P. (2016, August 24) University to freshmen: Don't expect safe spaces or trigger warnings. *The Chicago Maroon.* Retrieved December 14, 2017 from https://www.chicagomaroon.com/article/2016/8/24/university-to-freshmen-dont-expect-safe-spaces-or-trigger-warnings/

Gross, J. J. (2015). Emotion regulation: Current status and future prospects. *Psychological Inquiry, 26,* 1-26. doi:10.1080/1047840X.2014.940781

Gross, J. J., & Jazaieri, H. (2014). Emotion, emotion regulation, and psychopathology: An affective science perspective. *Clinical Psychological Science, 2,* 387-401. doi:10.1177/2167702614536164

Gross, J. J., & Levenson, R. W. (1997). Hiding feelings: The acute effects of inhibiting negative and positive emotion. *Journal of Abnormal Psychology, 106,* 95-103. doi:10.1037/0021-843X.106.1.95

Gross, J. J., & Thompson, R. A. (2007). Emotion regulation: Conceptual foundations. In J. J. Gross (Ed.), *Handbook of emotion regulation* (pp. 3-24). New York: Guilford Press.

Gust, O. (2016, June 14). I use trigger warnings – but I'm not mollycoddling my students. *The Guardian.* Retrieved September 3, 2016, from https://

www.theguardian.com/higher-education-network/2016/jun/14/i-use-trigger-warnings-but-im-not-mollycoddling-my-students

Hall, N. M., & Berntsen, D. (2008). The effect of emotional stress on involuntary and voluntary conscious memories. *Memory, 16,* 48-57. doi: 10.1080/09658210701333271

Hanlon, A. R. (2015, May 17). My students need trigger warnings—and professors do, too. *New Republic.* Retrieved May 19, 2015, from http:// www.newrepublic.com/article/121820/my-students-need-trigger-warnings-and-professors-do-too

Harvey, A. G., & Bryant, R. A. (1998). The effect of attempted thought suppression in acute stress disorder. *Behaviour Research and Therapy, 36,* 583-590. doi:10.1016/S0005-7967(98)00052-7

Hertel, P. T., & Mathews, A. (2011). Cognitive bias modification: Past perspectives, current findings, and future applications. *Perspectives on Psychological Science, 6,* 521-536. doi:10.1177/1745691611421205

Holmes, E. A., & Bourne, C. (2008). Inducing and modulating intrusive emotional memories: A review of the trauma film paradigm. *Acta Psychologica, 127,* 553-566 doi:10.1016/j.actpsy.2007.11.002

Horowitz, M. J., Wilner, N., & Alvarez, W. (1979). Impact of Event Scale: A measure of subjective stress. *Psychosomatic Medicine, 41,* 209-218. doi: 10.1097/00006842-197905000-00004

Hsee, C. K, & Ruan, B. (2016). The Pandora effect: The power and peril of curiosity. *Psychological Science, 27,* 659-666. doi: 10.1177/0956797616631733

James, E. L., Lau-Zhu, A., Clark, I. A., Visser, R. M., Hagenaars, M. A., & Holmes, E. A. (2016). The trauma film paradigm as an experimental psychopathology model of psychological trauma: intrusive memories and beyond. *Clinical Psychology Review, 47,* 106-142. doi:10.1016/j.cpr.2016.04.010

Jamieson, J. P., Nock, M. K., & Mendes, W. B. (2013). Changing the conceptualization of stress in social anxiety disorder: Affective and physiological consequences. *Clinical Psychological Science, 1,* 363-374. doi:10.1177/2167702613482119

Jarvie, J. (2014, March 3). Trigger happy. *New Republic.* Retrieved September 15, 2016, from https://newrepublic.com/article/116842/trigger-warnings-have-spread-blogs-college-classes-thats-bad

Johnson, M. K., Foley, M. A., Suengas, A .G., & Raye, C. L. (1988). Phenomenal characteristics of memories for perceived and imagined autobiographical events. *Journal of Experimental Psychology: General, 117,* 371-376. doi: 10.1037/0096-3445.117.4.371

Kamenetz, A. (2016, September 7). Half of professors in NPR Ed survey have used 'trigger warnings'. *National Public Radio.* Retrieved September 9, 2016, from http://www.npr.org/sections/ed/2016/09/07/492979242/half-of-professors-in-npr-ed-survey-have-used-trigger-warnings

Kane, M. J., & McVay, J. C. (2012). What mind wandering reveals about executive-control abilities and failures. *Current Directions in Psychological Science, 21,* 348-354. doi:10.1177/0963721412454875

Kensinger, E. A., & Schacter, D. L. (2008). Memory and emotion. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of Emotions* (pp. 601-617). New York: Guilford Press.

Kirsch, I. (1985). Response expectancy as a determinant of experience and behavior. *American Psychologist, 40,* 1189-1202. doi:10.1037/0003-066X.40.11.1189

Kirsch, I. (1997). Response expectancy theory and application: A decennial review. *Applied and Preventive Psychology, 6,* 69-79. doi:10.1016/S0962-1849(05)80012-5

Kirsch, I. (2004). Conditioning, expectancy, and the placebo effect: Comment on Stewart-Williams and Podd (2004). *Psychological Bulletin, 130*, 341-343. doi:10.1037/0033-2909.130.2.341

Kirsch, I., & Lynn, S.J. (1999). Automaticity in clinical psychology. *American Psychologist, 54*, 504-515. doi:10.1037/0003-066X.54.7.504

Kleinsorge, T. (2007). Anticipatory modulation of interference induced by unpleasant pictures. *Cognition and Emotion, 21*, 404-421. doi: 10.1080/02699930600625032

Knoll, A. R., Otani, H., Skeel., R. L., & Van Horn, K. R. (2017). Learning style, judgements of learning, and learning of verbal and visual information. *British Journal of Psychology, 108,* 544-563. doi:10.1111/bjop.12214

Koster, E. H. W., Fox, E., & MacLeod, C. (2009). Introduction to the special section on cognitive bias modification in emotional disorders. *Journal of Abnormal Psychology, 118,* 1-4. doi:10.1037/a0014379

Kumagai, A. K., Jackson, B., & Razack, S. (2017). Cutting close to the bone: Student trauma, free speech, and institutional responsibility in medical education. *Academic Medicine, 92,* 318-323. doi:10.1097/ACM. 0000000000001425

Lang, F. R., Staudinger, U. M., & Carstensen, L. L. (1998). Perspectives on socioemotional selectivity in late life: How personality and social context do (and do not) make a difference. *Journal of Gerontology: Psychological Sciences, 53B*, 21-30. doi: 10.1093/geronb/53B.1.P21

Leavitt J. D., & Christenfeld, N. J. S. (2011). Story spoilers don't spoil stories. *Psychological Science, 22,* 1152-1154 doi:10.1177/0956797611417007

Leavitt J. D., & Christenfeld, N. J. S. (2013). The fluency of spoilers: Why giving away endings improves stories. *Scientific Study of Literature, 3,* 93-104. doi: 10.1075/ssol.3.1.09lea

Lepore, S. J., Fernandez-Berrocal, P., Ragan, J., & Ramos, N. (2004). It's not that bad: Social challenges to emotional disclosure enhance adjustment to stress, *Anxiety, Stress, & Coping, 17,* 341-361, doi: 10.1080/10615800412331318625

Lesh, M. (2016, August 20). Warning: This article contains ideas that offend. *The Spectator Australia.* Retrieved September 2, 2016, from http:// spectator.com.au/2016/08/warning-this-article-contains-ideas-that-offend/

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181-1209. doi:10.1037/0003-066X.48.12.1181

Lonergan, M., Leclerc, M.-È., Descamps, M., Pigeon, S., & Brunet, A. (2016). Prevalence and severity of trauma- and stressor-related symptoms among jurors: A review. *Journal of Criminal Justice, 47,* 51-61 doi:10.1016/j.jcrimjus.2016.07.003

Lukianoff, G., & Haidt, J. (2015, September). The Coddling of the American Mind. *The Atlantic.* Retrieved August 12, 2015, from https://www.theatlantic.com/magazine/archive/2015/09/the-coddling-of-the-american-mind/399356/

Manne, K. (2015, September 19). Why I use trigger warnings. *The New York Times.* Retrieved September 21, 2015, from https://www.nytimes.com/2015/09/20/opinion/sunday/why-i-use-trigger-warnings.html

Manning, S., & Wace, C. (2016, May 7). Oxford law students too 'fragile' to hear about violent crime: Undergraduates given 'trigger warnings' before traumatic material. *Daily Mail.* Retrieved September 4, 2016, from http://www.dailymail.co.uk/news/article-3579086/Oxford-law-students-fragile-hear-violent-crime-Undergraduates-given-trigger-warnings-traumatic-material.html

Marcotte, A. (2013, December 30). The year of the trigger warning. *Slate.* Retrieved June 9, 2014, from http://www.slate.com/blogs/xx_factor/2013/12/30/trigger_warnings_from_the_feminist_blogosphere_to_shonda_rhimes_in_2013.html

Mathews, A., & MacLeod, C. (2005). Cognitive vulnerability to emotional disorders. *Annual Review of Clinical Psychology, 1,* 167-195. doi:10.1146/annurev.clinpsy.1.102803.143916

McManus, F., Sacadura, C., & Clark, D. M. (2008). Why social anxiety persists: An experimental investigation of the role of safety behaviours as a maintaining factor. *Journal of Behavior Therapy and Experimental Psychiatry, 39,* 147-161. doi:10.1016/j.jbtep.2006.12.002

McNally, R. J. (2014, May 20). Hazards ahead: The problem with trigger warnings, according to the research. *Pacific Standard.* Retrieved June 9, 2014, from https://psmag.com/education/hazards-ahead-problem-trigger-warnings-according-research-81946

Medina, J. (2014, May 17). Warning: The literary canon could make students squirm. *The New York Times.* Retrieved June 9, 2014, from https://www.nytimes.com/2014/05/18/us/warning-the-literary-canon-could-make-students-squirm.html

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of "far transfer": Evidence from a meta-analytic review. *Perspectives on Psychological Science, 11,* 512-534. doi: 10.1177/1745691616635612

Michael, R. B., Garry, M., & Kirsch, I. (2012). Suggestion, cognition, and behavior. *Current Directions in Psychological Science, 21,* 151-156. doi: 10.1177/0963721412446369

Mooneyham, B. W., & Schooler, J. W. (2013). The costs and benefits of mind-wandering: A review. *Canadian Journal of Experimental Psychology, 67,* 11-18. doi:10.1037/a0031569

National Coalition Against Censorship. (2015). *What's all this about trigger warnings?* Retrieved December 9, 2015, from http://ncac.org/wp-content/uploads/2015/11/NCAC-TriggerWarningReport.pdf

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231-259. doi: 10.1037/0033-295X.84.3.231

Oberlin College, Office of Equity Concerns. (2013). *Support resources for faculty.* Retrieved September 3, 2016, from https://web.archive.org/web/20131122144749/http://new.oberlin.edu/office/equity-concerns/sexual-offense-resource-guide/prevention-support-education/support-resources-for-faculty.dot

Ogle, C. M., Siegler, I. C., Beckham, J. C., & Rubin, D. C. (2017). Neuroticism increases PTSD symptom severity by amplifying the emotionality, rehearsal, and centrality of trauma memories. *Journal of Personality, 85,* 702-715. doi:10.1111/jopy.12278

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

Palmer, T. (2017, March 29). Monash University trigger warning policy fires up free speech debate. *Australian Broadcasting Corporation.* Retrieved April

18, 2017, from http://www.abc.net.au/news/2017-03-28/monash-university-adopts-trigger-warning-policy/8390264

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. *Psychological Science in the Public Interest, 9,* 105-119. doi:10.1111/j.1539-6053.2009.01038.x

Pichert J. W., & Anderson, R. C. (1977). Taking different perspectives on a story. *Journal of Educational Psychology, 69,* 309-315. doi: 10.1037/0022-0663.69.4.309

Planès, S., Villier, C., & Malleret, M. (2016). The nocebo effect of drugs. *Pharmacology Research and Perspectives, 4,* e00208. doi:10.1002/prp2.208

Porter, S., & Birt, A. R. (2001). Is traumatic memory *special*? A comparison of traumatic memory characteristics with memory for other emotional life experiences. *Applied Cognitive Psychology, 15,* S101-S117. doi:10.1002/acp.766

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112,* 160-164. doi:10.1037/0033-2909.112.1.160

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. doi:10.1146/annurev.psych.59.113006.095941

Qualtrics [survey software]. (2018). Retrieved from https://www.qualtrics.com/.

Rief, W., Glombiewski, J. A., Gollwitzer, M., Schubö, A., Schwarting, R., & Thorwart, A. (2015). Expectancies as core features of mental disorders. *Current Opinion in Psychiatry, 28,* 378-385. doi:10.1097/YCO.0000000000000184

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45,* 775-777. doi:10.1037/0003-066X.45.6.775

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,* 455-475. doi:10.1007/BF00977239

Rubin, D. C., Berntsen, D., & Bohni, M. K. (2008). A memory-based model of posttraumatic stress disorder: Evaluating basic assumptions underlying the PTSD diagnosis. *Psychological Review, 115*, 985-1011. doi:10.1037/a0013397

Rubin, D. C., Boals, A., & Berntsen, D. (2008). Memory in posttraumatic stress disorder: Properties of voluntary and involuntary, traumatic and nontraumatic autobiographical memories in people with and without posttraumatic stress disorder symptoms. *Journal of Experimental Psychology: General, 137,* 591-614. doi:10.1037/a0013165

Rubin, D. C., Dennis, M. F., & Beckham, J. C. (2011). Autobiographical memory for stressful events: The role of autobiographical memory in posttraumatic stress disorder. *Consciousness and Cognition, 20,* 840-856. doi:10.1016/j.concog.2011.03.015

Rubin, D. C., Schrauf, R. W., & Greenberg, D. L. (2003). Belief and recollection of autobiographical memories. *Memory & Cognition, 31,* 887-901. doi:10.3758/BF03196443

Schooler, J. W., Reichle, E. D., & Halpern, D. V. (2004). Zoning out while reading: Evidence for dissociations between experience and metaconsciousness. In D. T. Levin (Ed.), *Thinking and seeing: Visual*

metacognition in adults and children (pp. 203-226). Cambridge, MA: MIT Press.

Schwarz, K. A., Pfister, R., & Büchel, C. (2016). Rethinking explicit expectations: Connecting placebos, social cognition, and contextual perception. *Trends in Cognitive Sciences, 20,* 469-480. doi:10.1016/j.tics.2016.04.001

Seligowski, A. V., Lee, D. J., Bardeen, J. R. & Orcutt, H K. (2015). Emotion regulation and posttraumatic stress symptoms: A meta-analysis. *Cognitive Behaviour Therapy, 44,* 87-102. doi:10.1080/16506073.2014.980753

Sheskin, D. J. (2003). Inferential statistical tests employed with two independent samples (and related measures of association and correlation). In *Handbook of Parametric and Nonparametric Statistical Procedures* (3rd ed., pp. 373-423). Boca Raton, FL: CRC Press.

Smallwood, J., Nind, L., & O'Connor, R. C. (2009). When is your head at? An exploration of the factors associated with the temporal focus of the wandering mind. *Consciousness and Cognition, 18,* 118-125. doi:10.1016/j.concog.2008.11.004

Southwick, S. M., Morgan, C. A., III, Nicolaou, A. L., & Charney, D. S. (1997). Consistency of memory for combat-related traumatic events in veterans of Operation Desert Storm. *The American Journal of Psychiatry, 154*, 173-177. doi:10.1176/ajp.154.2.173

Suk Gersen, J. (2014, December 15). The trouble with teaching rape law. *The New Yorker.* Retrieved February 9, 2016, from http://www.newyorker.com/news/news-desk/trouble-teaching-rape-law

Sundin, E. C., & Horowitz, M. J. (2002). Impact of Event Scale: psychometric properties. *British Journal of Psychiatry, 180,* 205-209. doi:10.1192/bjp.180.3.205

Takarangi, M. K. T., Segovia, D. A., Dawson, E. & Strange, D. (2014). Emotional impact feedback affects how people remember an analogue trauma event. *Memory, 22,* 1041-1051, doi:10.1080/09658211.2013.865238

Takarangi, M. K. T., Smith, R. A., Strange, D., & Flowe, H. D. (2017). Metacognitive and metamemory beliefs in the development and maintenance of posttraumatic stress disorder. *Clinical Psychological Science, 5,* 131-140. doi:10.1177/2167702616649348

Takarangi, M. K. T., & Strange, D. (2010). Emotional impact feedback changes how we remember negative autobiographical experiences. *Experimental Psychology, 57,* 354-359. doi:10.1027/1618-3169/a000042

Takarangi, M. K. T., Strange, D., & Lindsay, D. S. (2014). Self-report may underestimate trauma intrusions. *Consciousness and Cognition, 27,* 297-305. doi:10.1016/j.concog.2014.06.002

Talarico, J. M., LaBar, K. S., & Rubin, D. C. (2004). Emotional intensity predicts autobiographical memory experience. *Memory & Cognition, 32,* 1118-1132. doi:10.3758/BF03196886

Talarico, J. M., & Rubin, D. C. (2003). Confidence, not consistency, characterizes flashbulb memories. *Psychological Science, 14,* 455-461. doi: 10.1111/1467-9280.02453

University of California, Santa Barbara, Associated Students Senate. (2014, February). *A resolution to mandate warnings for triggering content in*

academic settings (02262014:61). Retrieved on September 3, 2016, from https://www.as.ucsb.edu/senate/resolutions/a-resolution-to-mandate-warnings-for-triggering-content-in-academic-settings/

Vingiano, A. (2014, May 5). How the "trigger warning" took over the internet. *BuzzFeed.* Retrieved August 10, 2016, from https://www.buzzfeed.com/alisonvingiano/how-the-trigger-warning-took-over-the-internet

Waldman, K. (2014, May 7). Twitter is no place for trigger warnings. *Slate.* Retrieved August 10, 2016, from http://www.slate.com/blogs/xx_factor/2014/05/07/trigger_warnings_on_twitter_don_t_make_sense.html

Waldman, K. (2016, September 5). The trapdoor of trigger words. *Slate.* Retrieved September 8, 2016, from http://www.slate.com/articles/double_x/cover_story/2016/09/what_science_can_tell_us_about_trigger_warnings.html

Watson. D., & Clark. L. A. (1999). *The PANAS-X: Manual for the Positive and Negative Affect Schedule-Expanded Form.* Retrieved June 23, 2014, from University of Iowa, Department of Psychology website: http://ir.uiowa.edu/cgi/viewcontent.cgi?article=1011&context=psychology_pubs

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54,* 1063-1070. doi: 10.1037/0022-3514.54.6.1063

Wegner, D. M., Schneider, D. J., Carter, S. R., III, & White, T. L. (1987). Paradoxical effects of thought suppression. J*ournal of Personality and Social Psychology, 53,* 5-13. doi:10.1037/0022-3514.53.1.5

Wells, R. E., & Kaptchuk, T. J. (2012). To tell the truth, the whole truth, may do patients harm: The problem of the nocebo effect for informed consent. *The American Journal of Bioethics, 12,* 22-29. doi: 10.1080/15265161.2011.652798

Wenzlaff, R. M., & Wegner, D. M. (2000). Thought suppression. *Annual Review of Psychology, 51,* 59-91. doi:10.1146/annurev.psych.51.1.59

Wilson, R. (2015, September 14). Students' requests for trigger warnings grow more varied. *The Chronicle of Higher Education.* Retrieved September 15, 2015, from http://chronicle.com/article/article-content/233043/

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin, 116,* 117-142. doi:10.1037/0033-2909.116.1.117

Wyatt, W. (2016). The ethics of trigger warnings. *Teaching Ethics, 16,* 17-35. doi: 10.5840/tej201632427

Wythe, P. (2014, February 18). Trigger warnings needed in classroom. *The Daily Targum.* Retrieved September 4, 2016, from http://www.dailytargum.com/article/2014/02/trigger-warnings-needed-in-classroom

Appendix A

Below are the questions we asked student subjects in Experiment 1a who read the story "A Dark Brown Dog." The correct answers are in **bold**, but note that we varied which letter corresponded to the correct answer in the experiment.

1. The story can be interpreted as an allegory for the life awaiting newly freed slaves, symbolised by the dog. In this interpretation the father symbolises:
    a. The Northern states of America
    b. A progressive generation of white Americans
    c. The Southern states of America
    d. Another freed slave
2. What does the author say the dog feels when he is parted from the child at night?
    a. Boredom
    b. Excitement
    c. Despair
    d. Relief
3. How does the child probably feel at the end of the story?
    a. Gleeful
    b. Confused
    c. Indifferent
    d. Distraught
4. The author compares the relationship between the child and the dog to that between a monarch and a subject because:

a. The child is protected by the dog

b. The dog lords over the child

c. The child fulfils the dog's wishes

d. The dog is devoted to the child

5. Why does the dog become panicked while climbing the stairs?

a. The dog can see the father at the top

b. The dog is afraid of heights

c. The child is beginning to look menacing

d. The child is pulling him too quickly

Below are the questions we asked student subjects in Experiment 1a who read the story "The Veldt." The correct answers are in **bold**, but note that we varied which letter corresponded to the correct answer in the experiment.

1. Which of the following is NOT a piece of advice the psychologist gives George?

a. The children should spend more time outdoors

b. George should entirely demolish the nursery

c. The children should see the psychologist daily

d. George should switch off everything in the house

2. Why does Peter think it would be dreadful if the whole house got turned off?

a. He is afraid his parents will become too busy with housework to look after him

b. He is concerned that it would look like his family is poor

   c. **He is unwilling to do things for himself**
   d. He is reluctant to have to interact with his sister to play games

3. Which piece of furniture apologises to George?
   a. An armchair
   b. His bed
   c. **The dining table**
   d. The stove
4. The psychologist says that one of the original uses of rooms like the nursery was to:
   a. **See into children's minds in order to better help them**
   b. Create a fun environment for children to play in
   c. Reduce the need for babysitters
   d. Encourage families to spend more time together
5. Readers can infer that at the end of the story:
   a. The parents go on vacation without their children
   b. **The parents are killed by the lions in the nursery**
   c. The children lock themselves in the nursery
   d. The children begin therapy with the psychologist

Below are the questions we asked MTurk subjects in Experiment 1b who read the story "A Dark Brown Dog." The correct answers are in **bold**, but note that we varied which letter corresponded to the correct answer in the experiment.

1. The story can be interpreted as an allegory for the life awaiting newly freed slaves, symbolised by the dog. In this interpretation the father symbolises:
    a. The Northern states of America
    b. A progressive generation of white Americans
    c. The Southern states of America
    d. Another freed slave
2. The child first hits the dog because:
    a. The dog wouldn't stop barking
    b. The dog became overexcited
    c. The dog growled at the child
    d. The dog looked ugly
3. How does the child probably feel at the end of the story?
    a. Gleeful
    b. Confused
    c. Indifferent
    d. Distraught
4. The father's decision to keep the dog exemplifies his:
    a. Generosity
    b. Fondness for animals
    c. Spitefulness
    d. Recklessness
5. What does the author say the dog had nightmares about?
    a. Other dogs

b. His former owner

c. The family

d. The child

Below are the questions we asked MTurk subjects in Experiment 1b who read the story “The Veldt.” The correct answers are in **bold**, but note that we varied which letter corresponded to the correct answer in the experiment.

1. Readers can infer that the voices Lydia and George repeatedly hear screaming in the nursery are:

    a. The children’s voices

    b. Their own voices

    c. The voices of characters in the children’s storybooks

    d. None of the above

2. Which of the following is NOT stated as a side effect of the technology in the “Happylife Home”?

    a. The children becoming more compassionate

    b. The parents having fewer physical childcare duties to do

    c. George becoming more dependent on sedatives

    d. Lydia having less housework to do

3. What evidence does Lydia present to George to suggest that the automated house is bad for him?

    a. He spends more time at work

    b. He is grumpier with her and the children

    c. He is avoiding the nursery

d. He is smoking and drinking more

4. Which piece of furniture apologises to George?

    a. An armchair

    b. His bed

    c. The dining table

    d. The stove

5. Readers can infer that at the end of the story:

    a. The parents go on vacation without their children

    b. The parents are killed by the lions in the nursery

    c. The children lock themselves in the nursery

    d. The children begin therapy with the psychologist

Appendix B

At the end of Experiment 1a, we asked student subjects a few questions about the experiment and about themselves. The questions [and response options] were:

Briefly list specific details from the story (the first passage you read) that you found disturbing. [text entry box]

Had you read the story (the first passage) before? [Yes/No]

Had you read the article (the second passage) before? [Yes/No]

Did you see a warning before you read the first story? [Yes/No]

(if "Yes" selected) What did the warning say? [text entry box]

We told you that the purpose of this study was to examine the factors that affect comprehension of different writing styles. Do you think it could have been looking at anything else? [text entry box]

What results do you think we are expecting to find? [text entry box]

Outside of this experiment, have you encountered "trigger warnings"? [Yes/No]

What do you think a "trigger warnings" is? What do you think their purpose is? [text entry box]

Are there any other comments you would like to make about this experiment? [text entry box]

What is your age? [number entry box]

Are you: [Male/Female]

Have you lived in New Zealand since at least age 3? [Yes/No]

At the end of Experiment 2a, we asked student subjects the same questions as above, apart from the following changes: The wording was adapted where needed so that the questions referred to watching a film clip rather reading a story (and to the slightly different cover story we had used regarding the purpose of this experiment); and they answered the additional questions [with response options]:

Did you play the video more than once? [Yes/No]

Did you look away while the video was playing? [Yes/No]

Some subjects in Experiment 2a also completed an exploratory measure, just prior to the end of the experiment, which asked them to make ratings regarding thought suppression strategies. But, upon reflection, we decided that in adapting these items, their wording had become too confusing, and so we would not attempt to analyse or interpret those data.

Appendix C

At the start of Experiments 1b, 2b, 3, and 4, we gave MTurk subjects instructions about the conditions under which they should complete the experiment as follows:

> **During this experiment, we ask that you comply with the following experiment requirements:**
>
> 1. Please **maximize the size of your web browser** so that it covers your entire screen. Complete this experiment on a desktop computer, laptop computer, or large tablet, not on a mobile phone or similar device.
> 2. Please complete the experiment in a single session, and **do not leave the experiment to engage in other tasks**. So don't check your mail, look at Facebook, send or read a text message, get up for a drink, etc.
> 3. Please **do not use your web browser's back or refresh buttons** at any point during the experiment.
> 4. Because this experiment requires your close attention, we ask that you **complete the experiment in an environment that is free of noise and distraction**. Please do not speak to anyone, or have anyone near you. Ideally, you would be alone in a quiet room, or in a room where other people are quiet (such as a library).
>
> The reason we ask you to follow these instructions is to ensure the quality of the information you give us. We know from previous research that if you do take a break, chat with others, etc, it will impair your ability to do the tasks set in this experiment.

I understand these instructions, and agree to comply with them for the duration of the experiment [checkbox]

During Experiments 1b, 2b, 3, and 4, we included several attention checks. Specifically, MTurk subjects in Experiment 1b got the following checks:

At the end of the PANAS-X measure, visually presented as though they were additional items, the instructions "choose box 1" and "choose box 5" appeared (and the same thing, but different numbers, appeared at the end of the second PANAS-X). Subjects responded on the same rating scale as they had for the other, real items.

At the bottom of the page with the story on it, visually presented as though it was another sentence in the story, they saw the instruction "That is the end of the story. To show that you read all the way to the end, type the random word 'grain' on the next page." On the next page there was the instruction "Please type the random word here, or if you do not know it, just proceed to the next page." [text entry box]

At the end of the four-alternative forced-choice questions about the story they had read, subjects saw two additional questions. These were visually presented in the same way as the previous questions, but each had an instruction embedded in the question text. For example, "Near the end of the story ignore this question and simply choose option three below. Who was responsible?" Subjects selected one of the four options presented below the question.

MTurk subjects in Experiments 2b, 3, and 4 saw the same attention checks as subjects in Experiment 1b, apart from the following changes: They did not get the "random word" question, and the four-alternative forced-choice attention checks were adapted to be ostensibly about the article they had read. For example, "Near the start of the article ignore this question and simply choose option three below. Whom was it attributed to?"

At the end of Experiments 1b, 2b, 3, and 4, we asked MTurk subjects questions about the experiment, and about themselves—including how well they complied with our earlier instructions. Specifically, MTurk subjects in Experiment 1b saw the same end-of-experiment questions as student subjects in Experiment 1a (see Appendix B), apart from the following changes: They did not see the two open-ended questions about the purpose and expected results of the experiment, or the question about having lived in New Zealand; and they saw the additional questions [with response options]:

> Did you read the entire story (the first passage), from start to finish? [Yes/No]
>
> Did you read the story (the first passage) carefully, giving it your full attention? [Yes/No]
>
> Did you read the entire article (the second passage), from start to finish? [Yes/No]
>
> Did you read the article (the second passage) carefully, giving it your full attention? [Yes/No]

Did you use a search engine during the experiment to look up the story, the article, or answers to the comprehension questions? [Yes, I did use a search engine to look up the story and/or the article and/or answers to the comprehension questions / No, but I did use a search engine to look up information unrelated to the experiment / No, I did not use a search engine at any time during the experiment]

Did you maximize the size of your web browser so that it covers your entire screen? [Yes/No]

Did you complete the experiment on a mobile phone (or a similar device with a small screen)? [Yes/No]

Did you complete the experiment in a single session, without stopping? [Yes/No]

Did you pause or leave the experiment to engage in other tasks, even if they were other computer tasks? [Yes/No]

Did you complete the experiment without anyone helping you? [Yes/No]

Did you complete the experiment in an environment that is free of noise and distraction? [Yes/No]

Did you speak with anyone at any time during the experiment? [Yes/No]

Is English your first language? [Yes/No]

Have you ever studied psychology? [Yes/No]

What is your nationality? (i.e. which country or countries are you a citizen of) [text entry box]

MTurk subjects in Experiments 2b, 3, and 4 saw the same end-of-experiment questions as subjects in Experiment 1b, apart from the following changes: The

wording was adapted where needed so that the questions referred to watching a film clip rather reading a story (and to the slightly different cover story we had used regarding the purpose of this experiment); they also answered the same additional questions as student subjects in Experiment 2a (see Appendix B), as well as the following questions [with response options]:

Could you hear the video? [Yes/No]

Was the video image clear? [Yes/No]

Did you look away while the video was playing? [Yes/No]

Did you pause the video while it was playing? [Yes/No]

Did you play the video more than once? [Yes/No]

Did you play the video in full screen? [Yes/No]

Did you watch the video on the Youtube website (rather than in the survey window)? [Yes/No]

Did you turn on subtitles during the video? [Yes/No]

Did you change the video quality? [No, I left it at the default quality / Yes, changed it to 240p ／ Yes, I changed it to 360p ／ Yes, I changed it to 480p]

Table S1

Descriptive Statistics for Ratings of Material Classified by Presence of Warning and Negativity of Material

Rating	Experiment	Warning and negative material			No warning and negative material			Warning and control material			No warning and control material		
		M	95% CI LL	95% CI UL	*M*	95% CI LL	95% CI UL	*M*	95% CI LL	95% CI UL	*M*	95% CI LL	95% CI UL
Positive	1a	–	–	–	–	–	–						
	1b	1.84	1.62	2.07	2.00	1.75	2.25						
	2a	1.60	1.14	2.06	2.04	1.55	2.52	3.32	2.68	3.96	3.93	3.34	4.52
	2b	1.33	1.13	1.53	1.43	1.23	1.63	3.95	3.54	4.37	4.03	3.62	4.44
	3-pre$_a$	2.22	1.93	2.51	4.35	4.08	4.63	2.22	1.90	2.54	4.50	4.22	4.78
	3-post$_b$	1.34	1.17	1.51	1.36	1.16	1.57	4.44	4.04	4.83	3.96	3.55	4.37
	4	1.41	1.23	1.59	1.49	1.23	1.76	3.55	3.14	3.96	4.00	3.67	4.33
Surprising	1a	–	–	–	–	–	–						
	1b	5.05	4.72	5.39	5.45	5.08	5.81						
	2a	3.96	3.15	4.77	4.74	4.02	5.47	3.14	2.40	3.89	3.00	2.34	3.66
	2b	5.39	5.03	5.74	6.00	5.68	6.32	3.88	3.39	4.36	3.90	3.45	4.36
	3-pre$_a$	4.43	4.15	4.70	3.86	3.49	4.24	4.30	3.97	4.63	3.84	3.54	4.14
	3-post$_b$	4.73	4.35	5.12	6.07	5.79	6.35	4.62	4.18	5.05	4.12	3.70	4.54
	4	4.84	4.43	5.25	5.58	5.20	5.97	3.75	3.27	4.22	3.92	3.50	4.34
Interesting	1a	–	–	–	–	–	–						
	1b	4.95	4.56	5.34	5.46	5.06	5.87						
	2a	4.52	3.83	5.21	4.48	3.79	5.17	3.39	2.70	4.09	2.96	2.33	3.60
	2b	4.33	3.94	4.72	4.96	4.53	5.39	3.83	3.37	4.29	4.11	3.71	4.51
	3-pre$_a$	4.36	4.06	4.66	4.16	3.86	4.47	4.36	4.08	4.63	4.04	3.67	4.40
	3-post$_b$	4.07	3.67	4.48	4.30	3.87	4.73	3.47	3.05	3.88	3.67	3.30	4.04

	4	4.21	3.82	4.60	4.60	4.18	5.02	3.31	2.92	3.70	4.26	3.87	4.64
Unpleasant[c]	1a	–	–	–	–	–	–						
	1b	–	–	–	–	–	–						
	2a	5.52	4.91	6.13	5.70	5.19	6.22	2.79	2.09	3.48	2.32	1.67	2.97
	2b	6.33	6.12	6.54	6.29	5.99	6.60	2.75	2.30	3.20	3.04	2.60	3.48
	3-post[b]	5.90	5.59	6.22	6.36	6.13	6.60	2.12	1.78	2.47	2.62	2.22	3.01
	4	6.04	5.71	6.36	6.18	5.86	6.50	3.00	2.59	3.41	2.64	2.24	3.04
Distressing[c]	1a	–	–	–	–	–	–						
	1b	–	–	–	–	–	–						
	2a	5.20	4.45	5.95	5.04	4.38	5.69	2.82	2.11	3.54	2.54	1.84	3.23
	2b	6.09	5.81	6.36	6.11	5.77	6.45	2.78	2.33	3.23	3.00	2.53	3.47
	3-post[b]	5.52	5.17	5.88	6.26	5.99	6.52	2.30	1.92	2.69	2.80	2.36	3.24
	4	5.83	5.49	6.16	5.91	5.54	6.28	2.94	2.50	3.39	2.91	2.44	3.38
Disgusting[c]	1a	–	–	–	–	–	–						
	1b	–	–	–	–	–	–						
	2a	4.72	4.01	5.43	4.93	4.24	5.61	1.71	1.24	2.18	1.32	1.06	1.58
	2b	5.31	4.93	5.70	5.61	5.21	6.02	1.59	1.34	1.85	1.88	1.52	2.23
	3-post[b]	5.03	4.65	5.42	5.22	4.81	5.62	1.29	1.11	1.47	1.49	1.24	1.74
	4	5.16	4.75	5.57	5.56	5.16	5.95	1.77	1.47	2.08	1.36	1.17	1.55

Note. Dashes indicate that subjects in Experiment 1a did not make any ratings about the story they read, and that subjects in Experiment 1b did not make "unpleasant," "distressing," or "disgusting" ratings about the story they read.

[a]Ratings made pre-exposure to the film, but after "warning" subjects had seen the trigger warning about it. [b]Ratings post-exposure to the film (as for the other experiments). [c]Pre-exposure ratings are reported in-text, with the results of Experiment 3.

Appendix E

No student subjects in Experiments 1a and 2a chose to withdraw prior to the end of the experimental session.

Experiment 1b

Eighty-five MTurk subjects began but did not complete this experiment. Of those subjects, 21 (25%) quit at some point prior to being told they would next read a story, 25 (29%) quit while on the survey page displaying the story, and 39 (46%) quit at some point after having read the story.

Breaking it down by condition, 44 subjects dropped out of the "no warning" condition (14% before told story, 34% during story, and 52% after story); similarly, 41 dropped out of the "warning" condition (37% before told story, 24% during story, and 39% after story).

Notably, no "warning" subjects quit on the "trigger warning" survey page, that is, no one decided—specifically upon reading the warning—not to proceed with the experiment.

Experiment 2b

One-hundred-and-twenty-seven MTurk subjects began but did not complete this experiment. Of those subjects, 27 (22%) quit at some point prior to being told they would next watch a film clip, 12 (10%) quit while on the survey page telling them they would next watch a film clip, 9 (7% of total) subjects quit while on the "trigger warning" survey page, 11 (9%) quit while on the survey page displaying the film clip, and 68 (54%) quit at some point after having watched the film clip.

Breaking it down by condition, 61 subjects dropped out of the “no warning” condition (13% prior to film, 2% upon told film, [0% upon warning; wrong condition], 12% during film, and 74% after film); similarly, 66 subjects dropped out of the “warning” condition (29% prior to film, 17% upon told film, 14% upon warning, 6% during film, and 35% after film).

Very few “warning” subjects (9 of them) quit on the “trigger warning” survey page; a similar number dropped out on the page prior, upon being told they would next watch a video (without knowing what it would be about).

Experiment 3

One-hundred-and-seventeen MTurk subjects began but did not complete this experiment. Of those subjects, 25 (21%) quit at some point prior to being told they would next watch a film clip, 12 (10%) quit while on the survey page telling them they would next watch a film clip, 6 (5% of total) subjects quit while on the “trigger warning” survey page, 3 (3%) quit while on the survey page displaying the ratings regarding what they expected the film clip would be like, 10 (9%) quit while on the survey page displaying the film clip, and 61 (52%) quit at some point after having watched the film clip.

Breaking it down by condition, 57 subjects dropped out of the “no warning” condition (23% prior to film, 14% upon told film, [0% upon warning; wrong condition], 4% during pre-exposure ratings, 12% during film, and 47% after film); similarly, 60 subjects dropped out of the “warning” condition (20% prior to film, 7% upon told film, 10% upon warning, 2% during pre-exposure ratings, 5% during film, 57% after film).

Very few "warning" subjects (6 of them) quit on the "trigger warning" survey page, and a similar number dropped out on the page prior, upon being told they would next watch a film (without knowing what it would be about).

Experiment 4

One-hundred-and-fifty-two MTurk subjects began but did not complete this experiment. Of those subjects, 36 (24%) quit at some point prior to being told they would next watch a film clip, 19 (13%) quit while on the survey page telling them they would next watch a film clip, 1 (<1% of total) subjects quit while on the "trigger warning" survey page, 16 (11%) quit while on the survey page displaying the film clip, and 80 (53%) quit at some point after having watched the film clip.

Breaking it down by condition, 81 subjects dropped out of the "no warning" condition (28% prior to film, 12% upon told film, [0% upon warning; wrong condition], 7% during film, and 52% after film); similarly, 71 subjects dropped out of the "warning" condition (18% prior to film, 13% upon told film, 1% upon warning, 14% during film, and 54% after film).

Very few "warning" subjects (1 of them) quit on the "trigger warning" survey page; many more dropped out on the page prior, upon being told they would next watch a film clip (without knowing what it would be about).