Flinders
UNIVERSITY
ADELAIDE · SOUTH AUSTRALIA

Is forewarned forearmed? An investigation into the emotional and behavioral effects of trigger warnings

By

Victoria Mary Everest Bridgland

Thesis
Submitted to Flinders University
for the degree of

Doctor of Philosophy

College of Education, Psychology, and Social Work
25th October 2021

Table of Contents

Summary

Despite widespread use, there is a dearth of research specifically investigating the effects trigger warnings have on people's emotional and behavioral reactions. My thesis aimed to bridge gaps left by the first wave of trigger warning research to help determine *how* and *why* these warnings may or may not change emotion and behavior.

My thesis makes a substantial contribution to our knowledge about the emotional effects of trigger warnings. First, I found that upon viewing a trigger warning, people experience an anxious anticipatory period that does not seem to reflect mental preparation to cope with negative content (Chapters 3, 4, 5, and 6). Second, using novel stimuli that are ambiguous and neutral, rather than explicitly negative, my work replicates previous evidence that trigger warnings have little subsequent effect on immediate emotional reactions towards material (Chapter 3). Indeed, overall, my thesis and the work of others unanimously suggests that trigger warnings *do not* mitigate distressing reactions. Rather, it is more likely that trigger warnings lead to harm and my thesis suggests three possibilities for when and how this harm is likely to occur. First, it is possible that trigger warnings have the potential to exacerbate distressing reactions when expectations match with experiences (Chapter 3). Second, the negative effects of trigger warnings may only emerge over time (Chapter 4). Third and finally, the negative effects of trigger warnings may not occur for immediate emotional reactions, but rather for other kinds of appraisals more closely linked with negative memories (Chapter 4).

Second, my thesis significantly contributes to our knowledge about the behavioral effects of trigger warning messages. Thus far, research has focused almost exclusively on how trigger warnings may or may not change emotional reactions towards material. Chapters 6 and 7 help to confirm that, despite both critics' and advocates' claims, trigger warnings do little to foster the avoidance of potentially upsetting material. In fact, Chapter 6 shows that

people seem eager to view distressing material marked with a trigger warning (vs. a neutral message), and Chapter 7 provides evidence that vulnerable people—such as those with mental health concerns—may be the least likely to be deterred by a warning message. My work therefore provides preliminary support that trigger warnings may actually foster a "Forbidden Fruit effect" (Ringold, 2002)—where a restricted behavior becomes more desirable—and encourage morbid curiosity about distressing content (Oosterwijk, 2017). Moreover, Chapter 7 is the first empirical investigation of trigger warnings in the applied context of social media, finding that warnings do not seem effective in preventing people from consuming negative content online. These findings have critical implications for current policies that effect over 1 billion people worldwide via Instagram.

Taken together, my thesis adds to a growing body of literature showing that trigger warnings seem to be ineffective in achieving their purported goals. Further work should focus on how to develop alert systems or strategies that do achieve these commendable aims.

Declaration

I certify that this thesis:

1. does not incorporate without acknowledgment any material previously submitted or degree or diploma in any university; and
2. to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text.

Signed: Victoria M. E. Bridgland

Acknowledgement of country

I would like to acknowledge that this body of work was produced on the lands of the Kaurna nation. I recognize the Traditional Custodians of the land where my research was conducted and pay my respects to their Elders past, present, and emerging.

Acknowledgments

First and foremost, I would like to thank my supervisor, Melanie Takarangi. You were not just a mentor—that goes without saying. You were also a captain taking me to unchartered waters on intrepid adventures into the unknown. You were a sports coach cheering me on to push myself to new limits. You were a lighthouse keeper helping me make it safely to harbour when things got rough. You were a high priestess helping me work magic and discover new and exciting things. You were Athena (goddess of wisdom) who always knew what to do or how to figure things out. You were a warrior fighting for me and crushing obstacles. And lastly and most importantly, you were and are a true friend who treated me as an equal. Words are not enough to express how thankful I am to have had you as my guide along this journey. I would not be the same person I am today without you. I'm excited for many more research and non-research related adventures together in the future.

My next round of thanks goes to the never-ending and enduring support of the Takarangi Lab. Although sometimes frustrating, it's no mistake that we are regularly bestowed with names relating to powerful groups of women (Charlies Angels being my particular favourite). We are a force to be reckoned with. To my 'big sisters' Deanne, Ella, Sasha, Di, and Jacinta—thank you for taking me under your wings. I was a mere fledgling when we first met and you taught me how to fly. To my 'little sisters' Taylor, Nadine H, Lucy, Nadine S, Erin, Catherine, and Jorja—thank you for allowing me to be a part of your journeys and 'pass the torch' along.

I'd like to thank my overseas friends and collaborators who I've been lucky enough to meet and work with along the way. Thanks to Rich and the McNally lab who have welcomed me with open arms from across the world. Thanks especially to Ben my fellow trigger warning companion and collaborator.

Lastly, I’d like to thank my family and friends outside of academia. To my sister Trudy who is an endless source of comfort and support—thank you for around the clock chats, graveyard dog walks, and your unconditional and unfaltering love. You are my lifeline and guide through all facets of life and I would not survive without you. Thank you to Phoebe and Chloe my other ‘sisters’ who are always ready to laugh and cry with me. Lastly, thank you to mum and dad who have always provided a safe harbour for me to come home to.

1 Literature review

The sadistic torture of Barbara Gordon by The Joker in DC's *The Killing Joke* (Finkel, 2017). The live suicide of veteran Ronnie McNutt on TikTok (Bedo, 2020). The abusive relationship between Daisy and Tom Buchanan in *The Great Gatsby* (Medina, 2014). Footage of the murder of George Floyd by Minneapolis Police Officer Derek Chauvin (Simonpillai, 2021). These are just a few examples of the fictional and non-fictional horrors that often lurk behind trigger warnings. Also commonly called *content warnings, trigger warnings* are alerts that upcoming material may contain potentially distressing themes, content, or images. Trigger warnings originated in the early days of the internet on feminist message boards, before becoming widespread with the advent of social media in the late 2000s, and migrating to college campus culture in the early 2010s (Vingiano, 2014). Advocates argue that trigger warnings help people by emotionally preparing them to view or completely avoid content they may not want to see (e.g., Gust, 2016). However, critics of trigger warnings argue that instead of preparation warnings may exacerbate negative reactions, and that encouraging avoidance might be harmful rather than beneficial. Despite the widespread use of trigger warnings, there is a dearth of research specifically investigating the effects such warnings have on people's emotional and behavioral reactions. In an effort to understand if trigger warnings are helpful or harmful, a handful of studies emerged during 2018 to 2020; these studies focused on examining what happens when someone sees a trigger warning and then views distressing material (Bellet, Jones, & McNally, 2018; Boysen, Issacs, Tretter, & Markowski, 2021; Gainsburg & Earl, 2018; Bellet et al., 2020; Jones, Bellet, & McNally, 2020; Sanson, Strange, & Garry, 2019). The take home from this first wave of research is that trigger warnings seem *trivially* impactful—they do not exacerbate distressing reactions, but do not help alleviate them either. My thesis aimed to bridge gaps left by this first wave of trigger warning research to help determine *how* and *why* these warnings may or may not change emotion and behavior.

1.1 Trigger warnings: A history

In 1957 the notion of *informed consent*, or the practice of warning someone about the potential risks inherent in a medical procedure, entered public and legal consciousness (in the case of *Salgo vs. Leland Stanford, Jr University Board of Trustees;* Fries & Loftus, 1979) and in 1968 the modern movie rating system was born (e.g., MA 15+: Strong themes, Strong Sex Scenes, Sexual Violence; The Classification & Rating Administration, 2021). Therefore, warnings about potential risks, or potentially encountering distressing content, have existed in different forms for many decades now. Yet in the 1990s a new type of warning rose from relative internet obscurity into the mainstream. Feminist community forum users—such as those on the *Ms. Magazine* website—started adding the phrase "*trigger warning"* on posts discussing potentially distressing topics (e.g., sexual assault and eating disorders; Vingiano, 2014). The 2000s saw the spread of trigger warnings to a fanfiction site called LiveJournal; possibly the earliest mention of the term appears in a post from 2003 by user "Morbid-thoughts" titled "*What Type of Self-Mutilation Are You? (Warning: Triggering Pictures)*," and from there trigger warnings migrated to Twitter (2006), Tumblr (2007), and Facebook (2008). These early forms of trigger warnings were simple: they were lines of text added at the top of a post by the user who created the content.

But since their initial debut, trigger warnings have evolved significantly in form, function, and domain. Individual users have continued to include text-based trigger warnings at the top of their posts, but social media platforms have instituted more official warning systems—primarily to guard against graphic *visual* content. For example, in 2017 Instagram introduced *sensitive content screens* (Instagram, 2017). Photos and videos that do not violate community guidelines (e.g., unlike posts explicitly supporting terrorism)—but are nonetheless deemed “sensitive” by moderators—are presented blurred, with a warning (“Sensitive content: This photo/video may contain violent or graphic content”), and with an option to unveil the content (‘see photo’). Facebook, Twitter, Reddit, YouTube and Buzzfeed have added similar warning systems. Even the BBC has experimented with sensitive content screens on its homepage and has run small scale experiments to investigate the effectiveness of allowing readers to blur sensitive news stories (Miller & Grandjean, 2019). Moreover, while the streaming service Netflix occasionally issues text-based trigger warnings in addition to traditional film/television advisory warnings, a recent petition argues for mandated trigger warnings on all potentially distressing content (Medhora, 2021). In summary then, trigger warnings originated and have continued to evolve and flourish online. Interestingly however, they have not only remained in this domain.

The early 2010s saw trigger warnings become inextricably linked to college campuses. In 2013, Oberlin University was thrust into the spotlight when it issued an official document to staff advising them to “understand triggers, avoid unnecessary triggers, and provide trigger warnings”—a proposal ultimately abandoned after several dozen professors expressed concern that this initiative would limit academic freedom (Flaherty, 2014; Flood, 2014). Then in 2015 a group of four undergraduate students from Columbia University wrote an op-ed calling for the use of warnings for any “triggering and offensive material” that may be included in their courses (Johnson, Lynch, Monroe, & Wang, 2015). Around the same time, similar requests were made at the University of California Santa Barbara, Rutgers University, the University of Michigan, George Washington University and likely many others (Flood, 2014). Since these early calls for trigger warnings in university/college settings, as many as 51% of college professors issue warnings on their class content (NCAC, 2015). But the introduction of trigger warnings into university/college culture was not without intense controversy. In 2016, the University of Chicago released a now infamous welcome letter to incoming students, stating: “we do not support so-called trigger warnings.” But it was an article from 2015 by Lukianoff and Haidt—*The Coddling of the American Mind*—that crystalized trigger warnings as a mainstay of the 2010 culture wars. Lukianoff and Haidt argued that trigger warnings, along with “safe places” and “microaggressions,” were part of a cultural shift towards “vindictive protectiveness.” That is, trigger warnings suggest that students should be shielded from difficult or distressing topics and as a result, college students were being increasingly infantilized and mollycoddled. After the publication of this article hundreds of think-pieces and essays emerged online either criticizing or defending the use of trigger warnings (e.g., Flaherty, 2019; Gust, 2016).

One reason why trigger warnings became so controversial is because of changes to the content to which they were applied. The term “trigger warning” originates from PTSD research showing that stimuli (e.g., flashing lights) with characteristics similar to a traumatic event (e.g., a car crash involving bright headlights) can “trigger” a person to re-experience that trauma (Ehlers, Hackmann, & Michael, 2004). Re-experiencing symptoms include vivid thoughts, feelings and flashbacks about the event (Ehlers et al., 2004). Trigger warnings were first conceptualized as protecting victims of trauma or people suffering from PTSD on the feminist forums where they made their debut; for example, a victim of sexual violence could mitigate being “triggered” (e.g., experiencing a panic attack or PTSD symptoms) if they were warned prior to reading a post about sexual assault (e.g., by avoiding the post completely or “preparing” themselves to read it; Haslam, 2017). These ideas about the original purpose of trigger warnings persist in some informational materials disseminated today. For instance, The Innocent Lives Foundation (2020)—a source cited by social media influencers who use trigger warnings—claims that “memories for trauma are worse without warning” and that “trigger warnings are simple ways to help survivors avoid reliving the event.” Aside from traumatic experiences (e.g., domestic violence), early trigger warnings also warned about mental health disorders such as depression, suicidal ideation and self-harm (Vingiano, 2014).

Despite having roots in trauma and mental illness, the topics covered by trigger warnings have expanded far and wide. For instance, the Oberlin college students recommended that warnings should be added to any material that “marginalizes student identities in the classroom” and/or depicts “histories and narratives of exclusion and oppression” (Johnson, Lynch, Monroe, & Wang, 2015). Indeed, trigger warnings are now widely used for any potentially provocative or sensitive material that students may encounter (e.g., racism, pregnancy, needles etc.), including issues of injustice, discrimination, and oppression (Palmer, 2017; Walker, 2021). Mirroring this shift in content covered, even the very term “trigger warning” has been eschewed in favor of “content warnings”/ “content notes” to acknowledge the idea that although people may not be “triggered” in the clinical sense of the term, they may still be distressed by the content (Vingiano, 2014). Moreover, some have argued that the term “trigger warning” should be abandoned altogether, in favor of the benevolent term “content forecast,” because the word “trigger” alludes to violent weaponry and a “warning” creates a sense of threat (Doney, 2019; Stringer, 2016).

Taken together, since their introduction, trigger warnings have changed in form: from simple lines of text to official blurring screens; function: from warning about trauma/mental illness related content to warning about anything that may be potentially distressing; and domain: from online feminist message boards to college campuses. With this evolution, trigger warnings have attracted controversy and criticism. However, perhaps the biggest controversy surrounds what trigger warnings reportedly do (or do not do).

1.2 Are trigger warnings helpful or harmful? The debate versus the evidence

Two camps have emerged from the controversy surrounding trigger warnings: people who believe that trigger warnings are helpful (e.g., Gust, 2016) and people who believe that trigger warnings may actually cause harm (e.g., Lukainoff & Haidt, 2015). Here I will outline the central opposing themes of the debate—noting that this is by no means an exhaustive account of the claims made about trigger warnings—and discuss relevant empirical work and psychological theory. The first two themes focus on secondary cultural effects that may arise due to the widespread use of trigger warnings: first, how trigger warnings may shape people's perceptions of vulnerable populations, and second, how trigger warnings may shape academic environments. The second two themes focus primarily on what might happen when someone actually encounters a trigger warning directly—first, what might happen if that person stays and views the upcoming content, and second, what happens should they decide to avoid the content. While it is necessary to discuss the first two themes for completeness, my thesis is focused primarily on the latter two themes and therefore, these themes comprise the majority of my literature review.

The cultural impact of trigger warnings. Debate swirls in both public and academic domains about the general cultural impact of trigger warnings. First, there are opposing arguments about the way trigger warnings may shape cultural perceptions about vulnerable populations (e.g., people with mental illness, PTSD, or who have been victims of trauma). Advocates argue that trigger warnings communicate a culture of support for vulnerable people (Boysen, Wells, & Dawson, 2016; George & Hovey, 2020). Indeed, Doney (2019) describes trigger warnings "not as outcome-based boxes to tick, but as components of a greater whole"—that is, when someone chooses to use a trigger warning, they are also conveying a greater message about recognizing the individual journeys of a person and their lived experiences (i.e., which could include trauma). Critics, however, suggest that trigger warnings are part of a cultural shift that promotes sensitivity, fragility and overidentification with victimhood (Boysen et al., 2016; Lukainoff & Haidt, 2015; Robbins, 2016; Vigo, 2018). Other critics are concerned that trigger warnings may promote the stereotype that victims of trauma are not resilient, an idea that may hinder psychological recovery in trauma victims (Boysen et al., 2016; McNally, 2014).

Only three studies have examined the impact of trigger warnings on cultural perceptions towards trauma survivors, with mixed results. Using a non-trauma exposed Mechanical Turk sample, Bellet et al. (2018) found that participants who viewed a trigger warning (vs. no warning) had a stronger belief that both themselves and others would be more at risk of long-term emotional harm (e.g., PTSD) following exposure to a traumatic event. However, this pattern did not replicate in a follow-up study with college students (Bellet et al., 2019) or in a trauma-exposed sample (Jones, et al., 2020). It is possible that the results from Bellet et al. (2018) did not replicate because the original results were a false positive or because the findings were unique to the older trauma naïve MTurk sample used (i.e., rather than younger college students). Indeed, Copoc (2021) found that viewing a trigger warning did not increase college students perceived societal stigma or personal stigma towards people with mental illnesses (e.g., belief that people with a mental illness cannot take care of themselves) relative to participants who viewed a control message. Therefore, at this early stage of research it appears that trigger warnings have little effect on cultural perceptions about trauma survivors and mental illness.

Second, specific debates surround the use of trigger warnings within academic and educational contexts. Advocates state that warnings in the classroom help to foster a safe environment for trauma survivors, allowing them to prepare for distressing material and therefore enhancing their learning outcomes (DeBonis, 2019; George & Hovey, 2020). In one study, students interviewed about the use of warnings within the classroom believed warnings increased academic freedom and helped to frame emotive issues in an academic context (Bentley, 2017). In contrast, critics argue that trigger warnings have a *chilling* effect on academic freedom that places warnings within a greater cultural censorship movement

(Essig, 2014; Kamenetz, 2016; Klugman, 2017; Lukainoff & Haidt, 2015; McNally, 2014; NCAC, 2015). In particular, trigger warnings have been criticized for inhibiting the academic freedom of staff who are worried about harming students via potentially distressing material (Essig, 2014; NCAC, 2015). Furthermore, some critics posit that trigger warnings may reduce students' critical thinking ability, by providing them with the option to censor themselves from learning about certain uncomfortable topics (NCAC, 2015; Klugman, 2017). In line with this idea, some students have also suggested that warnings increased their awareness of triggering material and made them more apprehensive and anxious when attending class, inhibiting learning (Bentley, 2017).

Only two studies have specifically investigated the way that warnings might enhance or reduce learning outcomes. Across three experiments, Boysen et al. (2021) found that trigger warnings did not improve performance on a multiple-choice test for factual lecture content—including for participants with prior personal experience with the distressing topics discussed (e.g., sexual assault). In contrast, Bruce and Roberts (2020) found that participants who were victims of sexual and physical violence had poorer reading comprehension of articles about sexual assault labelled with trigger warnings (vs. unlabeled articles, and vs. participants without a history of violence). Therefore, the current evidence suggests that at best, warnings have no effect on educational outcomes, while at worst trigger warnings may impair educational outcomes in vulnerable populations (e.g., people with a trauma history).

Taken together, warning advocates claim that warnings represent a positive cultural shift benefiting vulnerable populations, while critics argue the opposite may be true. Scant empirical evidence exists for these claims and the available evidence provides conflicting data suggesting that warnings at best may have no positive effect and at worst may lead to increased perceptions of vulnerability and impair learning outcomes. However, here, I am more interested in the claims made about what actually happens when someone comes across a trigger warning message, rather than their greater cultural impact. That is, what might happen if someone sees a warning and then stays to view the warned of content? And what happens if someone decides to avoid the warned of content? These questions relate to the primary reasons why trigger warnings were created and why they are widely implemented (i.e., to help people when encountering distressing content). Therefore, answering these questions will help to unravel many of the key debates surrounding the use of trigger warnings.

Trigger warnings and emotional reactions. If someone sees a trigger warning, and decides to stay and view the warned of material, how might trigger warnings change emotional reactions towards that material?

Coping strategies and emotion regulation. Advocates claim that warnings help people "mentally prepare" to cope with distressing material, and thus mitigate the negative reactions that could otherwise occur if people were to encounter such material unaware (Bentley, 2017; Cares, Franklin, Fisher, & Bostaph, 2017; DeBonis, 2019; George & Hovey, 2020). To "prepare" is defined as "mak[ing] (someone) ready or able to do or deal with something" (Oxford Languages, 2021). To prepare in a trigger warning context could therefore be via *coping strategies*—a conscious effort to manage the demands of a stressful situation (e.g., receiving a cancer diagnosis), using thoughts and behaviors (e.g., seeking social support; Folkman & Moskowitz, 2004). Work on approach-based coping strategies (i.e., coping strategies focused on a stressor itself and a person's reaction to it; Littleton, Horsley, John, & Nelson, 2007), such as emotion regulation, offers insight into how this process *could* work (see Gross, 2002, for review). For instance, the emotional regulation technique known as cognitive reappraisal is similar to the intended purpose of a trigger warning. Cognitive reappraisal works by asking people to alter their appraisal of a situation to change the emotional impact it causes. Several studies have demonstrated that instructing participants to use reappraisal instructions prior to being exposed to distressing material can reduce negative emotional reactions. For example, participants given acceptance (experience emotions without judgment) or reappraisal (reframe thoughts in unemotional or positive terms) instructions experience significantly lower levels of negative emotions when watching negative films compared to when they are given no instructions (Shiota & Levenson, 2012; Troy, Shallcross, Brunner, Friedman, & Jones, 2018; Wolgast, Lundh, & Viborg, 2011). These data suggest that it is possible for people to emotionally prepare themselves against negative reactions if given prior instructions about how to interpret upcoming negative content.

However, although trigger warnings and cognitive reappraisal strategies share similar goals, it seems unlikely that trigger warnings have the same emotional benefits as reappraisal strategies. Reappraisal and acceptance strategies ask participants to examine content unemotionally and non-judgmentally. Trigger warnings, however, usually describe potential negative emotional reactions that will likely occur when viewing the material (e.g., distress, anxiety etc.) and do not describe *how* to view content to lessen its impact. Thus, although advocates argue that trigger warnings should increase emotional preparedness and reduce subsequent negative emotional reactions, it seems unlikely that trigger warnings and cognitive reappraisal exert the same effects.

Bracing for the worst. Aside from actively “emotionally preparing” someone to face distressing content, advocates also argue that warnings mitigate unwanted “surprises” (Cares, Franklin, Fisher, & Bostaph, 2018; Mosseri, 2019a). Further, many health websites that have information on PTSD/trauma triggers advocate for the use of trigger warnings under the proposition that “triggers are more distressing if they come as a surprise” (Cuncic, 2020; Sullivan, 2019; Good Therapy, 2018), or similarly, that “vivid memories of trauma are more distressing if they happen without any warning” (The Innocent Lives Foundation, 2020). The origins for these ideas are dubious; these sites either provide no empirical basis for their claims or refer ambiguously to the American Psychological Association (APA) as a source. It is possible that these ideas originate from the commonly held notion that negative outcomes are worse if they are unexpected than expected. This notion has some basis in empirical evidence: negative outcomes are generally reported as more aversive if they are perceived as unexpected (i.e., come as a shock/surprise) than expected (Dugdale, Eklund, & Gordon, 2002; Siemer, Mauss, & Gross, 2007). Given that the unexpected is aversive, does this evidence suggest that warnings may reduce negative reactions relative to coming across content unaware?

An area of literature that may help determine if warnings mitigate distress by setting up negative outcomes as expected rather than unexpected, is research on *bracing for the worst*. When anticipating a negative outcome, people often demonstrate a marked decline in optimism and adopt a pessimistic outlook (Shepperd, Oulette, & Fernandez, 1996). For instance, participants estimate lower test performance scores (Gilovich, Kerr, & Medvec, 1993; McKenna & Myers, 1997; Savitsky, Medvec, Charlton, & Gilovich, 1998; Sweeny, Shepperd, & Carroll, 2009), become less optimistic when anticipating medical test results (Taylor & Shepperd, 1998), and become more pessimistic when facing a possible financial strain (Shepperd, Findley-Klein, Kwavnick, Walker, & Perez, 2000) as the moment of resolution draws near. There are several possible reasons for this decline in optimism, such as the possibility that people recalibrate their mood when they acquire new information that suggests a negative outcome is likely (see Carroll, Sweeny, & Shepperd, 2006). However, when a decline in optimism reflects an attempt to ready oneself and avoid disappointment and prepare for the worst outcome, this is known as *bracing* (Shepperd et al., 2000). Bracing is theorized to be a type of preparedness—an adaptive state of readiness to respond to uncertainty (Sweeny & Shepperd, 2007). Similar to the goals of a trigger warning, bracing is a way of managing and taking control of one’s emotional state with the goal of reducing the psychological impact should the aversive event come true. Rather than managing anticipation in a positive manner however, according to the bracing hypothesis, it is always better to *expect the worst*. That is, bracing for the worst is said to be beneficial because negative outcomes are less aversive if they are expected (vs. unexpected) and positive outcomes are more positive when they are unexpected (vs. expected). Trigger warnings may therefore promote a bracing technique because they encourage people to *prepare for the worst* when anticipating how they may react to upcoming content (e.g., the content is distressing, and I will have a distressing reaction).

However, although it is a commonly held belief that bracing for the worst will help to ameliorate negative reactions in the face of a negative outcome (relative to holding a positive expectation) mixed evidence exists for this claim. For instance, breast cancer survivors asked to retroactively think about how their current quality of life differed to what they had expected when they were first diagnosed, reported higher levels of current negative affect when they believed their life was worse than expected (Bettencourt & Manning, 2016). In other words, remembering expecting the worst in the *past* can enhance positive affect in the *present* if someone believes their present situation is better than they originally estimated (See also; Wilson & Ross, 2001). Similarly, immigrants expressed higher current life satisfaction when their actual experiences in their new home country exceeded expectations they reported before they immigrated (e.g., discrimination; Mähönen, Leinonen, & Jasinskaja-Lahti, 2013). Yet, a recent longitudinal study on the experience of daily stressors (such as arguments with significant others and work problems)—which monitored participants several times a day over 7 days—found that negative affect was *not* lower for anticipated than unanticipated stressors when they occurred (Neubauer et al., 2018).

Evidence from more experimental paradigms also provides mixed evidence for the bracing hypothesis. Participants given a series of probabilities of winning a lottery (Mellers, Schwartz, Ho, Katty, & Ritov, 1997; van Dijk & van der Pligt, 1997), or who rated higher probabilities for making a basketball shot (McGraw, Mellers, & Ritov, 2004), reported more intense emotional responses for unexpected outcomes. Moreover, participants falsely identified as being "low risk" versus "high risk" of disease susceptibility (Shepperd & McNulty, 2002), and of dangerous toxin exposure (Sweeny & Dillard, 2013) reported higher negative affect when a result came back as testing positive (e.g., when it was a surprise rather than expected for participants told they were at a "low risk"). Therefore, unexpected outcomes are often more intense than expected ones, meaning that when the unexpected outcome is negative bracing for the worst *can* lessen the emotional blow.

However, other research suggests that the benefits of expecting the worst may only occur under limited circumstances. Marshall and Brown (2006) found that although participants who received unexpected (vs. expected) feedback of failure on a cognitive test were more surprised, these participants did not report higher negative affect in the face of that unexpected surprise. In fact, participants who expected success *generally* (i.e., were optimistic), expressed more positive emotion in the face of success *and* failure—while the opposite was true for participants who generally expected failure. Marshall and Brown suggested that optimistic participants may have viewed both successes and failures in a more positive light—a conclusion supported by a second experiment where optimistic participants were more likely to attribute successes and less likely to attribute failures to personal ability.

As well as optimism, other factors appear to influence expectancy related emotional outcomes. For instance, participants who have been primed with a prevention (safety concerns) or morality focus, or who are under high cognitive load, feel more positive about expected rather than unexpected outcomes regardless of the positive or negative implications of the outcomes (Noordewier & Stapel, 2009). Furthermore, emotions appear to be amplified or attenuated by expectancies only when an expectation is active in memory *very* shortly after the event before the expectancy fades from consciousness (e.g., Shepperd & McNulty, 2002; Sweeny & Dillard, 2013; Golub, Gilbert, & Wilson, 2009). For instance, in Golub et al. (2009), there was no difference in negative affect for participants who received expected versus unexpected feedback when rating emotional responses just a short time (two minutes) after receiving feedback or after a longer delay (after 24-hours). However, in a replication of this experiment, Sweeny and Shepperd (2010) found that participants who had expected more positive results than what they received, felt worse when affect was measured *immediately* after feedback. These results help to explain previous findings (e.g., Shepperd & McNulty, 2002; Sweeny & Dillard, 2013) because participants rate their emotional responses immediately after receiving feedback in typical bracing paradigms. In sum, the negative affect experienced post feedback from over optimistic estimates appears to dissipate rather quickly.

Thus, if the “benefits of pessimism’” are so fleeting, are the costs of negative expectancies worth this payoff? Indeed, while Sweeny and Dillard (2013) found that participants told they were “high risk” for toxin exposure felt less negative after receiving expected negative test results (vs. those falsely told they were “low risk”), these participants also indicated they would be less likely to take action to prevent the threat in the future. Therefore, the positive affect experienced in the aftermath of disconfirmed negative expectancies may a) not be very beneficial in the long term, and b) may actually prevent adaptive behaviors.

Evidence from other research areas suggests that the costs of bracing for and therefore anticipating a negative event can be akin to experiencing it. For instance, participants expecting to perform a cold pressor task (submerging one’s hand in ice water) report decreased frustration tolerance and increased blood pressure (Spacapan & Cohen, 1983). Moreover, Neubauer et al. (2018) found that anticipating a stressor in the next few hours was associated with prolonged elevated negative affect that lasted for two to three hours after they first reported expecting the stressor prior to experiencing it.

Taken together, the fleeting and limited benefits of expecting the worst may not outweigh the costs and "may be a sucker's bet" (Golub et al., 2009). Applying the bracing literature to trigger warnings, it seems possible that expecting to have a negative reaction towards content may not translate into any emotional benefits when someone actually goes on to view the content, despite the emotional costs.

Expectancy effects. While the literature on bracing demonstrates that negative expectancies appear costly in the lead up to a negative experience, literature on *expectancy effects* demonstrates that negative expectancies can also worsen the experience itself (Kirsch, 1985). Indeed, critics argue that trigger warnings have the potential to *increase* rather than decrease negative emotional reactions towards content. For instance, warnings may do so by instilling fears (Lesh, 2016), forcing an interpretation (Waldman, 2016), or skewing perceptions (Filipovic, 2014), that material will cause harm, all of which may not have existed in the absence of a warning. Other critics have argued that labelling content with a trigger warning may reinforce other harmful reactions, such as the perception that trauma is central to one's identity (McNally, 2014)—a concept linked with increased PTSD severity (Berntsen & Rubin, 2006). These criticisms likely have merit—we know that setting up an expectation of negative physical health symptoms such as pain, itch, and other side effects can cause or exacerbate those very outcomes; known as the nocebo effect (Bartels, van Laarhoven, van de Kerkhof, & Evers, 2016; Benedetti, Lanotte, Lopiano, & Colloca, 2007; Myers, Cairns & Singer, 1987). A nocebo effect is a type of *response expectancy* (Kirsch, 1985)—people anticipate automatic responses and behaviors to environmental cues, leading them to internally generate those anticipated responses, which alters their subjective experience and physiological function. In one example, participants given information about gastrointestinal side effects were six times more likely to withdraw from an angina treatment due to this complaint (vs. participants not told about this side effect; Myers et al., 1987). Applying this idea to a trigger warning context, it is possible that when people view a warning, they begin to anticipate and expect the negative reactions they may have when they actually view the material (e.g., distress, anxiety, panic etc.), and subsequently manifest these reactions when they actually see the content.

A scant number of studies that have directly examined the effects of film rating warnings on reactions to graphic media generally show support for the idea that warnings may worsen reactions. For instance, participants told that a film they were viewing was rated R and contained possibly violent content were significantly more scared prior to viewing and more distressed while viewing the film compared to participants told that the film was rated PG with graphic content cut (de Wied, Hoffman, & Roskos-Ewoldsen, 1997). Relatedly, participants given explicit or vague warnings reported being more upset and frightened while watching negative films than participants who were not warned (Cantor, Ziemke, & Sparks, 1984). Similarly, participants given detailed consent information highlighting possible negative effects reported higher negative evaluations of sexually explicit photographs compared to procedural only information (Senn & Desmarais, 2006). Thus, past research on forewarning about graphic media demonstrates that warnings could lead to nocebo effects; creating negative expectancy and exacerbating negative reactions towards material.

Taken together, advocates claim that trigger warnings are used to “emotionally prepare” people to view distressing material, reducing negative affect, while critics claim that warnings may actually increase distressing reactions. Although research on cognitive reappraisal shows that it is possible to give people instructions that help them to mitigate the effect of negative material, it seems more likely that warnings may exacerbate negative reactions.

Trigger warnings and emotional reactions: The evidence so far. Thus far, the largest body of trigger warning work has focused on how trigger warnings change emotional reactions in the lead up to, and subsequent viewing, of negative material. Two previous studies have demonstrated that trigger warnings appear to increase negative expectancies and create a noxious anticipatory period characterized by anxiety and negative affect prior to viewing content. Gainsburg and Earl (2018) found that participants reported anxiousness and nervousness more than any other emotion when asked to imagine encountering content with a trigger warning message; participants also reported higher levels of anticipated negative affect for video and essay titles accompanied by a trigger warning (vs. no warning). Similarly, Sanson et al. (2019) found that participants who saw a trigger warning prior to watching a distressing film believed the film would be more negative than participants who saw no warning. These findings therefore *might* lend support both to a nocebo or to a bracing account; that is, perhaps trigger warnings make people feel worse in the lead up to consuming negative content, and then either exacerbate (nocebo) or alleviate (bracing) negative reactions when people actually come to view that content.

However, the handful of studies that have investigated these prospects have found mixed results. Sanson et al. (2019) found that trigger warnings (vs. no warning) had trivial effects on levels of negative affect, intrusions, and avoidance symptoms following exposure to negative text passages and film clips. Similarly, Boysen et al. (2021) found warnings had little effect on emotional reactions to negative lecture content—including among people with personal experiences that matched the topics (e.g., sexual assault).

Other studies point to the possibility that trigger warnings may only lead to negative emotional outcomes for certain groups of people. For example, Bellet et al. (2018) found that

MTurk participants who viewed a trigger warning (vs. no warning) reported higher anxiety when reading distressing text passages—but only when those participants held the belief that words can cause harm. Bellet et al. (2019) failed to replicate this pattern of results in a college student population but did find that warnings caused an increase in anxiety when participants read distressing text passages. Jones et al. (2020) also failed to replicate the Bellet et al. (2018) finding, but participants in their sample with *higher PTSD symptoms* had increased anxiety when viewing content accompanied by a trigger warning. Moreover, Jones et al. (2020) found that participants with a history of trauma reported that their traumatic event was central to their identity when they were exposed to trigger warnings (vs. no warnings). Finally, Gainsburg and Earl (2018) found that trigger warnings slightly *reduced* negative reactions towards distressing essay content. However, this was only true for participants who believed trigger warnings were *coddling* in nature; participants who viewed trigger warnings as *protective* against harm actually reported more negative affect toward content with a warning (vs. content without a warning). These results suggest that certain populations—such as people with high PTSD symptoms or people who believe that trigger warnings are protective against harm—seem especially susceptible to the negative effects of warnings.

Taken together, early research suggests that at worse, trigger warnings cause anticipatory anxiety in the lead up to distressing content and, in some cases, increase negative emotional reactions towards material. At best, warnings appear to have little effect on reactions towards material. Thus, even when warnings appear not to be doing any harm, they also do not appear to do any good. These results therefore do not support the idea that trigger warnings operate via a bracing or cognitive reappraisal framework: if they did, then the evidence should show that warnings alleviate negative reactions. The available studies also suggest mixed support for a nocebo account: some studies suggest trigger warnings may exacerbate negative reactions while others suggest they have little to no effect. Although the majority of existing trigger warning research has focused on how trigger warnings might affect emotional reactions, many questions remain.

First, what are the most likely cognitive and emotional reactions when someone sees a trigger warning? Answering this question is key to understanding why emerging research suggests trigger warnings fail to reduce negative reactions. One avenue is to take a closer look at the ways that people claim that trigger warnings work. Specifically, in studies asking participants about their opinions about trigger warnings, responses commonly reflect a belief that warnings help people to "prepare" for distressing material (Bentley, 2017; Cares et al. 2017; DeBonis, 2019; George & Hovey, 2019; NACA, 2015). As discussed above, preparation could be understood as referring to the use of coping strategies such as cognitive reappraisal—i.e., reappraising the way a situation is construed to decrease emotional impact. However, no studies to date have investigated the ways that trigger warnings may or may not change people's use of coping strategies.

Second, given that trigger warning use is widespread and covers a diverse range of topics, it is possible that trigger warnings not only exacerbate people's reactions to overtly negative material, but also their reactions to neutral or ambiguous material. For instance, consider a television program that warns of a sexual assault scene—it is possible a viewer may interpret other scenes of a sexual nature as negative because they are expecting an assault to occur. This idea is not without empirical support: research on priming effects shows it is also possible warnings may cause people to interpret *neutral* material in a negative way. According to spreading activation theory, when a concept has been primed, associated concepts and knowledge in memory become more accessible (Collins & Loftus, 1975). For instance, exposure to the word "death" leads participants to respond quicker to semantically related words like "distress" (semantic priming; Janiszewski & Wyer, 2014), and reading negative news articles increases memory for negative information in subsequent news articles (affective priming; Baumgartner & Wirth, 2012). Affective priming can also change how subsequent neutral or ambiguous information is interpreted. For instance, participants primed with negative adjectives (e.g., mean), versus positive adjectives (e.g., sincere), rated an unknown person in a photograph higher on a number of negative traits (Ferguson, Bargh, & Nayak, 2005), and participants primed with ethics-related words (vs. neutral) were more likely to categorize morally ambiguous behavior as unethical (Welsh & Ordonez, 2014). Therefore, it seems possible that warnings about distressing content may not only exacerbate negative reactions towards distressing material but could also prime participants to interpret a negative meaning from *neutral* material.

Third, existing research has focused on a relatively narrow definition of a trigger warning—despite the fact that the term “trigger warning” has evolved and changed over time. Specifically, the typical *popular* definition of a trigger warning is quite vague: an alert that upcoming material may be distressing. Prior work has focused on this definition, examining people’s general emotional reactions when they encounter various types of *novel* stimuli, such as negative films (Sanson et al. 2019) and text passages (Bellet et al. 2018). However, no research has investigated trigger warnings as they were *originally* defined—which was as a warning that people might encounter material that could “trigger” them to re-experience a traumatic event. Trigger warnings, therefore, were originally intended to mitigate the “triggering” process by alerting viewers that upcoming content may spark the recall of traumatic memories, specifically, not just that provocative or sensitive material may be encountered (Haslam, 2017). These ideas about the original purpose of trigger warnings are therefore central to the debate about the use of trigger warnings for people suffering from PTSD, and/or trauma survivors. Therefore, it is important to investigate if trigger warnings may change personal appraisals such as how people recall a negative event, not only the immediate emotional reactions they might experience when encountering novel impersonal stimuli.

Trigger warnings and avoidance. Aside from emotional reactions, the complete avoidance of potentially distressing material is paradoxically argued by advocates as a benefit of trigger warnings and claimed by critics to be a harmful effect. On the one hand, advocates claim that warnings help people completely avoid content that may “trigger” a severe emotional reaction (Manne, 2015) and that avoidance may be the best tool available when exposure to trauma stimuli occurs outside a therapeutic setting (e.g., public environment; Boysen et al., 2021). Additionally, some advocates have argued that warnings do not always signal someone to avoid content altogether, but rather to let them confront the content in a different time and place in a safe environment, which could be with a therapist (e.g., DeBonis, 2019) or in their own home with a cup of tea (University of St. Thomas, 2015). On the other hand, critics argue that trigger warnings may encourage the complete avoidance of any material that someone deems distressing, reducing resilience (Lukainoff & Haidt, 2015; Medina, 2014). Indeed, critics point out that the behavioral avoidance of trauma related content is a coping strategy known to prolong PTSD symptoms (McNally, 2014; Lukainoff & Haidt, 2015).

Avoidance based coping. Avoidance-based coping involves avoiding a stressor and reactions to it. That is, avoidance could involve the complete behavioral avoidance of a situation (known as situation selection) such as leaving a lecture or turning off the TV after seeing a trigger warning. However, staying to view content could also constitute avoidance, if a person chooses to engage with trauma-related content after a trigger warning but tries to avoid their emotions and reactions while they do so (e.g., denying that the stressor exists or disengaging from or trying to suppress thoughts and feelings and emotions, or engaging in fantasy). Avoidance based coping is generally considered maladaptive and is associated with emotional distress following a stressful/traumatic event (Littleton et al., 2007). Additionally, avoidance behaviors are a symptom of many anxiety-based clinical disorders such as posttraumatic stress disorder (PTSD; Ehlers & Clark, 2000) and Generalized Anxiety Disorder (GAD; Salters-Pedneault, Tull, & Roemer, 2004). Furthermore, decreasing avoidance is key to the most efficacious therapy for PTSD—exposure therapy (Rauch, Eftekhari, & Ruzek, 2012). However, little to no research has addressed if *warnings* promote or do not promote avoidance type behaviors.

The Forbidden Fruit and Pandora Effects. One possibility is that trigger warnings increase anxiety and apprehension about upcoming content and therefore promote the avoidance of warned of material. For instance, a trauma survivor may see a trigger warning relating to their traumatic experience and avoid the warned of content the same way they would avoid other trauma related stimuli (e.g., people, places or objects associated with the original trauma; Ehlers & Clark, 2000). However, available research suggests that warnings might *increase* rather than decrease the attractiveness of content. Psychological reactance, “boomerang effects” or the “forbidden fruit effect” occurs when people’s freedom to engage in an experience is restricted, and that experience becomes more attractive (Ringold, 2002). For instance, “No Diving” signs increased the likelihood that students with a history of risky diving behaviors dove into the shallow end of the pool (deTurck & Goldhaber, 1989), warning labels (vs. no label) on cigarette packages increase existing smokers desire to smoke cigarettes (Hyland & Birrell, 1979), and exposure to anti-drug advertisements was associated with less negative attitudes towards using amphetamines and barbiturates (Feingold & Knapp, 1977). Particularly relevant for trigger warnings, warning labels about graphic content increase the desire to watch violent television shows (Bushman & Stack, 1996), and play violent video games (Bijvank, Konijn, Bushman, & Roelofsma, 2009). It is therefore possible that trigger warnings are not only ineffective in promoting avoidance behaviors, but actually attract people towards distressing content.

Another closely related area of research that suggests trigger warnings might not be effective in promoting the avoidance of potentially aversive stimuli is the *"Pandora effect*." According to the Pandora effect, in an effort to resolve uncertainty people will open a sealed box even if the contents of the box are expectedly negative. In fact, in a series of experiments, Hsee and Ruan (2016) demonstrated that people are *more* likely to engage with stimuli (i.e., open the box) if the consequences of such engagement are *uncertain* (vs. certain) and *negative* (vs. neutral) in nature (e.g., electric shocks, unpleasant sounds and disgusting images). These results may also reflect morbid curiosity or the tendency for people to seek out negative information. For instance, Oosterwijk (2017) found that participants willingly subjected themselves to negative images over neutral alternatives. In summary, the Pandora effect demonstrates that people are generally drawn towards negative material and the forbidden fruit effect shows that warnings can enhance this attraction, rather than deter it.

Of further concern is the idea that vulnerable people—the very people trigger warnings were originally designed to protect—might in fact be *least* deterred by warning messages or least likely to use warning messages as an avoidance tool. Evidence for this idea comes from research showing that vulnerable populations (e.g., people with mental illness) are often attracted towards negative content. For instance, people with prior lifetime exposure to violence, and fear of future terrorism, are more likely to seek out and watch disturbing content online, such as the graphic ISIS beheading video (Redmond, Jones, Holman, & Silver, 2019) and some trauma survivors engage in self-triggering and seek reminders of their traumatic experience (e.g., graphic imagery and media), a behavior that is associated with PTSD symptom severity (Bellet, Jones, & McNally, 2020). Similarly, people with or at risk of depression often choosing to expose themselves to negative rather than positive imagery

(LeMoult et al., 2018; Millgram, Joormann, Huppert, & Tamir, 2015). Given that vulnerable populations may be attracted towards negative content, it is possible that at worst labelling negative content with a warning message may increase engagement with this content and at best, have no effect on promoting avoidance.

But why might people with mental health vulnerabilities may be attracted, rather than deterred by, warnings? First, in line with Zillmann’s (1988) Mood Management Theory, we know that people often use media to regulate mood. Although we might expect that people would typically select positive media to repair negative mood, people may instead seek other emotional goals beyond immediate mood repair and engage in “counter-hedonistic” consumption behavior. Viewers may be driven by a desire to obtain information, or gain insight or justify one’s own feelings and situation; in other words, to make meaning (Loewenstein, 1994). Indeed, the desire to make meaning of a traumatic experience was the best predictor of how often participants self-triggered (Bellet et al., 2020). Furthermore, clinically depressed people (vs. non-depressed), are more likely to use emotion regulation strategies to maintain or increase their level of sadness rather than to alleviate it (Millgram et al., 2015), perhaps because sad moods are familiar to depressed people. Based on these ideas it seems possible that people with a tendency towards negative mood states—perhaps due to mental health vulnerabilities—would be more less likely to use trigger warnings as a tool for avoidance.

Trigger warnings and avoidance: The evidence so far. Only a handful of studies have examined the avoidance of material accompanied by warning messages. Kimble et al. (2021) found that only a small minority of participants (< 6%)—including those with a history of trauma or with probable PTSD—avoided reading potentially triggering text when provided with the option of reading a non-triggering alternative. However, participants were not issued with a trigger warning, they were instead made aware about the distressing nature of the reading via the informed consent procedure. Only two studies have explicitly examined the effects of trigger warnings on the avoidance of material. Gainsburg and Earl (2018) found that participants were slightly less likely to select a film title if it was accompanied by a trigger warning (probability of selection = 0.56, vs. the same title with no warning = 0.44)—however this difference was not statistically significant (*n* = 240). Similarly, Bruce and Roberts (2020) found no preference for articles labelled with trigger warnings compared to the same titles without warnings—including for participants who had experienced a past history of trauma matching the article. Therefore, the limited number of studies on the effects of trigger warnings on behavioral avoidance suggest that warnings have little to no impact or at the very least do not seem to encourage avoidance.

Due to the minimal number of studies specifically examining trigger warnings and behavioral reactions, many open questions remain. First, there has been a narrow exploration of avoidance coping behaviors. The three previous studies on trigger warnings and avoidance focused on a very narrow definition of avoidance coping—the complete behavioral avoidance of stimuli also known as *Situation Selection* (Gross, 2002). Participants were given the choice to pick news headlines (Bruce & Roberts, 2020), film titles (Gainsburg & Earl, 2018), and essay readings (Kimble et al., 2021) accompanied with or without trigger warnings. Therefore, no research has explored if warnings might promote other kinds of avoidance behaviors (e.g., the suppression of thoughts and feelings) if someone stays to view content following a trigger warning.

Second, previous studies have assessed avoidance using a narrow range of experimental stimuli. In fact, both Gainsburg and Earl (2018) and Bruce and Roberts (2020) examined *titles* (film and news articles) accompanied by trigger warnings. Therefore, participants were already given information about the stimuli via the information conveyed in the title and therefore were already somewhat warned about what the film/article might contain. This methodological feature could explain why there was no difference between the warning and no warning conditions—participants may have based their choice to approach/avoid the article on the title itself. However, we do not know what might happen in the absence of such information, for instance, in the case of the vague warnings used on social media websites such as Instagram. These warnings warn of negative content but do not provide any information about that content (i.e., "*Sensitive content: This photo/video may contain violent or graphic content*"). It is possible, based on previous research on the "Pandora Effect," that participants may be *more* likely to want to engage with content that is marked by a vague warning—compared to no warning—so they can close an information gap (e.g., "*what is the negative content the warning is referring to*?").

Third, previous studies have only examined a narrow range of vulnerable populations. Bruce and Roberts (2020) and Kimble et al., (2021) surveyed trauma survivors. However, prior research (e.g., LeMoult et al., 2018; Milgram et al., 2015; Redmond et al., 2019) suggests that it may be necessary to examine how other vulnerable populations (e.g., people with depression or lowered wellbeing) approach or avoid content with warning messages. Specifically, as stated above, warning messages may unnecessarily attract, rather than deter, certain vulnerable populations towards negative material.

Fourth, research has only examined if warnings change the avoidance of material *prior* to being exposed to it; no research has investigated if warnings might change the way someone avoids material after they are exposed. For instance, do trigger warnings change how long someone spends viewing negative material once they have decided to consume it? These ideas parallel with other online strategies to protect vulnerable populations, such as enabling people to close webpages quickly—e.g., the "Quick Exit" button on the new South Australia Victim Support Webpage.

1.3 Conclusion

In sum, advocates claim trigger warnings are helpful in reducing negative emotional reactions towards material, via mental preparation, and mitigating surprise, as well as promoting helpful avoidance behaviors. Despite these claims, evidence from literature on emotion regulation, bracing for the worst, expectancy effects, priming, psychological reactance, and the Pandora effect suggest that it is unlikely warnings exert these beneficial effects. Instead, existing psychological theory and empirical evidence lends more support to the claims of critics, suggesting that trigger warnings may actually be harmful or otherwise ineffective. However, research specifically investigating the effects trigger warnings on emotional and behavioral reactions is lacking. Therefore, my thesis aims to further investigate the emotional and behavioral effects of trigger warnings and determine if they are helpful or harmful.

2 Overview of Thesis Studies

Despite widespread use, there is a dearth of research specifically investigating the effects of trigger warnings on emotional and behavioral reactions. My thesis aims to bridge gaps left by the first wave of trigger warning research to help determine *how* and *why* these warnings may or may not change emotion and behavior.

Chapter 3 — Study 1a-1e

In Chapter 3 (Studies 1a-1e) I wanted to know *how* trigger warnings might cause negative expectations and prime participants to interpret neutral or ambiguous material (i.e., material that is not explicitly negative) in a negative way. Prior to this work, trigger warning research had focused on examining participants' reactions following a warning for a narrow range of negative stimuli: i.e., text passages (Bellet et al., 2019; Gainsburg and Earl, 2018) and films (Sanson et al., 2019). Across five experiments, participants viewed, or did not view, a message that photo material would be distressing, before rating the pleasantness and arousal of negatively valenced (paired with negative headlines), partially-ambiguous (paired with neutral headlines), or completely ambiguous (no accompanying headline) photos. I found, in line with previous work, that although trigger warnings foster negative expectancies about upcoming content, they have trivial effects on reactions to subsequent stimuli—including neutral or ambiguous material. That is, warnings seem to affect *anticipatory* anxiety, but not people's negative *reactions* towards content. However, part of trigger warnings' trivial impact means that they also do not seem to *reduce* people's negative reactions towards material.

Chapter 4 — Study 2

In Chapter 4 (Study 2), I had two main aims. First, I wanted to further expand the narrow range of stimuli employed in previous work. Specifically, I wondered *how* warnings might change emotional reactions towards *personally relevant* content—such as someone's own negative memories. The original purpose of a trigger warning message is to alert people to material that upcoming content may spark ("trigger") the recall of traumatic memories—not just that provocative or sensitive material may be encountered (Haslam, 2017). Therefore, I wondered if trigger warnings would change reactions towards *negative memories* themselves. To investigate this idea, I asked participants to recall a recent negative event over two sessions a fortnight apart. Prior to initial recall in the first session, participants were assigned either to a warning message—informing them that the negative memory task was distressing—condition or an unwarned control condition. I found that the emotional impact of the negative memory (the frequency of experiences related to the event such as "I had trouble staying asleep"), subsided less over a two-week period for participants who were *warned* in the first session. Second, I explored one possibility for *why* previous research has found that trigger warnings do not seem to help reduce distressing reactions—the strategies used to cope with negative content. That is, trigger warning advocates claim that trigger warnings enable people to "*prepare*" to cope with potentially distressing content. One way to operationalize preparation is to consider coping strategies—a conscious effort to manage the demands of a stressful situation (e.g., receiving a cancer diagnosis), using thoughts and behaviors (e.g., seeking social support; Folkman & Moskowitz, 2004). I found that warning participants about the distressing nature of the recall task did not increase reported coping strategies. My findings therefore suggest that warning messages may prolong the negative characteristics associated with memories over time, rather than prepare people to recall a negative experience.

Chapter 5 — Study 3

In Chapter 5 (Study 3), I focused specifically on how trigger warnings may or may not change the coping strategies that someone brings to mind when they encounter potentially triggering content. In Study 2, using a coping questionnaire, I did not find any evidence that trigger warnings changed the types of coping strategies participants used when asked about their memories for a negative event. In Study 3, I expanded this investigation to ask more generally about ways that someone might cope with encountering potentially negative material related to their personal stressful/traumatic experience in everyday life (e.g., TV, lecture, social media, etc.). I also used an open-box format so that participants could describe *what they would do* if they came across a trigger warning or content (i.e., with no warning) related to this experience (e.g., in the news, in a lecture, etc.), without prompting from a questionnaire. I found that thinking of encountering a trigger warning did not appear to change the coping strategies people brought to mind or people's expected emotional reactions compared to imagining encountering trauma-related content directly. This included both approach (e.g., focused on the stressor itself) and avoidance based (e.g., avoiding the stressor) strategies (Littleton, Horsley, John, & Nelson, 2007). Therefore, it is likely that warnings do not reduce emotional reactions towards negative material because they are doing little to change the way that someone "*prepares*" to encounter it.

Chapter 6 — Study 4

In Chapter 6 (Study 4), I conducted a conceptual replication of Chapter 5 to address key limitations and also to home in on *behavioral* reactions towards trigger warnings. Thus far, my research (Studies 1-3) and others' research has focused almost exclusively on how warnings may or may not change emotional reactions towards material. Furthermore, although intentions (e.g., "I plan to exercise") *generally* map onto future behavior (e.g., actually exercising; $r = 0.53$; Sheeran, 2002), they may sometimes be inconsistent with actual behavior—the intention-behavior gap (Sheeran & Webb, 2016). Therefore, in Study 4, rather than asking about participants' most stressful/traumatic experience, participants watched a traumatic film, and rather than asking participants to report on hypothetical avoidance behaviors towards trauma related material, I measured actual behavioral avoidance towards film related stimuli presented with and without trigger warning messages. I found that participants rarely avoided negative stimuli and did not avoid negative stimuli more when it was preceded by a trigger warning versus a neutral instruction screen—supporting other work that participants rarely avoid negative study material, even when given an option to do so (Kimble et al., 2021).

Chapter 7 — Studies 5a and 5b

In Chapter 7 (Studies 5a and 5b), I wanted to further explore how trigger warnings may or may not encourage the avoidance of negative material in a specific applied context that closely mirrors real world trigger warning use. Instagram's sensitivity screen initiative—where images are obfuscated with a blur and accompanied by a trigger warning—aims to allow people, and in particular "vulnerable people" with mental health concerns (e.g., depression, Posttraumatic Stress Disorder), to *avoid* potentially distressing content. In Study 5a, I asked participants how likely they would be to uncover a blurred image if they came across it on Instagram. In Study 5b, I presented participants with a mock Instagram photo viewing task where participants had the option to click to uncover ("see photo") a single blurred image or select "next photo" to skip uncovering the image. I found however that sensitivity screens are ineffective at deterring vulnerable and non-vulnerable users from approaching potentially graphic content. My findings suggest that alternative, empirically grounded methods for flagging potentially negative content on social media may be necessary.

Summary

Taken together, my thesis adds to a growing body of literature showing that trigger warnings seem to be ineffective in achieving their purported goals. First, trigger warnings cause anticipatory anxiety but do not ameliorate emotional reactions towards material, likely because they do little to "prepare" people to view negative content. Second, warnings also do not seem to promote avoidance behaviors—even within populations that have traditionally called for trigger warning messages for the express purpose to avoid content (e.g., trauma survivors). Further work should focus on how to develop alert systems or strategies that do achieve these commendable aims.

3 Investigating the effects of trigger warnings on ambiguous stimuli

Chapter 3 is published as:

Bridgland, V. M. E., Green, D. M., Oulton, J. M., & Takarangi, M. K. T. (2019). Expecting the worst: Investigating the effects of trigger warnings on reactions to ambiguously themed photos. *Journal of Experimental Psychology: Applied*, 25(4), 602-617. https://doi.org/10.1037/xap0000215

Author Contributions: I developed the study design with the guidance of MKTT, DMG, and JMO. I collected the data, and performed the data analysis and interpretation, and drafted the manuscript. MKTT, DMG, and JMO, contributed equally by making critical revisions to the manuscript. All authors approved the final version of the manuscript for submission.

Abstract

Trigger warnings are messages alerting people to content containing themes that could cause distressing emotional reactions. Advocates claim that warnings allow people to prepare themselves and subsequently reduce negative reactions towards content, while critics insist warnings may increase negative interpretations. Here, we investigated (a) the emotional impact of viewing a warning message, (b) if a warning message would increase or decrease participants' negative evaluations of a set of ambiguous photos, and (c) how participants evaluated overall study participation. We meta-analyzed the results of 5 experiments (N = 1,600) conducted online, and found that trigger warnings did not cause participants to interpret the photos in a more negative manner than participants who were unwarned. However, warned participants experienced a negative anticipatory period prior to photo viewing that did little to mitigate subsequent negative reactions.

Introduction

From forewarning of graphic sexual violence in Batman: The Killing Joke (Finkel, 2017), to violence portrayed in the Royal Opera house performance of Donizetti's Lucia di Lammermoor (Maddocks, 2016), trigger warning use has exploded far and wide. Although trigger warnings can be defined in different ways, here we adopt the common definition—consistent with academic (Bellet, Jones, & McNally, 2018; Sanson, Strange, & Garry, in press) and public (e.g., Harper, 2018; Malervy, 2018) use—that trigger warnings are "a statement at the start of a piece of writing, video, etc. alerting the reader or viewer to the fact that it contains potentially distressing material—*often used to introduce a description of such content*" (Oxford Dictionaries, 2018; our emphasis). Advocates claim that such warnings allow people to completely avoid *or* emotionally prepare themselves to reduce negative reactions towards content and protect mental health (e.g., Lockhart, 2016). However, critics insist warnings may have adverse effects, such as encouraging avoidance behaviors—known to increase distress and maintain PTSD (Ehlers & Clark, 2000; Littleton, Horsley, John, & Nelson, 2007)—and negative expectancies about sensitive topics that normalize fearful responses (e.g., Lukianoff & Haidt, 2015). We know that at least one-third of first year students around the world screen positive to an anxiety, mood, or substance abuse disorder (DSM-IV; Auerbach et al., 2018). Moreover, many universities now mandate the use of trigger warnings as part of mental health initiatives (e.g., Harris, 2016; Palmer, 2017). Thus, it is vital to empirically examine critics' claims that warnings may actually *increase* negative reactions to potentially negative content. Here, our aim was not to test the assumption that trauma survivors benefit from completing avoiding content (although, as noted, this assumption carries its own set of criticisms), but rather to investigate what may happen if someone views a warning and then *continues* to view content. Our overall focus was to investigate the effects of trigger warnings on emotional reactions. This was managed by examining emotional reactions in three key ways: (a) the emotional impact (e.g., mood and anxiety) of a warning message, (b) whether that warning message would increase participants' negative expectations about and emotional evaluations of a set of target ambiguous stimuli, and (c) how participants evaluated participating in the study overall. We present a meta-analysis of five experiments where we either warned or did not warn participants that a series of photographs would be distressing. In reality, the photos were ambiguous, depicting scenes that could be interpreted as positive, negative, or neutral. Additionally, we paired the photos with negative, neutral, or no news headlines (between subjects) that matched the photos. At the end, participants judged the costs and benefits of participation.

Little published research has examined the effects of warning messages, despite their widespread use. For instance, warnings in consent forms that give explicit forewarning about the nature of graphic films (vs. vague or no information) have been shown to cause participants to be more scared about seeing something unwanted (De Wied, Hoffman, & Roskos-Ewoldsen, 1997) and also to be more frightened and upset while watching (Cantor, Ziemke, & Sparks, 1984). We are aware of only two studies that have examined the effects of trigger warnings directly.

Bellet et al. (2018) found that participants who read a warning (vs. no warning) experienced greater anxiety while reading distressing text—but only when they believed words could cause harm. Warned participants also perceived themselves and trauma survivors as more vulnerable to emotional distress following a traumatic event. However, these effects were small, and warned and unwarned participants did not differ on anxiety ratings when exposed to less distressing content, or on implicit self-identification with vulnerable versus resilient traits. Sanson et al. (in press) found that although warned (vs. unwarned) participants expected that a film would be more negative prior to viewing, the warning had little impact on emotional distress, intrusive thoughts, or avoidance behaviors after film exposure. Some of these findings (e.g., Bellet et al., 2018) may be explained by response expectancy (Kirsch, 1985) or nocebo effects; when negative expectancies (e.g., expecting pain) lead to the exacerbation of negative outcomes (e.g., symptoms; Benedetti, Lanotte, Lopiano, & Colloca, 2007). Yet, other findings (e.g., Sanson et al., in press) suggest that the effects of trigger warnings are trivial; neither helpful nor harmful.

However, we should note a critical difference between previous research and our research. Here, rather than assessing how warnings may exacerbate participants' reactions to overtly graphic stimuli, we aimed to examine participants' evaluation of *ambiguous* stimuli. While we suspect that warnings may lead people to create an expectancy about their reaction to negative content, warnings could also act as a *prime* for subsequent content. For instance, it seems plausible that a warning about distressing content could prime participants to interpret a negative meaning from a set of ambiguous and technically non-threatening photographs (e.g., because the warning suggests something is negative here, it must be). If this were so, such findings would have major applications for the current use of warning messages. We assessed the role of priming in the current series of experiments.

According to spreading activation theory, when a concept has been primed, associated concepts and knowledge in memory become more accessible (Collins & Loftus, 1975). For instance, exposure to the word "death" primes participants to respond quicker to semantically related words like "distress" (semantic priming; Janiszewski & Wyer, 2014), and reading negative news articles increases memory for negative information in subsequent news articles (affective priming; Baumgartner & Wirth, 2012).

Importantly, priming effects also affect how people interpret *ambiguous* stimuli. For instance, when primed with negative (e.g., mean, selfish, rude) versus positive trait (e.g., sincere, creative, wise) adjectives, participants rated an ambiguous person in a photograph

(Ferguson, Bargh, & Nayak, 2005) higher on a number of negative traits. Furthermore, when primed with ethics-related words (vs. neutral), participants were more likely to categorize morally ambiguous behavior as unethical (Welsh & Ordonez, 2014). Applying such findings to the use of warnings, it is possible that viewing a trigger warning message may prime a negative mindset and cause people to interpret neutral or informational content (e.g., lecture content, news articles, etc.) in a negative way.

In the present experiments, participants viewed, or did not view, a message that photo material would be distressing, before rating the pleasantness and arousal of negatively valenced (paired with negative headlines), partially-ambiguous (paired with neutral headlines), or completely ambiguous (no accompanying headline) photo stimuli. We also assessed participants' expectations about photo content and their perception of the costs and benefits of participating in the study.

Because warnings lead to the development of negative expectancies and can make people more fearful of upcoming material (e.g., De Wied, et al., 1997; Sanson et al., in press), we predicted that warned participants would have lower positive affect accompanied by higher negative affect and anxiety prior to viewing the photos compared to unwarned participants. Because warnings can exacerbate emotional reactions to negative material (e.g., Bellet et al.; Cantor et al., 1984), we predicted that warned participants who viewed negatively valenced photos (ambiguous photos paired with negative headline) would report increased negative mood and anxiety, and rate the photos as more emotionally arousing, and negative in valence. Additionally, we know that priming concepts makes associated ideas more accessible and also changes the manner in which ambiguous stimuli are interpreted. Therefore, we predicted that warned participants would evaluate ambiguous stimuli (ambiguous photos paired with neutral headlines and photos without headlines) in line with a negative activation; for example, they may think "there must be something distressing happening in this photograph." Finally, in relation to participants' reactions towards study participation as a whole, we drew on findings from Yeater, Miller, Rinehart, and Nason (2012), which suggest that engagement with distressing topics causes a decrease in reported mental costs and can foster the belief that study participation is more beneficial for the self and others. Thus, if warnings do increase negative reactions, we predicted that warned participants would also rate lower costs and higher benefits associated with study participation.

We conducted five experiments. In Study 1a we gave an intense warning about graphic photo stimuli, replicated in Study 1b. In Study 1c, we reduced the intensity of the warning message to examine if removing extreme elements would change its impact, replicated in Study 1d. In Study 1e, we replicated Study 1a and 1b with a new set of photos and a pre and post-stimuli expectancy scale.

Method

This experiment was approved by the Flinders University Social and Behavioral Research Ethics Committee. We preregistered Study 1e (osf.io/zb2rw/registrations/), and the data, supplementary files, and materials for all five experiments can be found under this project: osf.io/zb2rw/. For all five experiments, we have reported all measures, conditions, and data exclusions.

Participants

Studies 1a-1d. Because prior research has found that the effect of warning messages on negative reactions to material ranges from negligible (e.g., Sanson et al., in press) to medium-large (e.g., Cantor, et al., 1984), we elected to detect a small-medium effect size of f = .18 (i.e., mid-way between the small and medium benchmarks .10-.25). An a priori power analysis with a power of .80 for a 2 x 3 between subjects' ANOVA (Fixed effects, special, main effects and interactions: numerator df = 2, groups = 6) found a sample size of 301 participants was required (G*Power; Faul, Erdfelder, Lang, & Buchner, 2007). We achieved this target *N* in almost every experiment.

Study 1e. We recalculated the power analysis using the $\eta_p{}^2$ value (.024) of the main effect of the trigger warning message on photo valence ratings (our main variable of interest) from Study 1a where this effect was the largest (*f* = .16). An a priori power analysis with a power of .80 revealed an ideal sample size of 395 participants (65-66 per cell), which we achieved in Study 1e.

Total meta-analysis sample size. After we completed Study 1e, we also examined power for our meta-analysis sample size using an R-Studio script from Quintana (2017) based on formulas from Valentine, Piggott, and Rothstein (2010). For a random effects model accounting for moderate heterogeneity, our sample size of an average of 150 per group (for our main variable of interest: warning conditions), for five effect size comparisons, yielded a power of .78 for detecting a small (*d* = 0.2) effect.

Across five experiments we recruited a total of 1,867 participants online through Amazon's Mechanical Turk. Participants received a payment of $1.50. We excluded 103 participants for failing at least one of two instructional attention checks (Oppenheimer, Meyvis, & Davidenko, 2009), 51 for recognizing the photos or people pictured in them, 16 for leaving the task during the study, 18 for self-reported issues with the survey (e.g., photos not loading, warning video not playing, etc.), and 79 for guessing the key hypothesis.[1] Our analyses focused on the remaining 1,600 participants (see Table 3.1).

[1] Warning conditions only. We excluded these participants on the basis that participants who detected our deception likely responded to our emotional measures in any number of disingenuous ways—given they guessed what we were expecting. For instance, upon detecting the deception, participants may have been more likely to report that the stimuli was not distressing because they were now contrasting the warning message with the stimuli. Alternatively, participants may have responded in line with demand effects, and reported that the stimuli was more distressing than what they actually thought. See osf.io/7pnbv/for full details of hypothesis guess responses.

Table 3.1

Characteristics of the 5 studies included in the meta-analysis

Study	Warning message	Materials shown	N excluded	N retained	Age			% Female	% Caucasian
					Range	Mean	SD		
1a	Intense warning	6 neutral IAPS photos	44	293	19-70	35.12	10.64	50.5	80.2
1b	Intense warning	6 neutral IAPS photos	42	306	18-75	39.44	12.85	58.8	73.2
1c	Reduced warning	6 neutral IAPS photos	64	302	19-79	35.04	11.18	49.3	73.5
1d	Reduced warning	6 neutral IAPS photos	60	302	19-76	39.08	12.77	58.3	76.2
1e	Intense warning	6 neutral Shutterstock photos, 2 NAPS photos	57	397	19-72	35.66	10.29	57.7	70.8

Design

In all experiments, we used a 2 (Trigger warning condition: trigger warning, no trigger warning) x 3 (Headline valence: negative, neutral, no headline) x 2 (Time: pre-photos, post-photos) mixed design. Participants were randomly assigned to one of six possible conditions. All studies were conducted online using Qualtrics survey software (Qualtrics, Provo, UT).

Measures

Studies 1a-1b: Trigger warning message. Participants in the trigger warning condition viewed the following message on their screen and simultaneously listened to it as an audio clip:[2]

[2] Warned participants had to identify the sound played in an audio clip among four multiple-choice options. Participants were given two chances to complete this task successfully, otherwise participation was terminated.

“*Warning: This study involves viewing photographs that show emotional events. Some may be very graphic and very negative in nature (e.g., trauma, war, torture, maltreatment and death). Some people may find this material distressing. Please do not proceed if you do not want to be exposed to this material or think that you may be adversely affected by being exposed to this material.”*

Photo ratings. Participants rated each photo using the valence (1 = *smile*, to 9 = *frown*) and arousal (1 = *aroused*, to 9 = *calm*) dimensions of the Self-Assessment Manikin (SAM; Bradley & Lang, 1994; Appendix A), as used in the original International Affective Picture System (IAPS) ratings (Lang, Bradley, & Cuthbert, 1997). The SAM depicts five figures along a continuum that allows for ratings that fall in-between each figure.

Image stimuli. Participants viewed six photos, depicting “ambiguous” scenes (e.g., image of passengers boarding a plane, image of two people embracing; Appendix B) from the IAPS3 (Lang et al., 1997). These photos were selected because of their prior use in a similar paradigm also using headlines to manipulate photo valence (Porter, ten Brinke, Riley, & Baker, 2014). Based on IAPS normative ratings, (ranging from 1 = *positive* to 9 = *negative*), the photos were neutral in valence (M = 3.28, SD = 0.86) and only moderately arousing (M = 4.14, SD = 0.65). Each image appeared for 5 seconds to allow participants time to view details of the photograph but not so long that ambiguity might be completely extinguished.

Headlines.[4] In two of the conditions, a negative headline (e.g., “‘*I’ve lost everything’ Mother takes photo of sons boarding plane shortly before fiery crash killing all*”; Appendix B) or a neutral headline (e.g., “*Boeing starts shipping their new Dreamliners to airlines*”) accompanied each photo. We selected headlines using data from two Mechanical Turk pilot studies. One group of participants (n = 42-45) rated “*how well does this headline match with*

[3] Selected image numbers: 4598, 7620, 8117, 8190, 8300 and 8400.

[4] See osf.io/6xstf/ for data on our headline piloting, and the headline materials we used in Studies 1a–1d and 1e.

the above picture?" (1 = *a very poor match*, 7 = *a very good match*), while another group (*n* = 45-52) rated the headlines using the valence SAM dimension. We averaged the total number of ratings to form a single match and valence rating for each headline. For our final headlines, we selected those with the highest valence and match ratings.

An independent samples t-test confirmed that the mean match rating for the neutral (*M* = 4.92, *SD* = 0.35) and negative headlines (*M* = 4.90, *SD* = 0.71) did not differ significantly, *t*(10) = -0.036, *p* = .97, *d* = 0.036. Similarly, the mean valence rating for the neutral headlines (*M* = 4.85, *SD* = 0.47) was significantly lower than the negative headlines (*M* = 8.00, *SD* = 0.39; *t*(10) = 12.75, *p* <.001, *d* = 7.29).

Positive Affect Negative Affect Schedule (PANAS; Appendix C). Participants completed the 20-item PANAS (Watson, Clark, & Tellegen, 1988) at three points (warning condition) or two points (no warning) during the study to assess mood. Participants rate how much they are currently experiencing 10 items measuring positive affect (e.g., excited) and 10 measuring negative affect (e.g., distressed) on a 5-point scale (1 = *very slightly or not at all*, 5 = *extremely*). The internal consistency is .85 for the Negative and .89 for the Positive Affect subscale, while the intercorrelation between these subscales is -.15 (Watson et al., 1988).

Short form Spielberger State-Trait Anxiety Inventory (STAI-6; Appendix D). Participants completed the STAI (Marteau & Bekker, 1992) at three points (warning condition) or two points (no warning) during the study. Responders rate how they feel at that present moment for three anxiety-present items (e.g., "*I am worried*") and three anxiety-absent items (e.g., "*I feel calm*";1 = *not at all*, 4 = *very much*). The scale has good internal consistency (.82; Marteau et al., 1992).

Posttest-reactions questionnaire (Appendix E). We assessed participants' reactions to the study using two subscales from the post-test reactions questionnaire by Yeater et al.

(2012).[5] Participants rated their agreement (1 = *I strongly disagree*, 7 = *I strongly agree*) with 10 "perceived benefit" statements (e.g., "*This study gave me insights into myself*") and 5 "mental costs" statements (e.g., "*This study was mentally exhausting*"). The internal consistency is .77 for the perceived benefit items and .69 for the mental cost items. Across our five studies, internal consistency for the benefit subscale (α = .85-.88) and cost subscale (α = .79-.85) were good.

Post-stimuli expectancy rating. To assess how participants found the experience of the photo imagery versus what they expected at the beginning of the study, we asked participants, "compared to what I expected before viewing the photos, the photos were: 1 = *much more negative*, 2 = *somewhat more negative*, 3 = *slightly more negative*, 4 = *as I expected*, 5 = *slightly more positive*, 6 = *somewhat more positive*, 7 = *much more positive*."

Studies 1c-1d

Studies 1c and 1d employed the same paradigm and stimuli as the previous two studies but we modified the warning to reduce its negativity. The new warning removed all content that would not be directly relatable to our photo set (e.g., mentions of torture and maltreatment), and therefore it more closely matched the content of the photos. Secondly, we swapped out the line "*some may be very graphic and negative in nature*" for "*the photographs contain negative themes that some people may find upsetting*".

Study 1e

Study 1e replicated Studies 1a-1b with three modifications. First, we used a new set of photos and accompanying headlines that more closely matched the description given in the original warning message (Appendix B). We used six photos from Shutterstock.com (under a

[5] These questions were a subset of a larger set of questions gauging participants' reactions to research participation (based on Carter-Visscher, Naugle, Bell, & Suvak, 2007; Cromer, Freyd, Binder, DePrince, & Becker-Blease, 2006; Edwards, Kearns, Calhoun, & Gidycz, 2009; Newman, Willard, Sinclair, & Kaloupek, 2001), which we collected with the intent of conducting an exploratory factor analysis. We do not report this analysis here. However, the data for all of these questions are available on OSF: osf.io/zb2rw/

standard image license) and two photos from the Nencki Affective Picture System (NAPS; Marchewka, Żurawski, Jednoróg, & Grabowska, 2014; Appendix B). While the old set of photos (used in Studies 1a-1d) depicted predominately extreme-sports related scenes, the new photos contained a variety of ambiguous scenes including images of war, accidents, childbirth, etc.

We selected new headlines—based on news media—based on two pilot studies with Mechanical Turk participants. Again, one group of participants (*n* = 40-46) rated how well the headline matched with the photo, while another group (*n* = 45-47) rated headline valence. For our final headlines, we selected those with the best match and valence ratings. An independent samples t-test confirmed that the mean match rating for the neutral (*M* = 5.21, *SD* = 0.64) and negative headlines (*M* = 5.67, *SD* = 0.45) did not differ significantly, *t*(14) = 1.59, *p* = .135, *d* = 0.14. Similarly, the mean valence rating for the neutral headlines (*M* = 4.97, *SD* = 0.59) was significantly lower than the mean rating for the negative headlines, (*M* = 7.87, *SD* = 0.21; *t*(14) = 13.10, *p* <.001, *d* = 6.55). Second, headlines appeared on a page *before* photos (rather than underneath) for a duration of five seconds, to prevent participants from attending to the headlines only. Third, we measured expectancy (negative and positive) about the study, by asking participants how negative/positive they expected the photos to be (1 = *very slightly or not at all,* 5 = *extremely*) *before* photo presentation on a number of verbal descriptors (e.g., frightening, sad, inspiring etc.; Bartsch & Mares, 2014). After all eight photos were presented, we asked participants to rate how they thought the photos were overall, using the same scale. We then compared scores from pre to post-stimuli presentation to assess how participants' initial expectancy about the photos differed from how they experienced the photos.

General procedure

Our cover story was that we were interested in evaluating people's judgments of news headlines and photos. Following consent, all participants completed demographic questions, the PANAS, and the STAI. Participants in the trigger warning condition then viewed and listened to the trigger warning message and completed the PANAS, STAI, and, in Study 1e, the pre-stimuli expectancy scale. Participants viewed the photos accompanied with negative headlines (negative valence conditions), neutral headlines (neutral valence conditions) or no headlines (no headline condition) in a randomized order. After each photo, participants had an unlimited time to rate the photo on emotional valence and arousal (in random order) on the next page. After rating all photos, participants rated how closely they paid attention to the photos,[6] completed the PANAS and STAI again, the post-test-reactions questionnaire, and the post-stimuli expectancy rating. Next, we gave participants one cue word unique to each photo (e.g., "soldier"), and asked them to list as many details as possible that they could recall about each photo (e.g., who/what was in the photo?). These data are not analyzed here.[7] To ensure response quality, we then asked participants if they left the task for any period of time, what they thought the study was investigating, if they recognized any of the photos/people pictured, and if they had any technical issues. We then debriefed participants.

Results

For all main between-subjects analyses, we used ESCI meta-analyses software (Cumming, 2016). For all main within-subjects analyses, we used Meta-Essentials software (Suurmond, van Rhee, & Hak, 2017). We have reported I^2 (the proportion of total variation in the estimates of effect that is due to heterogeneity between studies) and Tau (the estimated standard deviation between experiments) as measures of heterogeneity. We have used

[6] For all experiments, scores were close to '7 = extremeley closely' (vs. 1 = not at all closely) in the warned (*Ms* = 6.53-6.69 , *SDs* = 0.57-0.86), and unwarned conditions (*Ms* = 6.63-6.66, *SDs* = 0.56-0.71). There were no signifcant differences between warning conditions (*ps* = .102-.785, *ds* =0.03-0.18).

[7] See osf.io/zb2rw/.

random effects models for all meta-analyses to account for heterogeneity. For within-subjects analyses with heterogeneity present, we have reported the prediction interval (PI; a description of the range of observed effect sizes) as a better estimate of the true effect (Van Rhee, Suurmond, & Hak, 2015). Some measures were skewed and were not normalized by transformations so we have analyzed untransformed data.

All figures display the forest plots of effect sizes between studies. Each row represents one Study (1a-1e). For all between-subjects graphs (ESCI software; Cumming, 2016), the location of each square on the horizontal axis represents the effect size. The black lines extending either side of a square represent a 95 % confidence interval. The size of each square indicates the sample size and weighting an experiment is given in the meta-analysis. Finally, the diamond shows the result of the meta-analysis, with the center indicating the estimated effect size and the spread representing a 95% confidence interval. For all main paired-subjects analyses (Meta-Essentials software; Suurmond, et al., 2017), the location of each point on the horizontal axis represents the effect size. The black lines extending either side of a point represent a 95 % confidence interval. The size of each point indicates the sample size and weighting an experiment is given in the meta-analysis. Finally, the bottom row (6) represents the result of the meta-analysis, with the center of the point indicating the weighted average affect or combined effect size. The smaller black interval represents a 95% confidence interval while the larger grey interval is a prediction interval (description of the range of observed effect sizes).

Warning attrition

We first examined the dropout rate of participants who quit the survey directly after the warning presentation, or at the equivalent point in the non-warning condition—to evaluate whether a sub-group—e.g., of particularly sensitive people—were self-selecting out of the study. When we examined the data for everyone who started the survey[8] (total *n* = 2026, warning conditions *n* = 1091, no warning conditions *n* = 935), 2.7% (1.4% of total) participants opted out directly after the warning, while 1.2% (0.05% of total) participants opted out in the no warning condition. A Person Chi-square revealed that these percentages were significantly different $\chi^2(1) = 5.71$, $p = .017$, $\varphi = .053$ Thus, it is possible that a small proportion of sensitive participants opted out in the warning condition after viewing the warning. While small effects may not be very consequential in a single episode, they may matter once they start accumulating over time or when extrapolated to a population estimate (Funder & Ozer, 2019). Considering one example, 1.4% of a university population of 70,000 students would equate to 1,000 students avoiding material accompanied by a trigger warning message. However, this result should be interpreted with caution because the effect size is small, and the significant *p* value is likely due to the large sample size. Moreover, because there was no control condition where participants were presented with graphic photo imagery it is possible participants who dropped out at the warning message would have also dropped out once they actually came across these stimuli. This possibility would mean trigger warnings do not promote avoidance, but instead just cause people to avoid material that they would have avoided in any case.

The emotional impact of viewing a warning message.

We evaluated our first research aim—to examine the emotional impact of a trigger warning message—in two key ways, using mood and anxiety data. We first examined if the warnings made people feel more negatively aroused, by comparing *warned participants'* PANAS and STAI scores before and after the warning (see Table 3.2). Next, we compared warned participants with unwarned participants on mood and anxiety from baseline (i.e., the

[8] Here we included people who answered at least one question on the first PANAS after the consent form (e.g., excluding people who were disallowed entry to the survey because of a mobile device type, or failing the sound check).

first measurement taken prior to photo exposure and prior to the warning message for warned participants) to post-photo exposure. We conducted this analysis to ensure the warning and headline groups were equivalent at baseline, as well as to compare groups on any changes in scores over time scores from an equivalent baseline measurement. We also compared warned and unwarned participants before (post-warning for warned participants) and after photo exposure. We conducted this analysis to compare warned participants' reactions to unwarned participants immediately after the warned participants received the warning, and also to examine how emotional reactions changed over time from this post-warning point. We conducted several 2 (Trigger warning condition: trigger warning, no trigger warning) x 3 (Headline valence: negative, neutral, no headline) x 2 (Time: baseline, post-photos), and (Time: pre-photos, post-photos) repeated-measures ANOVAs on PANAS and STAI scores (Tables 3.3-3.9). We then meta-analyzed these data across our five studies.

Table 3.2

Summary of mean PANAS positive scores from pre-warning to post-warning by study

Study	Pre-warning		Post-warning		*n*	*d*	*t*	*p*
	M	*SD*	*M*	*SD*				
Positive affect								
1a	29.53	9.21	28.34	8.89	145	0.13	3.35	.001
1b	30.09	9.78	29.2	9.97	152	0.09	2.72	.007
1c	28.47	9.60	26.89	9.41	152	0.17	4.54	<.001
1d	29.55	9.1	28.98	9.38	150	0.06	1.8	.074
1e	28.24	9.27	27.16	9.48	200	0.12	4.11	<.001
Negative affect								
1a	14.37	7.09	15.12	7.31	145	0.1	-2.93	.004
1b	12.3	4.62	13.56	5.76	152	0.24	-4.13	<.001
1c	13.03	6.02	13.91	6.1	152	0.15	-3.99	<.001
1d	12.01	4.01	13.41	5.46	150	0.29	-4.92	<.001
1e	12.68	4.82	13.62	5.73	200	0.18	-4.38	<.001
State anxiety								
1a	10.38	3.71	11.41	4.1	145	0.26	-5.1	<.001
1b	9.68	3.26	11.17	4.08	152	0.4	-6.89	<.001
1c	10.14	3.59	11.41	3.97	152	0.34	-5.95	<.001
1d	9.89	3.51	10.79	4.09	150	0.24	-4.34	<.001
1e	10.4	3.96	11.39	4.28	200	0.24	-5.74	<.001

Table 3.3

Summary of mean PANAS positive scores for warning conditions from baseline to pre-photo, to post-photo exposure by study

Time	Warning condition	*M*	*SD*	*n*
Study 1a				
Baseline	TW	29.53	9.21	145
	No TW	30.02	9.21	148
Pre-photos	TW	28.34	8.89	145
Post-photos	TW	27.69	9.11	145
	No TW	29.58	9.6	148
Study 1b				
Baseline	TW	30.01	9.78	152
	No TW	29.43	9.21	154
Pre-photos	TW	29.2	9.97	152
Post-photos	TW	28.68	10	152
	No TW	29.04	9.48	154
Study 1c				
Baseline	TW	28.47	9.60	152
	No TW	31.26	9.26	150
Pre-photos	TW	26.89	9.41	152
Post-photos	TW	27.41	9.56	152
	No TW	30.45	9.07	150
Study 1d				
Baseline	TW	29.55	9.10	150
	No TW	29.11	9.10	152
Pre-photos	TW	28.98	9.38	150
Post-photos	TW	27.88	9.36	150
	No TW	28.21	9.13	152
Study 1e				
Baseline	TW	28.24	9.27	200
	No TW	28.50	9.04	197
Pre-photos	TW	27.16	9.48	200
Post-photos	TW	24.46	8.39	200
	No TW	24.86	8.63	197

Table 3.4

Summary of mean PANAS negative scores for warning conditions from baseline to pre-photo, to post-photo exposure by study

Time	Warning condition	*M*	*SD*	*n*
Study 1a				
Baseline	TW	14.37	7.09	145
	No TW	13.32	5.65	148
Pre-photos	TW	15.11	7.31	145
Post-photos	TW	14.83	7.5	145
	No TW	13.99	6.21	148
Study 1b				
Baseline	TW	12.30	4.62	152
	No TW	11.92	4.33	154
Pre-photos	TW	13.56	5.76	152
Post-photos	TW	13.13	5.79	152
	No TW	12.79	5.01	154
Study 1c				
Pre-photos	TW	13.03	6.02	152
	No TW	12.79	5.65	150
Pre-photos	TW	13.91	6.1	152
Post-photos	TW	13.87	6.31	152
	No TW	13.76	6.46	150
Study 1d				
Baseline	TW	12.01	4.01	150
	No TW	11.81	3.54	152
Pre-photos	TW	13.41	5.46	150
	No TW	11.81	3.54	152
Post-photos	TW	12.8	4.66	150
	No TW	12.61	4.36	152
Study 1e				
Pre-photos	TW	12.68	4.82	200
	No TW	12.82	6.02	197
Pre-photos	TW	13.62	5.73	200
Post-photos	TW	15.69	6.28	200
	No TW	16.34	7.56	197

Table 3.5

Summary of mean state anxiety scores for warning conditions from baseline to pre-photo, to pos- photo exposure by study

Time	Warning condition	*M*	*SD*	*n*
Study 1a				
Baseline	TW	10.38	3.71	145
	No TW	10.50	3.96	148
Pre-photos	TW	11.41	4.1	145
Post-photos	TW	11.16	4.03	145
	No TW	11.26	3.92	148
Study 1b				
Baseline	TW	9.68	3.26	152
	No TW	9.94	3.68	154
Pre-photos	TW	11.17	4.08	152
Post-photos	TW	11.07	3.81	152
	No TW	10.76	3.87	154
Study 1c				
Baseline	TW	10.14	3.59	152
	No TW	9.63	3.50	150
Pre-photos	TW	11.41	3.97	152
Post-photos	TW	11.2	3.82	152
	No TW	10.79	3.86	150
Study 1d				
Baseline	TW	9.89	3.51	150
	No TW	9.50	3.01	152
Pre-photos	TW	10.79	4.09	150
Post-photos	TW	10.99	3.96	150
	No TW	10.49	3.42	152
Study 1e				
Pre-photos	TW	10.40	3.96	200
	No TW	9.89	3.60	197
Pre-photos	TW	11.39	4.28	200
Post-photos	TW	13.27	4.54	200
	No TW	13.08	4.38	197

Table 3.6

Summary of mean ANOVA results for PANAS positive and negative subscale and state anxiety scores for the interaction between the trigger warning conditions and time (baseline to post-photos) by study

Study	*F*	*df*	*p*	η_p^2
Positive affect				
1a	6.12	1,287	.014[9]	.021
1b	3.65	1,300	.057	.012
1c	0.26	1,296	.612	.001
1d	2.17	1,296	.141	.007
1e	0.05	1,391	.821	.000
Negative affect				
1a	0.30	1,287	.584	.001
1b	0.04	1,300	.833	.000
1c	0.09	1,296	.766	.000
1d	0.002	1,296	.964	.000
1e	1.29	1,391	.257	.003
State anxiety				
1a	0.005	1,287	.943	.000
1b	3.34	1,300	.069	.011
1c	0.10	1,296	.845	.001
1d	0.12	1,296	.729	.000
1e	0.77	1,391	.380	.002

[9] Warning conditions were equivalent at baseline (*p* = .661, *d* = 0.05), and post-photo exposure (*p* = .094, *d* = 0.20). Warned particiopants reported lower possitive affect from baseline to post photo exposure (*p* < .001, *d* = 0.20), while there was no difference for unwarned participants (*p* = .248, *d* = 0.05).

Table 3.7

Summary of mean ANOVA results for PANAS positive and negative subscale and state anxiety scores for the interaction between the trigger warning conditions and time (pre-photos to post-photos) by study

Study	*F*	*df*	*p*	$\eta_p{}^2$
Positive affect				
1a	0.1	1,287	.750	.000
1b	0.03	1,300	.862	.000
1c	6.35	1,296	.012[10]	.021
1d	0.14	1,296	.712	.000
1e	2.81	1,391	.094	.007
Negative affect				
1a	6.16	1,287	.014	.021
1b	12.02	1,300	.001	.039
1c	5.74	1,296	.017	.019
1d	10.81	1,296	.001	.035
1e	9.24	1,391	.003	.023
State anxiety				
1a	10.67	1,287	.001	.036
1b	9.27	1,300	.003	.03
1c	20.68	1,296	<.001	.065
1d	6.17	1,296	.014	.02
1e	13.01	1,391	<.001	.032

[10] Warned particpnats expressed lower positive affect than unwarned (p <.001, d = .47) at pre photo exposure. Recall, trigger warning conditions for Study 1c were significantly different at baseline, likely accounting for this difference. See osf.io/842gv/ for full pattern of results.

Table 3.8

Summary of mean ANOVA simple effects analyses for PANAS negative subscale scores for warning conditions from pre to post-photo exposure by study

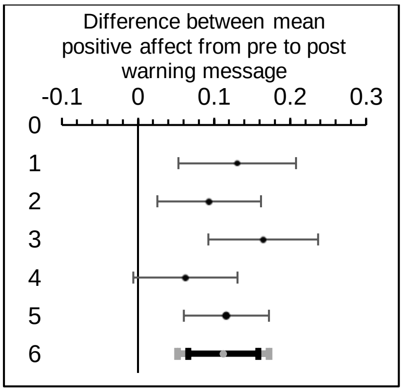
Condition	Condition	*p*	*d*
Study 1a			
Pre-photos	TW vs No TW	.022	0.27
Post-photos	TW vs No Tw	.330	0.12
TW	Pre vs post photos	.268	0.04
No TW	Pre vs post photos	.017	0.11
Study 1b			
Pre-photos	TW vs No TW	.005	0.32
Post-photos	TW vs No Tw	.609	0.06
TW	Pre vs post photos	.084	0.07
No TW	Pre vs post photos	.002	0.19
Study 1c			
Pre-photos	TW vs No TW	.099	0.19
Post-photos	TW vs No Tw	.867	0.02
TW	Pre vs post photos	.93	0.01
No TW	Pre vs post photos	.001	0.16
Study 1d			
Pre-photos	TW vs No TW	.003	0.35
Post-photos	TW vs No Tw	.718	0.04
TW	Pre vs post photos	.036	0.12
No TW	Pre vs post photos	.011	0.20
Study 1e			
Pre-photos	TW vs No TW	.173	0.14
Post-photos	TW vs No Tw	.34	0.09
TW	Pre vs post photos	<.001	0.34
No TW	Pre vs post photos	<.001	0.52

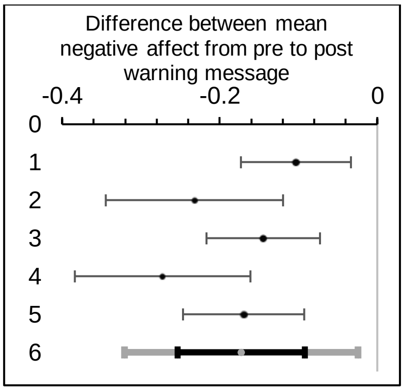
Table 3.9

Summary of mean ANOVA simple effects analyses for state anxiety scores for warning conditions from pre to post-photo exposure by study

Time	Warning condition	*p*	*d*
Study 1a			
Pre-photos	TW vs No TW	.057	0.23
Post-photos	TW vs No Tw	.745	0.03
TW	Pre vs post photos	.216	0.06
No TW	Pre vs post photos	.001	0.19
Study 1b			
Pre-photos	TW vs No TW	.006	0.32
Post-photos	TW vs No Tw	.498	0.08
TW	Pre vs post photos	.552	0.03
No TW	Pre vs post photos	<.001	0.22
Study 1c			
Pre-photos	TW vs No TW	<.001	0.48
Post-photos	TW vs No Tw	.354	0.11
TW	Pre vs post photos	.343	0.05
No TW	Pre vs post photos	<.001	0.32
Study 1d			
Pre-photos	TW vs No TW	.002	0.36
Post-photos	TW vs No Tw	.228	0.14
TW	Pre vs post photos	.425	0.05
No TW	Pre vs post photos	<.001	0.31
Study 1e			
Pre-photos	TW vs No TW	<.001	0.38
Post-photos	TW vs No Tw	.681	0.04
TW	Pre vs post photos	<.001	0.43
No TW	Pre vs post photos	<.001	0.77

Pre-warning to post-warning within-subjects analyses. Participants reported lower positive affect (Figure 3.1a, *d* = 0.11, 95% CI [0.07, 0.16], z = 6.76, *p* <.001, I^2 = 14.19%, Tau = 0.01), higher negative affect (Figure 3.1b, *d* = 0.17 95% CI [-0.25, -0.09], z = -5.94, *p* <.001, Tau = 0.04, I^2 = 52.27%, PI = 95% CI [-0.32, -0.02]), and higher state anxiety after receiving the trigger warning compared to before the warning (Figure 3.1c, *d* = 0.28, 95% CI [-0.37, -0.20], z = -9.52, *p* <.001, Tau = 0.04, I^2 = 37.09%, PI 95% CI [-0.42, -0.15]). Thus, the trigger warning seemed to have a small negative effect on mood (decreasing positive affect and increasing negative affect) and state anxiety ratings (increasing anxiety).





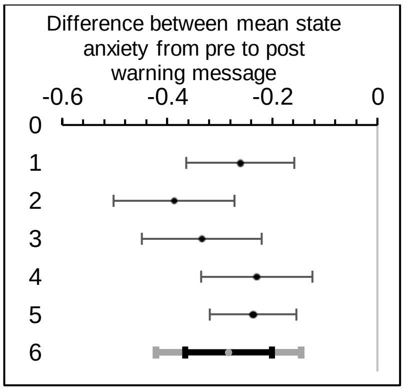


Figure 3.1a. The difference between mean positive affect ratings from before to after the trigger warning message. Positive values indicate a decrease in PA scores from pre to post-warning, i.e., participants rated lower positive mood from pre to post-warning.

Figure 3.1b. The difference between mean negative affect ratings from before to after the trigger warning message. Negative values indicate an increase in NA scores from pre to-post warning, i.e., participants rated higher negative mood from pre to post-warning.

Figure 3.1c. The difference between mean state anxiety ratings from before to after the trigger warning message. Negative values indicate an increase in state anxiety scores from pre to-post warning, i.e., participants rated higher state anxiety from pre to post-warning.

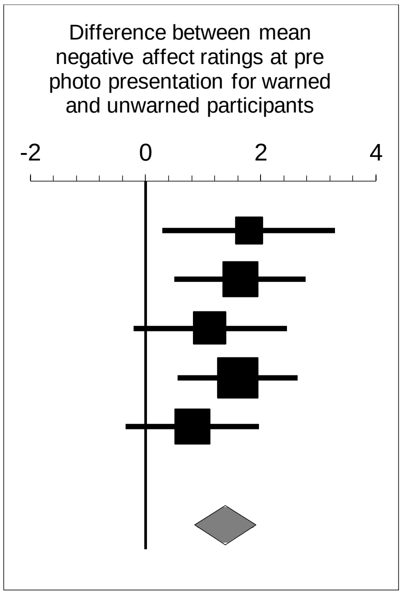
Baseline to post-photos between-subjects analyses. First, we examined if the headline conditions were equivalent at baseline and if headline valence manipulations were consistent with expectations. The interaction between headline conditions and time (from baseline to post-photo exposure) was significant for PANAS positive and negative subscales and state anxiety for all studies (*ps* <.001, $\eta_p{}^2s$ = .04-.166). There were no significant baseline differences for headline conditions, with the exceptions of Study 1a and 1e for PANAS negative affect only.[11] This exception is not unexpected: due to “the dance of the mean” (the sampling distribution of sample means), differences at baseline are not unusual (Cumming, 2012, p. 58). Thus, we have used meta-analyses to increase the precision of our estimates. In most cases, headline interactions suggest valence manipulations were successful (i.e., negative conditions experienced as more negative).[12] However, there were also no interactions between headline condition, warning condition, and time for negative affect and state anxiety in any of the studies (*ps* = .072-.893, $\eta_p{}^2$ = .001-.013)—suggesting that the six

[11] See https://osf.io/ud2ze/for complete inferential and descriptive results pertaining to baseline differences.
[12] See supplemental material at osf.io/ud2ze/ for complete inferential and descriptive statistics.

cells of our design were not significantly different at baseline, or, against predictions, at post-photo exposure. Moreover, there were no significant interactions between trigger warning condition and time (from baseline to post-photo exposure) for PANAS positive or negative affect subscales, or state anxiety, with the exception of Study 1a for positive affect (see Tables 3.3, 3.4, 3.5, and 3.6), suggesting the warning had little effect on emotional reactions over the course of the study. Thus, we did not meta-analyze these comparisons.

Pre-photos (post warning for warned participants) to post-photos between-subjects analyses. Against predictions, there were no interactions between headline condition, warning condition, and time for positive affect, negative affect and state anxiety in any of the studies (*ps* = 0.059-.909, $\eta_p{}^2$ = .001-.014). Moreover, the interactions between the trigger warning and time conditions (from pre-photos to post-photos) were not significant for positive affect, except in Study 1c (Table 3.7). Because the trigger warning appeared to have little influence on positive affect over time, we did not investigate these data further. Thus, going forward, we have focused on the interactions between the warning condition and time for PANAS negative affect and state anxiety.

For negative affect and state anxiety, the interaction between the trigger warning and time conditions was significant in all studies (Table 3.7). Prior to viewing the photos (and after the warning message for warned participants), warned (vs. unwarned) participants reported higher negative affect (Figure 3.2a, effect size for this group difference = 1.38, 95% CI [0.85,1.92], t = 5.05, p <.001, I^2 = 0%, Tau = 0, 95% CI [0,0.68]) and state anxiety (Figure 3.2b, effect size for this group difference = 1.36, 95% CI [0.99,1.74], t = 7.10, p <.001, I^2 = 0%, Tau 0, 95% CI [0,0.56]). The ANOVA results for all five studies revealed no significant differences between warned and unwarned participants on negative affect or state anxiety after viewing the photos, so we did not meta-analyze these data (see Tables 3.8 and 3.9).



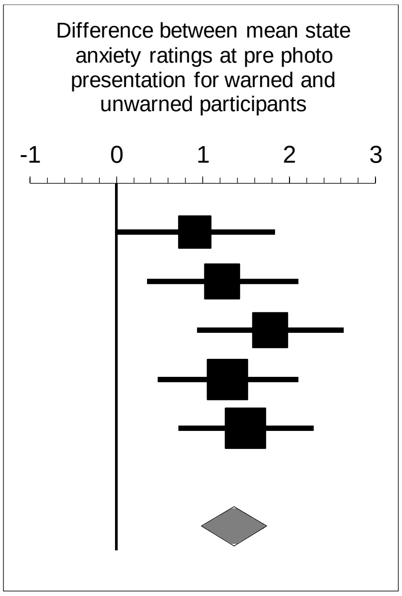
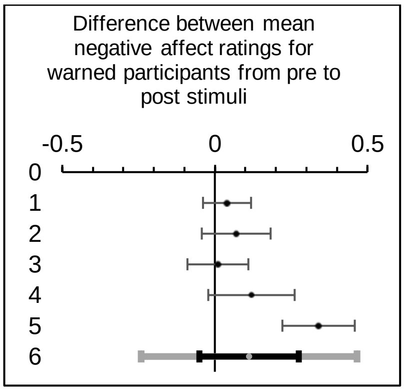


Figure 3.2a. The difference between mean negative affect ratings in the trigger warning and no warning conditions before photo presentation. Positive values indicate a higher mean score for the warning condition, i.e., negative affect scores were rated higher when participants were warned.

Figure 3.2b. The difference between mean state anxiety ratings in the trigger warning and no warning conditions before photo presentation. Positive values indicate a higher mean score for the warning condition, i.e., state anxiety were rated higher when participants were warned.

We then examined warned versus unwarned participants separately. Warned participants did not experience any significant change in negative affect (Figure 3.3a, *d* = 0.11 95% CI [-0.05, 0.28], z = 1.19, *p* = .056; Tau = 0.11, I^2 = 81.51%, PI = 95% CI [-0.24, 0.47]) or state anxiety (Figure 3.3b, *d* = 0.12 95% CI [-0.09, 0.33], z = 1.61, *p* = .107, Tau = 0.15, I^2 = 86.29%, PI = 95% CI [-0.35, 0.59]) over the course of viewing the photos. However, unwarned participants reported significantly more negative affect (Figure 3.4a, *d* = 0.24 95% CI [0.03, 0.44], z = 3.23, *p* = .001, Tau = 0.15, I^2 = 89.44%, PI = 95% CI [-0.24, 0.71]) and state anxiety (Figure 3.4b, *d* = 0.36 95% CI [0.07, 0.65], z = 3.45, *p* = .001, Tau = 0.20, I^2 = 91.02%, PI = 95% CI [-0.28, 1]) from pre to post-photo viewing. However, recall that when we examined warned and unwarned participants from equivalent baseline estimates (i.e., from before the warning message for warned participants and prior to photo exposure for unwarned participants) we found no significant interactions between warning condition and time. Thus, the increase in negative affect we observed here for the unwarned participants likely arises from the fact that unwarned participants' initial negative affect and anxiety was lower than warned participants due to no warning being present.



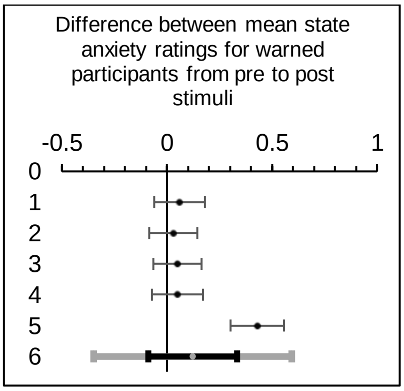
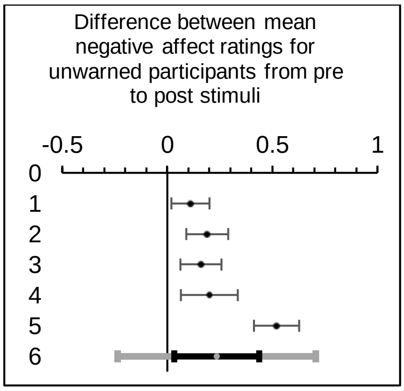


Figure 3.3a. The difference between mean negative affect from before to after the photo stimuli for warned participants. Positive values indicate an increase in NA scores from pre to-post stimuli i.e., warned participants NA ratings from pre to post stimuli were not found to be significantly different in the meta-analysis.

Figure 3.3b. The difference between mean state anxiety from before to after the photo stimuli for warned participants. Positive values indicate an increase in state anxiety scores from pre to-post stimuli i.e., warned participants state anxiety ratings from pre to post stimuli were not found to be significantly different in the meta-analysis.



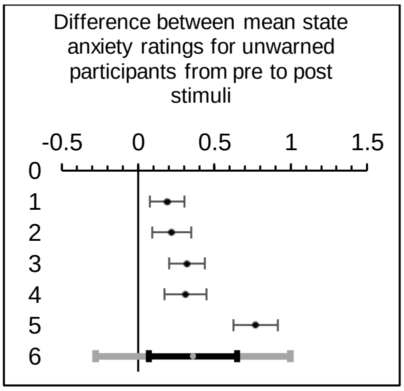


Figure 3.4a. The difference between mean negative affect from before to after the photo stimuli for unwarned participants. Positive values indicate an increase in NA scores from pre to-post stimuli i.e., unwarned participants rated higher negative mood from pre to post stimuli.

Figure 3.4b. The difference between mean state anxiety from before to after the photo stimuli for unwarned participants. Positive values indicate an increase in state anxiety scores from pre to post-stimuli i.e., unwarned participants rated higher state anxiety from pre to post stimuli.

Taken together, our data suggest the warning message created a noxious anticipatory period prior to photo viewing that lasted for the duration of the photo viewing (i.e., because negative affect and anxiety did not decrease significantly over time from the point of the warning onwards). Moreover, the change in mood from baseline (prior to the warning message for warned participants) to post-photo exposure was equivalent for warned and unwarned participants. Thus, while the warning message did not dramatically increase negative reactions it also provided no emotional benefits.

Effect of the warning message on photo expectancies and evaluations.

We next examined our second aim: whether emotional priming created by the warning—evidenced by the warning's initial effect on mood and anxiety—would increase participants' negative expectations about, and evaluations of ambiguous photos. First, we analyzed participants' expectations about the photos taken prior to (Study 1e), and the photos' consistency with expectations after (Studies 1a-1e), photo presentation. Second, we examined how (un)pleasant and arousing participants rated the photos.

Post-stimuli expectancy rating. We conducted several 2 (Trigger warning condition: trigger warning, no trigger warning) x 3 (Headline valence: negative, neutral, no headline) between groups ANOVAs on photo expectation ratings (Table 3.10). Across all studies, there was no significant interaction between the trigger warning and headline conditions (*ps* = .142-.667, $\eta_p{}^2$ = 002-.013), so in the following analyses we have only focused upon the main effect of the trigger warning condition.[13] Participants who received a warning reported the photos were significantly more positive than they had expected (Figure 3.5, effect size for this group difference = 0.81, 95% CI [.66, 0.97], *t* = 10.5, *p* <.001, I^2 = 0%, and Tau = 0, 95% CI [0,0.14]).

[13] Most headline main effects were consistent with manipulation expectations; negative headline conditions were rated closer to "more negative" than neutral and no headline conditions. See osf.io/842gv/ for complete inferential and descriptive statistics.

Table 3.10

Summary of mean photo expectation scores for trigger warning conditions by study

Study	Trigger warning			No trigger warning			*d*	*F*	*p*	η_p^2
	n	*M*	*SD*	*n*	*M*	*SD*				
1a	145	5.63	1.47	148	4.87	1.54	0.51	26.61	<.001	.085
1b	152	5.8	1.51	154	4.96	1.67	0.53	30.54	<.001	.092
1c	152	5.48	1.64	150	4.83	1.64	0.40	15.54	<.001	.05
1d	150	5.7	1.46	152	4.82	1.65	0.57	32.85	<.001	.10
1e	200	4.5	1.42	197	3.61	1.53	0.60	38.36	<.001	.089

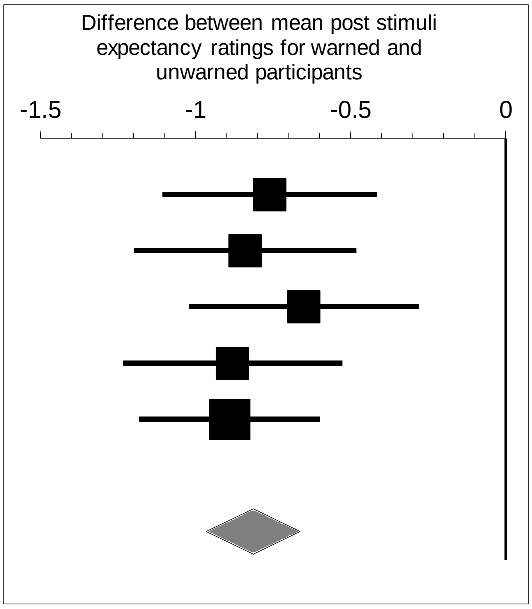


Figure 3.5. The difference between mean photo expectation ratings in the trigger warning and no warning conditions. Negative values indicate a higher mean score for the warning condition, i.e., photos rated more positive than expected, on a scale of 1 = Much more negative to 7 = Much more positive, when participants were warned.

Pre-stimuli and post-stimuli expectancy scale. We used a 2 (Trigger warning condition: trigger warning, no trigger warning) x 3 (Headline valence: negative, neutral, no headline) x 2 (Time: pre-photos, post-photos) repeated-measures ANOVAs on the pre-post photo ratings (Table 3.11). There was no interaction between warning, headline and time conditions. However, there was a significant interaction between trigger warning condition and time, *F*(2,391) = 76.1, *p* <.001, η_p^2 = .163. Pre-photos, warned participants (*M* = 2.89, *SD* = 0.68) expected the photos to be significantly more negative than unwarned participants did (*M* = 2.32, *SD* = 0.72; *p* <.001, *d* = 0.81). Post-photos, however, warned participants (*M* =

2.69, *SD* = 0.57) rated the photos significantly less negative overall compared to unwarned participants (*M* = 2.83, *SD* = 0.67; *p* = .009, *d* = .23), and also compared to their pre-photo rating (*p* = .001, *d* = 0.32); the opposite pattern was true for unwarned participants (*p* <.001, *d* = 0.73).

Table 3.11

Summary of expectation scores at pre-photo and post-photo time for trigger warning and no warning conditions in Study 1e

	Trigger warning			No trigger warning			*d*	*p*
	n	*M*	*SD*	*n*	*M*	*SD*		
Pre-photo stimuli								
	200	2.89	0.68	197	2.32	0.72	0.81	<.001
Post-photo stimuli								
	200	2.69	.57	197	2.83	0.67	0.23	.009

Note. Lower numbers indicate that participants found the photos more negative than expected while higher numbers indicate participants found the photos more positive than expected.

Evaluations of photos. We averaged the SAM valence and arousal scores for the six (Studies 1a-1d) and eight photographs (Study 1e) to create mean valence and arousal scores. For these data, we conducted several 2 (Trigger warning condition: trigger warning, no warning) x 3 (Headline valence: negative, neutral, no headline) between-groups ANOVAs (Table 3.12).

Table 3.12

Summary of mean photo valence and arousal scores by study for trigger warning and no warning conditions

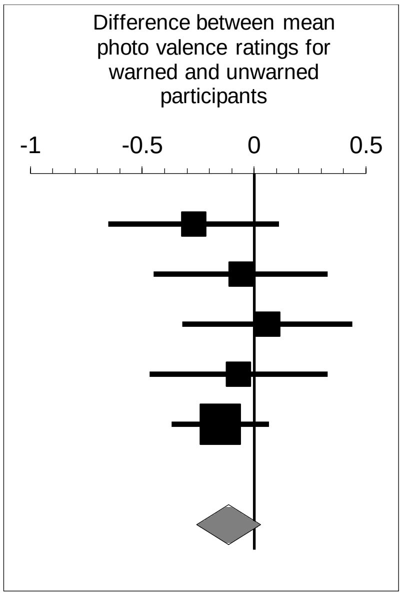
Study	Trigger warning			No trigger warning			*d*	*F*	*p*	η_p^2
	n	*M*	*SD*	*n*	*M*	*SD*				
Valence										
1a	145	5.02	1.72	148	5.29	1.6	0.16	7.17	.008	.024
1b	152	5.15	1.77	154	5.21	1.7	0.04	0.56	.455	.002
1c	152	5.21	1.65	150	5.15	1.71	0.04	0.29	.587	.001
1d	150	5.19	1.75	152	5.26	1.76	0.04	0.43	.515	.001
1e	200	6.42	1.11	197	6.57	1.1	0.14	3.33	.069	.008
Arousal										
1a	145	5.94	1.68	148	5.47	1.54	0.29	6.11	.014	.021
1b	152	5.8	1.65	154	5.36	1.45	0.28	5.94	.015	.019
1c	152	5.44	1.43	150	5.24	1.57	0.13	1.53	.217	.005
1d	150	5.3	1.56	152	5.45	1.69	0.09	0.69	.406	.002
1e	200	5.82	1.53	197	5.65	1.65	0.11	1.11	.294	.003

Photo valence. In all studies the interactions between the headline and warning conditions were not significant (*ps* = 0.425-0.96, η_p^2 = 0-.006), so we focused on the main effect of trigger warning condition.[14] The trigger warning seemed to have little to no effect on the perceived valence of the photographs (Figure 3.6a; group difference effect size = 0.11 95% CI [-0.26, 0.03], *t* = -1.55, *p* = .122, I^2 = 0%, Tau = 0, 95% CI [0,0.19]).

Photo arousal. The interactions between the headline and warning conditions were not significant in any of the studies (*p*s = .068-.977, $\eta_p{}^2$= 0-.018), so again we focused on the main effect of the trigger warning condition.[15] Warned participants rated the photos as less arousing (calmer) than unwarned participants (Figure 3.6b, group difference effect size = 0.23, 95% CI [0.02, 0.44], *t* = 2.1, *p* = .034, I^2 = 45.3%, Tau = 0.16, 95% CI [0,0.35]).

[14] Headline main effects were consistent with manipulation expectations; participants rated photos with negative headlines as more negative than photos with neutral or no headlines (*p*s <.001, *d*s = 1.44-3.01). See osf.io/txucj/ for complete inferential and descriptive statistics.

[15] Significant headline main effects in Studies b, c and e only (*ps* = .012-.034, η_p^2 = .017-.029), inconsistent with manipulation expectations (i.e., negative headlines not rated as more arousing; See osf.io/txucj/ for complete inferential and descriptive statistics.



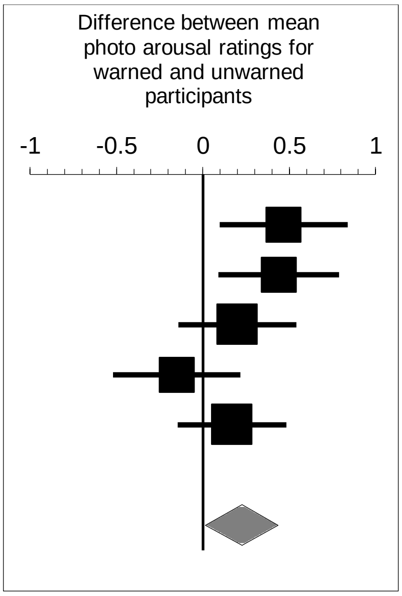


Figure 3.6a. The difference between mean photo valence ratings in the trigger warning and no warning conditions. Negative values indicate a lower mean score for the warning condition, i.e., valence scores were rated lower (less negative) when participants were warned, however this was not significantly different in the meta-analysis.

Figure 3.6b. The difference between mean photo arousal ratings in the trigger warning and no warning conditions. Positive values indicate a higher mean score for the warning condition, i.e., and arousal scores were higher (more calm) when participants were warned.

The effect of the warning message on holistic evaluations of study participation.

Our final aim was to examine whether warned participants would perceive higher benefits and decreased costs to overall study participation (Consistent with Yeater et al., 2012). We averaged the benefit and cost subscale items to create benefit and cost scores and conducted several 2 (Trigger warning condition: trigger warning, no warning) x 3 (Headline valence: negative, neutral, no headline) between-groups ANOVAs (Table 3.13).

Table 3.13

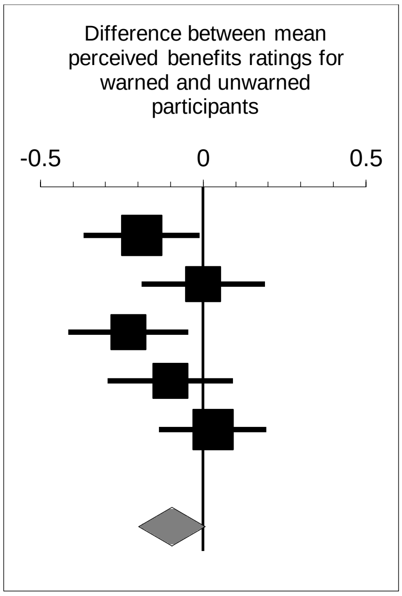
Summary of mean benefits and costs subscale scores by study for trigger warning and no warning conditions

Study	Trigger warning			No trigger warning			*d*	*F*	*p*	η_p^2
	n	*M*	*SD*	*n*	*M*	*SD*				
Benefits subscale										
1a	145	5.07	0.79	148	5.26	0.76	0.25	4.43	.036	.015
1b	152	5.2	0.83	154	5.2	0.85	0	0.004	.95	0
1c	152	5.1	0.78	150	5.33	0.85	0.28	6.52	.011	.022
1d	150	5.09	0.87	152	5.19	0.83	0.12	1.05	.307	.004
1e	200	5.11	0.87	197	5.08	0.8	0.04	0.12	.729	0
Costs subscale										
1a	145	2.61	0.95	148	2.53	0.84	0.001	0.55	.46	.002
1b	152	2.55	0.78	154	2.37	0.78	0.23	4.34	.038	.014
1c	152	2.47	0.78	150	2.56	0.79	0.12	0.9	.343	.003
1d	150	2.53	0.77	152	2.39	0.79	0.18	2.38	.124	.008
1e	200	2.8	0.79	197	2.81	0.9	0.01	0.01	.915	0

The interaction between the headline and warning conditions was not significant in any of the studies for the perceived benefits (*p*s = .234-.768, η_p^2 = .01-.002),[16] or costs associated with participation (*p*s = .129-.871 η_p^2 = .001-.014, with the exception of Study 1e, *p* = .03, η_p^2 = .018).[17] Thus, we only focus on the main effect of trigger warning condition. The warning seemed to have little to no effect on the perceived benefits (Figure 3.7a, group difference effect size = -0.1 95% CI [-0.2, 0.01], *t* = -1.84, *p* = .066, Tau = 0.07, 95% CI [0, 0.17], I^2 = 37.19%) or costs associated with study participation (Figure 3.7b, group difference effect size = 0.06 95% CI [-0.04, 0.16], *t* = 1.16, *p* = .244, Tau = 0.07, 95% CI [0, 0.16], I^2 = 34.15%).

[16] Significant headline main effects for Studies 1c and 1d only. See osf.io/pyzsk/ for complete inferential and descriptive statistics.

[17] Headline main effects were all significant (*ps*<.001, η_p^2 = .052-.092). Negative headlines were rated as significantly more costly than neutral and no headlines (*ps* = <.001-.005, *ds* = 0.46-0.73). See osf.io/pyzsk/ for complete inferential and descriptive statistics.



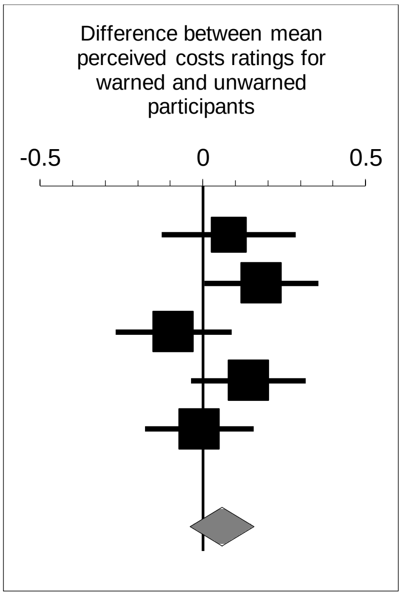


Figure 3.7a. The difference between mean benefit subscale ratings in the trigger warning and no warning conditions. Positive values indicate a higher mean score for the warning condition, i.e., study participation rated more beneficial, however this was not significantly different in the meta-analysis.

Figure 3.7b. The difference between mean cost ratings in the trigger warning and no warning conditions. Positive values indicate a higher mean score for the warning condition, i.e., study participation rated more mentally costly, however this was not significantly different in the meta-analysis.

Discussion

In the present studies, we aimed to investigate the effects of trigger warnings on emotional reactions. We explored this in three key ways: (a) the emotional impact of viewing a warning message, (b) if a warning message would increase participants' negative expectations about, and evaluations of, a set of negative and neutrally valenced ambiguous photos, and (c) if warned participants would evaluate the entire study experience negatively. Across our five studies, trigger warnings led to consistent increases in participants' negative mood, state anxiety, and negative expectations prior to photo presentation. These findings support the idea that trigger warnings prime participants to expect negative content. However, this negative priming did not increase negative interpretations of the ambiguous photos (i.e., valence). Warned participants rated the photos as significantly less arousing than unwarned participants, but this effect was small. Similarly, the warning did not change participants' perceived costs and benefits of participation.

Consistent with predictions, trigger warnings evoked an immediate decrease in positive affect, increases in negative affect and state anxiety, and primed a negative expectancy about the upcoming photo content. This effect also seemed to hold when we removed the more extreme elements of the warning message (Studies 1c and 1d). Moreover, a small but significantly higher percentage of people opted out directly after the warning message (vs. at the equivalent time in the no warning conditions), suggesting that the warning message may not only create negative arousal but may also encourage counter therapeutic avoidance behaviors for a small fraction of people (Littleton et al., 2007). However due to the small effect size (likely due to our large sample), future research should assess this conclusion more directly.

Our subsequent predictions that trigger warnings increase people's sensitivity or manifest fearful evaluations about towards content (e.g., Lukianoff & Haidt, 2015) remain unsubstantiated. Trigger warnings had little influence on the way participants evaluated the valence of the photos, and led to small but consistently calmer evaluations about photos. Perhaps the post-warning changes in negative affect and anxiety are evidence that participants devoted resources to emotional preparation or bracing (Shepperd, Findley-Klein, Kwavnick, Walker, & Perez, 2000). In fact, bracing often results in a negative anticipatory period characterized by high anxiety (Sweeny, Reynolds, Falkenstein, Andrews, & Dooley, 2015). Indeed, if the warning did activate emotional bracing, this might explain why the warning had no effect on participants' evaluations of neutral and no headline condition photos—these photos were neutral and therefore emotional preparation would not result in a significant reduction in negative evaluations. However, if trigger warnings were acting as a beneficial preparatory device, we should have seen a reduction in negative evaluations of photos with negative headlines over the course of photo viewing. However, warned participants rated negatively valenced photos no differently to unwarned participants—

despite that negative headline conditions were consistently rated as more negative overall than neutral or no headline conditions across multiple measures. Therefore, priming negative expectancies with a trigger warning prior to viewing negative stimuli resulted in an (un)pleasant state of anticipation with few subsequent benefits—a result supported by Sanson et al. (in press). Similarly, Golub, Gilbert, and Wilson (2009) found that while negative affect often increases in the anticipation of negative events, such “preparation” does little to attenuate negative affect in the face of the stressor. In fact, although Bellet et al. (2018) did not measure expectancies, their findings suggest that the warnings also have the potential to exacerbate negative responses.

However, there are other plausible interpretations of the findings. We know from the priming literature that new information is only likely to be assimilated into a primed category if the category is considered moderately extreme (e.g., a moderately hostile person such as “a boxer”; Herr, 1986). However, if the prime is an exemplar of an extreme category (e.g., Hitler, Stalin, etc.) participants contrast ambiguous stimuli to this category (e.g., rate ambiguous behavior as less hostile). Thus, because the warning was emotionally distressing, the ambiguous photos may have been contrasted with—rather than assimilated into—a negative category, resulting in positive evaluations.

Our own data support the idea of a contrast effect. In all studies, we asked participants to rate the experience of the photo stimuli in light of what they expected at the beginning of the study. The meta-analyses confirmed that warned (vs. unwarned) participants found the photos “somewhat more positive than expected”. In Study 1e, we found that directly before viewing the photos, warned participants expected the photos to be more negative than unwarned participants did. After photo viewing, warned participants rated the photos as significantly less negative than their initial rating, while unwarned participants actually rated the photos as significantly more negative than their initial rating. Moreover, unlike Bellet et al. (2018), who issued a warning immediately prior to each stimulus, here participants received a single overall warning. Thus, it is possible that participants in our experiments believed that the overall warning was not applicable to the neutral stimuli they viewed (e.g., perhaps they believed the negative stimuli would appear later on), and were subsequently relieved when the promise of negative content was never fulfilled, resulting in the small decreases in negative arousal we observed.

Whether a contrast effect is beneficial because it may make content appear more positive—and thus aligns with the goals of trigger warning advocates—is a question that still requires investigation. For instance, what would happen if the content was matched—rather than contrasted—to the expectation/primed category? Indeed, the warning message did elicit consistent levels of negative arousal across studies. Here it seems necessary to address several limitations within the present research: First we do not know if the negative affect caused by the warning would have a different impact on more negative content. But importantly, results from other studies suggest this outcome is likely (e.g., Bellet et al., 2018). Second, Bellet et al. found that participants who tended to believe that words could cause harm were particularly sensitive to the effects of warning messages. Because we did not assess participants' individual sensitivity levels, or beliefs about long term harm, we were not able to detect how these factors affected and interacted with the emotional effects of our warning message. Third, because trigger warnings were originally designed for trauma survivors, future research should explore how negative arousal created by warning messages may impact reactions towards *personally* distressing content or a diagnosis of PTSD. Moreover, here we employed the most common definition of trigger warning message, but future research should examine the effects of warnings in other various forms—for instance, more general warning messages, or warnings that give more or fewer details. Fourth, because we collected mood and anxiety data at three time points in the warning conditions (before and after the warning message and again after the stimuli), but only at two time points in the no warning conditions (pre and post stimuli), it is possible that the increases in negative arousal (e.g., pre to post warning measurements of negative mood and anxiety) were due to the fact that the warning acted as a notification about the task beginning. For example, the warning may have created arousal in response to thinking about the task participants were about to complete, rather than arousal from the warning alone. Moreover, this increase could have occurred because warned participants were asked to complete a second set of mood and anxiety questionnaires (e.g., regression towards the mean; Stigler, 1997). However, because we found that warned and unwarned participants did not increase from baseline to post-photo exposure, it seems likely that the increase observed from pre-warning to post-warning was caused by participants reading the message itself rather than seeing the set of mood and anxiety questionnaires again, or regression. Further, in Study 1e, warned participants did rate significantly higher negative expectations about the photos than unwarned participants. This pattern suggests that the increase in negative arousal observed after the warning was due to the warning creating a negative expectation about the task ahead, rather than due to the task more generally, which all participants read about in the consent form. Future research could address these issues more directly by collecting mood and anxiety data before and after a general set of instructions in the unwarned condition. Lastly, several studies have demonstrated that the quality of data from Mechanical Turk workers is as reliable (e.g., Buhrmester, Kwang, & Gosling, 2011; Casler, Bickel, & Hackett, 2013) or sometimes superior (e.g., fail less attention checks; Hauser & Schwarz, 2016) to participants sourced from traditional subject pools. Furthermore, research suggests that MTurk is an excellent source for studying clinical and subclinical populations: the prevalence of mental health disorders in MTurk populations has been found to match or exceed that of the general population, and clinical measures taken from MTurk participants demonstrate high reliability and validity (Shapiro, Chandler, & Mueller, 2013). Nevertheless, it would be useful to replicate this experimental paradigm in a traditional laboratory setting.

In sum, the present set of results demonstrate that trigger warnings are unlikely to create fearful or negative interpretations about material where none is warranted. Thus, the claims made by Lukianoff and Haidt (2015)—i.e., that warning messages would distort the interpretation of non-threatening material or exacerbate reactions towards negative content (e.g., information presented in lectures or news stories etc.) were unsubstantiated here. However, the warnings did create a negative anticipatory phase that did little to mitigate negative affect in our negative headline conditions. Thus, while the warning did not exacerbate negative reactions, it also did not appear to help people ‘mentally prepare’ to reduce the impact of content in any way (Lockhart, 2016). Overall, trigger warnings do not appear to be helpful when people are confronted with ambiguous material and in some instances, they may be harmful (e.g., Bellet et al., 2018). Future research should aim to disentangle expectancy disconfirmation and contrast effects from potential preparation effects.

4 Investigating the effects of trigger warnings on the reactions to and recall of negative memories

Chapter 4 is published as:

Bridgland, V. M. E., & Takarangi, M. K. T. (2021). Danger! Negative memories ahead: the effect of warnings on reactions to and recall of negative memories. *Memory*, 29(3), 319–329. https://doi.org/10.1080/09658211.2021.1892147

Author Contributions: I developed the study design with the guidance of MKTT. I collected the data, and performed the data analysis and interpretation, and drafted the manuscript. MKTT made critical revisions to the manuscript and approved the final version of the manuscript for submission.

Abstract

A trigger warning is an alert that upcoming material containing distressing themes might "trigger" the details and emotion associated with a *negative memory* to come to mind. Warnings supposedly prevent or minimize this distress. But, do warnings really have this effect? To simulate the experience described above, here, we examined whether warning participants—by telling them that recalling a negative event would be distressing—would change characteristics associated with the immediate and delayed recall of a negative event (such as phenomenology e.g., vividness, sense of reliving), compared to participants who we did not warn. Generally, we found that time helps to heal the "emotional wounds" associated with negative memories: negative characteristics—such as emotion, vividness etc.—faded over time. However, the event's emotional impact (the frequency of experiences related to the event such as "I had trouble staying asleep"), subsided less over a two-week delay for participants who were warned in the first session. Our findings suggest that warning

messages may prolong the negative characteristics associated with memories over time, rather than prepare people to recall a negative experience.

Introduction

A *trigger warning* is an alert that upcoming material (e.g., DC's 2019 film *Joker*) containing upsetting themes (e.g., graphic violence) might "trigger" intrusive memories about a related stressful event. "Triggered" memories can be very distressing, and trigger warnings supposedly help to prevent or minimize this distress (e.g., Friday, 2016). But do warnings actually have this effect? Although research has focused on the emotional effects of warnings when people encounter novel stimuli, we do not know whether warnings minimize the distress associated with bringing *a negative memory* to mind—the expected outcome of the "triggering" process. A worrying possibility is that warnings might distort negative memories in potentially harmful ways, for example by making memories *seem* more distressing. Indeed, we know that negative expectations—such as those that warning messages create—can cause or exacerbate negative reactions (i.e., the nocebo effect; Benedetti, Lanotte, Lopiano, & Colloca, 2007; Myers, Cairns & Singer, 1987). Moreover, the details and perceived impact of our personal memories—even very negative or traumatic memories—are, in general, highly susceptible to distortion (Pickrell, McDonald, Bernstein, & Loftus, 2017; Talarico & Rubin, 2003). Here, we sought empirical evidence for the idea that a warning would distort characteristics associated with the immediate and delayed recall of a negative event, including phenomenology (e.g., feeling like one is reliving the event), how central the event felt to people's identity, and its emotional impact (i.e., distressing symptoms). We also explored one potential mechanism that might help to explain how warnings help or harm; the coping strategies people use to cope with recalling the negative event.

In recent years, the topics potentially covered by trigger warnings have expanded far and wide (e.g., racism, blood, classism, pregnancy, etc.; LSA Inclusive Teaching Initiative, 2020), as have the range of emotional experiences such warnings are intended to help mitigate—from being mildly offended/distressed through to "re-traumatization" (Carter, 2015). The typical *popular* definition of a trigger warning is quite vague: an alert that upcoming material may be distressing. Prior work has focused on this definition, examining people's general emotional reactions when they encounter various types of *novel* stimuli, such as negative films (Sanson, Strange & Garry, 2019), images (Bridgland, Green, Oulton, & Takarangi, 2019), and text passages (Bellet, Jones, & McNally, 2018). From this research, we know that viewing a warning increases negative *anticipatory* reactions, but has little effect on subsequent *reactions,* towards potentially distressing material. Further, warnings do not seem to reduce distress among people with a trauma history, or for people who identify that study material (e.g., a description of a murder scene) reminds them of their most traumatic experience (Jones, Bellet, & McNally, 2020). Taken together, this initial work shows that general trigger warnings—which warn of upcoming distressing material—do not seem to help ameliorate negative reactions towards negative stimuli or stimuli that may have a connection to a stressful experience.

Here, however, we intended to investigate trigger warnings as they were *originally* defined—which has not yet been the subject of any empirical investigation. The term "trigger warning" originates from Posttraumatic Stress Disorder (PTSD) research showing that stimuli with characteristics similar to a traumatic event can "trigger" a person to re-experience the trauma (Ehlers, Hackmann, Michael, 2004). Re-experiencing symptoms include vivid thoughts, feelings and flashbacks about the event (Ehlers et al., 2004). Trigger warnings, therefore, were originally intended to mitigate the "triggering" process by alerting viewers that upcoming content may spark the recall of traumatic memories, specifically, not just that provocative or sensitive material may be encountered (Haslam, 2017). These ideas about the original purpose of trigger warnings are therefore central to the debate about the use of trigger warnings for people suffering from PTSD, and/or trauma survivors, and persist in informational materials disseminated today. For instance, The Innocent Lives Foundation (2020)—a source cited by social media influencers who use trigger warnings—claims that "memories for trauma are worse without warning" and that "trigger warnings are simple ways to help survivors avoid reliving the event."

Despite the prominence of these claims, no work has examined how trigger warnings may change how someone *remembers* a stressful/negative experience. Here, we aimed to simulate "triggering" the recall of a negative memory by specifically instructing participants to recall a negative event, and then examining whether warning participants about the potential for this process to be distressing would help (e.g., reduce distress) or harm (e.g., increase distress). To investigate one potential mechanism underpinning *why* a warning may change the ways in which a memory is recalled, we asked participants to report the strategies they used to cope with the negative event. The way we remember and relate to the past is critical for the maintenance of mental health and well-being (Adler & Pansky, 2019) and has implications for several clinical disorders (e.g., Posttraumatic Stress Disorder; (Oulton & Takarangi, 2017). Therefore, how warnings may change (or not change) how a negative event is recalled is central to assessing their use as an adaptive tool. We explore these ideas in more detail next.

How might a warning message affect the way that a negative event is initially recalled?

It is well established that setting up an expectation of negative *physical* health symptoms such as pain, itch, and other side effects can cause or exacerbate those very outcomes—known as the nocebo effect (e.g., Benedetti, et al., 2007). It is therefore possible that warnings may also affect psychological outcomes pertinent to *mental health*, such as exacerbating the emotional impact of a negative event. Indeed, we know that seeing a trigger warning leads to a noxious anticipatory period (Bridgland, et al., 2019) and that negative anticipatory information akin to warnings (e.g., that upcoming material is negative in nature) can enhance attention to negative stimuli, resulting in increased distress (Shafir & Sheppes, 2020). We also know that it is easier to recall memory details when someone is in the same emotional state as when the memory was encoded (Bower, 1987). Therefore, warning people about recalling a personal event may create a negative anticipatory period that, in turn, may change how a negative event is subsequently recalled.

The are several possible ways a warning might change the subsequent recall of a negative event. The warning might lead someone to retrieve an "*objectively" more negative* event (e.g., a Criterion A event in the DSM-5 involving actual or threatened death or serious injury; e.g., sexual assault, physical assault, loss of a loved one—although we do note that it is difficult and even controversial to define how objectively negative an event is, especially because people can have PTSD symptoms for events that do not meet Criterion A; Rubin & Feeling, 2013). Or, the warning may not change the event that is recalled but may enhance negative interpretations about the event. Either possibility should lead people to remember the negative event with more negative characteristics (such as emotional intensity, vividness), and to perceive that event as having greater emotional impact, and more centrality, compared to people who recall a negative event without a warning. Nevertheless, to increase the likelihood that participants would retrieve similarly negative events with and without a warning, and thus focus on participants' *interpretation* of those events, we constrained the recall period to events occurring during the past two weeks.

How might a warning change the way a negative experience is remembered over time?

While we know memories generally fade over time, we also know that external feedback about past events can change how we remember them. Typically, the details (Talarico & Rubin, 2003) and emotion (Walker & Skowronski, 2009) associated with negative events diminishes over time. Moreover, the mere act of thinking about and answering questions about a negative event on measures of memory characteristics (e.g., vividness, valence, sensory details etc.) can decrease negative reactions towards that memory (Rubin, Boals & Klein, 2010; Boals, Hathaway, & Rubin, 2011). Therefore, it is likely that participants will report an overall decrease in negative characteristics associated with their memory over the two-week period. However, it is possible that seeing a warning message at initial recall may reduce these general "healing effects" of time and warned participants may report a smaller reduction in negative memory characteristics.

Extant literature shows that exposure to misinformation about past events can change how we remember them (Loftus, 2005); including and perhaps even more so for negative events (e.g., Brainerd, Stein, Silveira, Rohenkohl, & Reyna, 2008). Warnings, therefore, may also affect how a negative memory is recalled over time. For instance, around 80% of military personnel who recently completed Survival School Training, endorsed misinformation for non-trivial event details such as the identity of their interrogator (Morgan, Southwick, Steffian, Hazlett, & Loftus, 2013). Importantly, however, false feedback can also change how we *feel* about past events. For example, participants who read reviews that a negative film was tolerable, reported fewer distress symptoms after a week than participants told the film was distressing, or neutral information (Takarangi, Segovia, Dawson, & Strange, 2014). Similarly, in Takarangi and Strange (2010), participants told their negative memory was *worse than* others' experiences reported greater stress, negative emotions, and vividness associated with the memory, a week later (vs. no feedback). Warnings could also distort memories for negative events over time by giving people more confidence that their memory was distressing and harmful or leading them to reconstruct their memory to align with negative appraisals. This process may also lead to an increase in the feeling that an event is central to one's identity—an outcome related to Post-traumatic Stress Disorder symptoms (PTSD; Berntsen, & Rubin, 2006).

Do trigger warnings change coping strategies?

To investigate a potential mechanism for the way trigger warnings may change the way a negative event is recalled, we also examined reported coping strategies. Unlike emotional reactions, coping strategies require an active effort to manage one's thoughts, emotions, and behaviors (Folkman & Moskowitz, 2004). Therefore, if warnings increase helpful coping strategies like proponents claim (McNiel, 2015; Palmer, 2017) we should find evidence that they are helping participants actively engage in strategies to assist in managing any distress associated with recalling the memory. For instance, a warning may remind someone to engage in emotional reappraisal (changing the way a situation is construed to decrease its emotional impact; Gross & John, 2003). Coping strategies may therefore help us understand how trigger warning messages may (or may not) affect the characteristics associated with the immediate and delayed recall of a negative event.

The present study

To investigate how warning messages may change how a negative event is initially recalled and remembered over time (e.g., emotional impact) and the strategies used to cope with the event, we asked participants to recall a recent negative event that had occurred in the past two weeks (Session 1); a fortnight later they recalled the same event again (Session 2). Prior to initial recall in Session 1, we randomly assigned participants to either view a warning message—informing them that the negative memory task was distressing—or an unwarned control condition. We had an additional exploratory aim; to examine if warnings might have accumulative effects (e.g., would a participant who was warned twice experience the smallest reduction in negative memory characteristics over time?). Although trigger warning messages are becoming increasingly prevalent in day-to-day life (e.g., on television, social media, in university etc.), no research has examined repeated exposure to warning messages across different experimental sessions. We therefore repeated our warning procedure in Session 2 (i.e., we randomized participants again to view or not view a warning message).[18]

In line with prior trigger warning research, we predicted that warned participants would experience a negative anticipatory period prior to completing the memory recall task (i.e., increases in negative mood and anxiety, and decreases in positive mood, from pre-to post-warning message). We hypothesized that in Session 1, participants given a trigger warning (vs. no warning) would report more negative memory characteristics (e.g., greater sense of reliving the event, greater emotional impact). Due to the healing nature of time, we predicted that participants' negative memory characteristics will likely diminish over the two-week delay. However, we predicted that this pattern will depend on whether participants received a trigger warning during Session 1 (i.e., an interaction between condition and time). Specifically, we anticipated that participants who received a warning in Session 1 would report a smaller decrease in negative characteristics over time (or possibly an increase in negative characteristics), compared to unwarned participants. We also anticipated that those who receive a warning in both Session 1 and Session 2 would report the smallest reduction (or largest increase) in negative responses over time due to the accumulated effect of the warning messages. Finally, it is possible that participants who were warned in Session 1 may have more negative mood and anxiety scores at the beginning of Session 2, due to anticipating feeling negative upon entering the testing room.

Method

The Flinders University Social and Behavioural Research Ethics Committee approved this experiment. Our preregistration, data, and supplementary files are located at: https://osf.io/dxnbp/. We have reported all measures, conditions, and data exclusions.

[18] These conditions were collapsed for our main analyses, but we report key findings here (below) and full results can be found at: https: https://osf.io/x6t7v/

Participants and Design

A total of 239 participants took part in Session 1. Of these, 24 did not return for Session 2 (8 = unwarned; 16 = warned), one had already completed the study previously, and one did not follow headphone instructions. Of the 213 participants who returned for Session 2, two failed to recall the same memory from Session 1, one did not follow headphone instructions, and, due to a technical error, one completed the wrong survey. Thus, 209 participants completed Sessions 1 and 2. Participants were predominantly female (80.9%), with an age range of 17-50 (*M* = 22.20, *SD* = 6.30); 45.9% were White/Caucasian/European, 23.4% were Asian, 11.5% other (unspecified, mixed-race, African, Middle Eastern, Hispanic), and 19.1% specified nationality ("Australian").

We departed from our pre-registered design and planned analyses.[19] We conducted a post-hoc sensitivity analysis to assess the power of our final sample (*n* = 209) for 2 (Session 1 warning condition: warned, unwarned) x 2 (Session Time: Session 1, Session 2) mixed ANOVA analyses. We found that our sample was adequate to reliably identify a small-medium effect size (*f* = 0.19) for an alpha level of 0.05, and a desired level of power = .80 (Faul, Erdfelder, Buchner, & Lang, 2009). Therefore, our sample size was adequate to detect our main interaction finding related to Impact of Event Scale Scores (small-medium η_p^2 = .036; in G*Power, *f*(U) = 0.19).

[19] We originally planned to analyse our dependenat variables using a 2 (Session 1 warning condition: warned, unwarned) x 2 (Session 2 warning condition: warned, unwarned) x 2 (Session Time: Session 1, Session 2) mixed design. We conducted an a priori power analysis for a 4 (between) x 2 (within) repeated measures ANOVA with the smallest effect size we would be interested in (*f* = .15), power of .95, and *r* = .48, based on a prior correlation between repeated measures of emotion about a recent negative event (Takarangi & Strange, 2010). The recommended sample size was 204. We calculated this power analysis because G*Power does not have the capability to calculate power for mixed designs beyond a single between subjects' level. However, a previous reviewer rightly pointed out that we were likely therefore underpowered for a 2 (between) x 2 (between) x 2 (within) subjects design. While we could have analyzed our variables using 4 (between) x 2 (within) subjects' analyses, we do not believe this analysis reflects the true nature of our design, because participants are only in two groups (warned or unwarned) in Session 1. Additionally, the repeated warning in Session 2 was a secondary interest. Therefore, we reframed our analyses to focus on the effects of the Session 1 warning condition and analyzed our variables using a 2 (Session 1 warning condition: warned, unwarned) x 2 (Session Time: Session 1, Session 2) design, collapsing the Session 2 warning condition. This change allowed us to reach suitable power. We report the analyses of the full 2x2x2 design here: https://osf.io/x6t7v/, and report any notable findings related to our secondary aim regarding the accumulative effects of the warning message for participants warned in Session 1 and Session 2 in our results section below.

Materials

Warning message. In the *warning present* conditions participants saw a warning message on screen and simultaneously heard it as audio (via headphones):

"Warning: This study involves recalling a negative personal experience. Some people find this process distressing. For example, you may experience negative mood and intrusive mental images. A small minority of people also experience distressing memories and reactions in the week after recalling negative events, although these reactions generally subside quickly. Please do not proceed if you do not want to take part in this task or think that you may be adversely affected by this task."

Participants warned in Session 2 also received this message, prefaced with: "*We wish to remind you."*

Recall task. In Session 1, we asked all participants to recall a negative event (see https://osf.io/2h6nw/ for full instructions; Appendix F) they had experienced in the past two weeks (Takarangi & Strange, 2010; see https:// https://osf.io/c6ubd/ for the full text responses with identifiable information redacted). In Session 2, participants recalled and wrote about this same event.

Positive Affect Negative Affect Schedule (PANAS; Watson, Clark & Tellegen, 1988; Appendix C). Participants rated how they felt in the current moment on 10 Positive Affect (e.g., present study: interested; α = .91-.93) and 10 Negative Affect (e.g., upset; α = .88-.91) items (1 = very slightly or not at all, 5 = extremely). Scores are summed for each subscale.

Six-Item short form of the State Scale of the Spielberger State-Trait Anxiety Inventory (STAI; Marteau & Bekker, 1992; Appendix D). Participants rated how they felt in the current moment (1= Not at all, 4 = Very much; present study: α = .82-.85) on three anxiety-present (e.g., "I am worried") and three anxiety-absent items (e.g., "I feel calm"; reverse scored). Scores for each item are summed to form a total state anxiety score.

Memory phenomenology (Appendix G). Participants rated their negative memory on a range of phenomenological characteristics. We selected items that would help us understand how a trigger warning may distort the way an autobiographical event is retained in memory and therefore if the memory would share more characteristics with a 'triggered' intrusive memory for trauma. Traumatic intrusions reportedly have a sense of "nowness" as if they are currently happening (captured by our questions relating to *reliving*, *vividness*, *emotional intensity* and *sensory details*; Hackmann, Ehlers, Speckens, & Clark, 2004), are highly accessible and are thought and talked about more than low-intensity memories (captured by our *accessibility* and *rehearsal* items; Berntsen, 1996), and are recalled as fragmented isolated details rather than a coherent narrative (captured by our *content* and *coherence* items; Talarico, & Rubin, 2003). Lastly, recent and highly emotional memories are more likely to be visualized via a person's own eyes (D'Argembeau, Comblain, & Van Der Linden, 2003), and memories recalled from a visual perspective matching how the event is stored in memory enhances believability (captured by our *imagination perspective* items; Marsh, Pezdek, and Lam, 2014).[20] To simplify the analyses, we combined and averaged items measuring related concepts based on categories: *reliving* (4-items based on Rubin, Deffer, and Umanath (2019): reliving the event, travelling back to the time it happened, and feeling the same emotions; present study; α = .75-.80); *imagination perspective* (4-items based on Rubin et al. (2019): believing the memory was real vs. imaginary, remembering the event vs. just knowing it happened, whether the memory has details specific to my life vs. general, seeing event from own eyes vs. outside observer; α = .56-.68); *vividness* (5-items based on Rubin et al. (2019): how vivid and clear is the memory, while remembering the

[20] When reconstructing events from autobiographical memory, a person's belief in the memory actually occurring (rather than being imaginary) is enhanced if the event is recalled from a visual perspective that matches how the event-related information is retained in memory (Marsh, et al., 2014). Recent memories are more likely to be recalled from a first-person rather than a third-person perspective—therefore when recalling a memory from the past-two weeks, someone would be more likely to believe that it has really occurred if it is recalled from an observer (first person) versus field (third-person) perspective (Marsh, et al., 2014).

event I can see/hear/smell/hear people talking; α = .66-.76); *content* (2-items based on Rubin et al. (2019): I know the setting/location of actions; α = .74-.76), time (2-items based on Sutin and Robins (2007): my memory for the day/hour the event took place is clear; α = .56-.65), *emotional intensity* (6-items based on Sutin and Robins (2007): while remembering the event/my emotions at the time were positive (reverse scored), while remembering the event/my emotions at the time were negative, while remembering the event my emotions I feel are intense, while remembering the event I had a physical reaction; α = .75-.77), *rehearsal* (3-items based on Rubin et al. (2003): the event has come to me out of the blue without trying to think of it, I have thought/talked about this event since it happened; α = .59-.74); *accessibility* (5-items based on Sutin and Robins (2007): e.g., this memory sprang to mind when I read the instructions; α = .74-.82; *coherence* (6-items based on Sutin and Robins (2007): e.g., my memory comes as a coherent story/in pieces(reverse scored)/in words, the order of actions/events is clear; α = .76-.78; see supplementary materials for full items: https://osf.io/kt8ap/). All items were rated on a 1-7 scale with higher scores indicating higher levels with one exception. We also asked about *sensory details* (5-items: does your memory contain sensory details? (yes/no) visual, auditory, olfactory, tactile, gustatory).

Centrality of Events Scale (CES; Berntsen, & Rubin, 2006; Appendix H). This 20-item questionnaire is designed to measure the centrality of a negative event for a person's identity and life story (i.e., a single factor that represents the extent a negative event is employed as a reference point for the organization of other mundane general life experiences and meaning). Participants rated items (e.g., "I feel that this event has become part of my identity") in relation to their negative memory (1 = totally disagree to 5 = totally agree; present study: α = .94-96). Scores are summed to form a total Centrality of Events score.

Correlations between CES and PTSD symptomology in the present study (assessed by the Impact of Events Scale) were *rs* = .52-.59, *p*s <.001.

Impact of Events Scale Revised (IES; Weiss, 2001; Appendix I). This 22-item questionnaire measures the emotional impact of stressful life events based on the DSM criteria for PTSD. Participants rated (0 = not at all, to 4 = extremely) how often they were distressed or bothered in the past seven days by a range of reactions (e.g., I had trouble staying asleep; present study; *α* = .94-.95). Scores are averaged and can be scored as a single factor, or as three subscales—avoidance, intrusions and hyperarousal.

Ways of Coping (Revised; Folkman & Lazarus, 1985; Appendix J). Participants rated the extent to which they engaged in a range of coping strategies for the negative event they recalled (e.g., [c]hanged or grew in a person in a good way; 0 = not used, to 3 = used a great deal) forming 8 subscales developed from a community sample measuring a range of stressful experiences (Folkman, Lazarus, Dunkel-Schetter, DeLongis, & Gruen, 1986): confrontive coping (6-items: present study; α = .65-.74), distancing (6-items: α = .65-.72), self-controlling (7-items: α = .57-.66), seeking social support (6-items: α = .78-.80), accepting responsibility (4-items: α = .67-.74), escape-avoidance (8-items: α = .77-.81), planning and problem-solving (6-items: α = .69-.76) and positive reappraisal (7-items: α = .78-.81). Items are summed to form each subscale.

Procedure

Session 1. Figure 4.1 depicts the procedure. The experiment (including all questionnaires etc.) was run using Qualtrics software (Provo, UT). We told participants we were interested in the relationship between autobiographical memory and personality. All participants were told that they would be asked to recall a negative autobiographical experience but were not told that this experience would be distressing. Following consent, participants completed initial measures of mood (PANAS) and state anxiety (STAI). We randomly allocated them to the warning or control (no warning) condition. Participants in the warning condition saw a warning message at this time, followed by demographic questions, and mood and anxiety measures a second time. The participants in the control condition only completed the demographic questions at this time. All participants completed the recall task and rated the phenomenological characteristics of their memory, followed by how central the memory felt to their identity (CES), coping strategies (WCS), and the emotional impact of the event (IES) in a randomized order.

Session 2. Participants returned two weeks after Session 1 at the same time (we allowed a 24-hour grace period before or after the scheduled return time—used by nine participants). The procedure was identical to Session 1 except participants recalled the same event that they recalled in Session 1. To address an exploratory aim about the possible accumulative effects of warning messages, we re-randomized participants again to either receive a second warning or no warning. We then fully debriefed and paid participants $25AUD (*n* = 98) or granted course credit (*n* = 111).

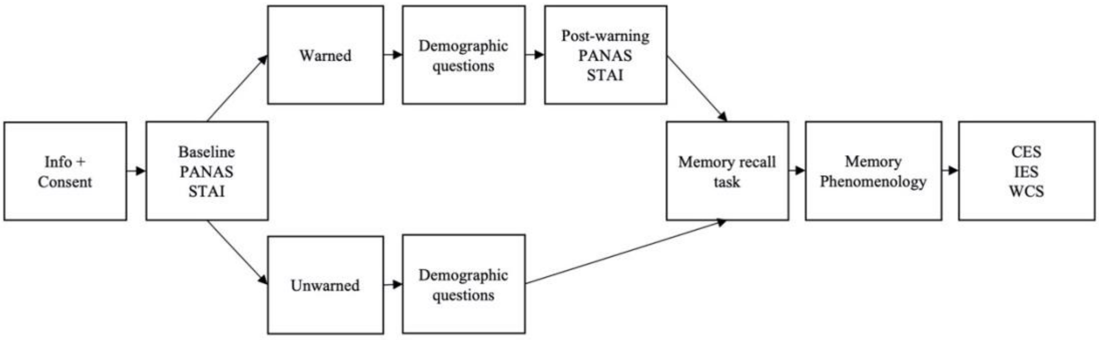


Figure 4.1. Chart depicting the procedure of Session 1 and Session 2

Results

Statistical Overview

Full descriptive and inferential statistics appear at: https://osf.io/7j5us/. Some measures were skewed and not normalized by transformations, so we have analyzed untransformed data. However, where variables violated homogeneity tests we ran analyses using transformed and untransformed scores and report changes in statistical patterns. For some measures, Box's Test of Equality of Covariance Matrices was significant, but because group sizes were similar we assumed Pillai's Trace to be stable (Field, 2005). All test statistics remained unchanged when corrected using Pillai's Trace.

To investigate our predictions that warned participants would experience a negative anticipatory period prior to completing the memory recall task (i.e., increases in negative mood and anxiety, and decreases in positive mood, from pre-to post-warning message), we conducted several paired samples t-tests; specifically, we compared PANAS mood measures and state anxiety measures (STAI) from pre- to post trigger warning presentation. For our main hypotheses that participants given a trigger warning (vs. no warning) would report more negative memory characteristics (e.g., greater sense of reliving the event, greater emotional impact) in Session 1, and that participants who received a warning in Session 1 would report a smaller decrease in negative characteristics over time (or possibly an increase in negative characteristics), compared to unwarned participants, we conducted several 2 (Session 1 warning condition: warned, unwarned) x 2 (Session Time: Session 1, Session 2) mixed ANOVA analyses.

Did the warning lead to a negative anticipatory period prior to the recall task?

We first confirmed that in *Session 1*, mood and anxiety ratings were not significantly different prior to randomization for participants in the warned and unwarned conditions. They were not (see Table 4.1; *ts* = 0.77-1.46, *ps* = .145-.442). We next compared mood and anxiety before and after the warning message in *Session 1*. In Session 1 the warning appeared to cause a negative anticipatory period: participants reported decreased positive affect (*t*(105) = 4.99, *p* <.001, d_z = 0.48, *95% CI* [0.28, 0.68]) and increased state anxiety (*t*(105) = -2.11, *p* = .037, d_z = -0.20, [-0.39, -0.01]) from pre- to post-warning message. However, participants reported similar negative affect from pre- to post-message (*t*(105) = 1.14, *p* = .259, d_z = 0.11, [-0.08, 0.30]).

In *Session 2*, we examined if participants' mood and anxiety scores prior to randomization into Session 2 warning conditions were influenced by their previous warning experience in Session 1. For instance, perhaps the previous feelings of anxiety and decreased positive affect returned to them when they were about to start the experiment at Session 2. However, the previous Session 1 warning did not seem to influence Session 2 anxiety prior to Session 2 condition randomization (*t*(207) = 0.94, *p* = .346, *d* = 0.13), positive affect (*t*(207) = -1.92, *p* = .056, *d* = 0.26), or negative affect (*t*(207) = -0.40, *p* = .687, *d* = 0.06). In sum, the warning message appeared to cause a negative anticipatory period prior to the recall task in Session 1.

Table 4.1

Summary of mean positive affect, negative affect and state anxiety ratings prior to randomization into warning conditions and pre- to post-warning message

Session 1 warning condition		*Warned* *(n = 106)*		*Unwarned* *(n = 103)*	
		M	*SD*	*M*	*SD*
Prior to randomization into warning conditions in Session 1	Positive affect	25.57	7.49	26.43	8.66
	Negative affect	16.18	6.07	15.50	5.90
	Anxiety	12.92	3.28	12.18	3.99
Post-warning Session 1	Positive affect	24.27	8.21	-	-
	Negative affect	15.89	6.27	-	-
	Anxiety	13.33	3.63	-	-
Prior to randomization into warning conditions in Session 2	Positive affect	23.47	8.24	25.66	8.20
	Negative affect	15.17	6.01	15.51	6.35
	Anxiety	12.38	3.86	11.86	4.00

Note: Positive Affect scale range 10-50, Negative Affect scale range 10-50, State anxiety scale range 6-24.

Characteristics associated with the memory

To examine the immediate effects of the warning message on memory recall (in Session 1) as well as how it may have affected the recall of the memory over time (in Session 2) we ran several 2 (Session 1 warning condition: warned, unwarned) x 2 (Session Time: Session 1, Session 2) mixed ANOVAs (full descriptive statistics in Table 4.2 and full inferential statistics Table 2 in the supplementary materials: https://osf.io/7j5us/). To investigate our predictions concerning the effects of the warning message on immediate and delayed recall, we applied a family-wise Holm-Bonferroni correction (for a total of four comparisons) for the results of each ANOVA to account for; 1) the main effect of Session 1 warning condition, 2) the interaction between Session 1 warning condition and Session Time, and any subsequent pairwise comparisons between 3) the effect of Session 1 warning condition in Session 1, and 4) the effect of session 1 warning condition in Session 2. Because we believed that time, as well as the act of completing the questionnaires would have an overall healing effect (a main effect of Time regardless of warning conditions) we did not include pairwise comparisons related to the change in each warning condition over Time in this correction.

Table 4.2

Summary of ANOVA results for 2 (Session 1 warning condition: warned, unwarned) x 2 (Session Time: Session 1, Session 2) mixed ANOVAs for memory characteristics and coping strategies

			Session 1				Session 2			
Session 1 warning condition		Scale range	Warned [21]		Unwarned		Warned		Unwarned	
			M	*SD*	*M*	*SD*	*M*	SD	*M*	*SD*
Phenomenology	Reliving	1-7	5.13	1.16	4.96	2.13	4.07	1.36	3.67	1.35
	Imagination perspective	1-7	5.94	0.95	5.99	0.79	5.28	1.21	5.16	1.15
	Vividness	1-7	4.59	1.10	4.53	0.98	3.98	1.24	3.65	1.08
	Content	1-7	6.29	1.02	6.16	0.81	5.82	1.29	5.57	1.14
	Time	1-7	5.27	1.39	5.26	1.27	4.25	1.52	4.19	1.43
	Emotional intensity	1-7	5.69	0.93	5.64	0.94	5.06	0.99	4.91	1.05
	Rehearsal	1-7	4.21	1.28	4.09	1.32	3.37	1.34	3.00	1.29
	Accessibility	1-7	5.71	1.08	5.55	1.26	4.90	1.50	4.80	1.30
	Coherence	1-7	5.02	1.13	4.99	0.98	4.18	1.23	3.84	1.18
CES		20-100	48.92	18.45	47.74	18.86	44.02	20.09	39.95	18.42
IES	Avoidance	0-4	1.58	0.89	1.52	0.87	1.28	0.94	1.05	0.87
	Intrusions	0-4	1.62	1.04	1.67	1.07	1.00	0.97	0.79	0.87
	Hyper-arousal	0-4	1.26	1.01	1.31	1.05	0.83	0.89	0.65	0.81
	Total	0-4	1.49	0.87	1.50	0.88	1.05	0.85	0.84	0.78
WCS	Confrontive coping	0-18	4.73	3.22	5.11	3.66	4.17	3.39	4.83	3.93
	Distancing	0-18	6.58	3.55	7.00	3.97	6.79	3.75	7.07	4.25
	Self-controlling	0-21	8.12	3.84	8.38	3.87	7.66	4.24	7.15	3.95
	Seeking social support	0-18	6.72	4.36	5.99	4.32	6.04	4.41	5.79	3.91
	Accepting responsibility	0-12	4.32	3.21	4.46	2.99	3.76	3.08	3.91	3.10
	Escape-avoidance	0-24	7.08	4.86	7.54	5.34	6.57	5.41	6.56	5.10
	Planful problem solving	0-18	6.18	3.57	6.96	3.98	5.81	3.74	6.53	4.10
	Positive reappraisal	0-21	4.40	3.58	4.50	4.62	4.23	3.96	4.56	4.67

[21] For one participant, Qualtrics failed to display the CES and some WCS items so some subscales could not be calculated (Self-controlling, Escape-avoidance, and Positive-reappraisal). Therefore, this participant was excluded from these analyses.

Healing effects of time

Before examining our main predictions in relation to the warning conditions, we first examined if there was an overall "healing" effect of time (i.e., decrease in negative reactions from Session 1 to Session 2) regardless of warning. As expected, the main effect of Session Time (Session 1, Session 2) was significant for all memory characteristics, ratings of event centrality, scores on the IES total (and subscales) reduced significantly over time (*Fs* = 46.61-198.34, *ps* <.001). Therefore, consistent with prior research, characteristics associated with recalling the negative events (e.g., phenomenological experiences, emotional impact, and event centrality) faded over time.

Did the warning change immediate recall experiences?

Despite the warning message creating a negative anticipatory period prior to memory recall in Session 1, no pattern of results shows support for the idea that the warning message made immediate recall experiences more negative. Additionally, the Session 1 warning did not seem to change recall experiences across the whole study, regardless of time—that is, there were also no main effects for Session 1 warning condition (*Fs* < 4, *ps* >.05).

Did the warning message distort delayed recall?

We next examined if the warning changed the way participants recalled the event over time. A significant interaction emerged between Session Time (1, 2) and Session 1 warning condition (warned or unwarned) for Impact of Event Total scores (*F*(1, 207) = 7.76, corrected *p* = .024, η_p^2 = .036). Follow-up simple effects tests revealed no significant differences in Session 1 for participants who were warned versus unwarned in Session 1 (*F* (1,207) = 0.93, corrected *p* >.999, *d* = -0.01, *95% CI* [-0.28,0.26]) or Session 2 (*F* (1,207) = 3.49, corrected *p* = .252, *d* = 0.26, *95% CI* [-0.01, 0.53]). The interaction appears to be driven instead by how the warning changed participants' scores over time. That is, IES total symptoms subsided *more* over time when participants were not warned in Session 1 (*F*(1,207) = 136.86, d_z =

1.12, [0.87, 1.36]) versus when they were warned (F(1,207) = 62.39, d_z = 0.79, [0.57, 1.02]). This finding suggests that the warning did indeed hamper the healing nature of time.

Contrary to predictions, there were no other interactions between Session 1 warning condition and Session Time for any other memory characteristics (Fs < 3, *ps* > .05). There were also no significant differences in the reporting of sensory characteristics (Y/N) for warned and unwarned participants in Session 1 or Session 2 (See https://osf.io/7j5us/). Thus, although warned—versus unwarned—participants did not experience more negative memory phenomenology, or perceive greater emotional impact, and event centrality *in Session 1*, consistent with our hypothesis, warning participants in Session 1 did result in a *smaller decrease* in some memory characteristics over time.

Coping strategies

The reported use of all coping strategy subscales reduced significantly over time (Fs = 4.34-17.28, *ps* <.001-.038), except for the distancing (F = 0.44, p > .05) and positive reappraisal (F = 0.08, p > .05) subscales. Likely as negative characteristics of the memory faded, so too did the need for coping behaviors. However, the warning had little impact on the coping strategies participants reported: we did not find any main effects for Session 1 warning condition or interactions between Session 1 warning condition and Session Time (*Fs* < 4, *ps* > .05).

Accumulative effects of warnings

Related to our second exploratory aim regarding the possible accumulative effects of warnings, we examined whether negative anticipatory effect of warnings accumulates over time as more people encounter an additional warning (see OSF for full descriptive statistics). Like Session 1, in Session 2 participants reported decreased positive affect (F(1,103) = 9.46, p = .003, η_p^2 = .084, d_z = 0.30, *95% CI* [0.10, 0.50]) and increased state anxiety (F(1,103) = 7.19, p = .009, η_p^2 = .065, d_z = -0.26, [-0.46, -0.07]) from pre- to post-warning message.

Again, participants' negative affect did not change (F< 1, p = .461, d_z = 0.07, [-0.12, 0.26]). There were no significant interactions between Session 1 warning condition and time (pre- vs. post-warning in Session 2; F = 0.31-1.62, 1, ps = .206-.581). Interestingly however, and in support of the idea that the negative effects of warning messages may accumulate, participants who were also warned in Session 1 reported lower overall positive affect than participants who were not also warned in Session 1. In other words we found a main effect of Session 1 warning condition among subjects who were also warned in Session 2 (F(1,103) = 7.08, p = .009, $\eta_p{}^2$ = .064). However, the main effect of Session 1 warning condition was not significant for negative affect (F (1,103) = 0.39, p = .531, $\eta_p{}^2$ = .004), or state anxiety (F(1,103) = 3.55, p = .062, $\eta_p{}^2$ = .033). We also found no results suggesting warnings had accumulative effects on any of our other measures (see https://osf.io/x6t7v/ for full details).

Discussion

Here, we investigated if it was possible for a warning message to distort the negative characteristics associated with the immediate and delayed recall of a negative event. We also explored whether warnings would change the strategies people used to cope with recalling the event. While the warning message caused a negative anticipatory period prior to the recall task, it did not change the way that people initially recalled their negative event (i.e., in Session 1). However, the warning message did appear to distort delayed recall experiences and hamper some of the healing effects of time. Ratings of event impact subsided less over the two-week delay for participants who heard a warning message in Session 1. Importantly, we did not find any evidence that warning messages were helpful in reducing negative emotional reactions or promoting the use of coping strategies.

Consistent with prior work (e.g., Bridgland et al., 2019), viewing the warning message in Session 1 (and in Session 2) led to a negative anticipatory period marked by increases in state anxiety and decreases in positive affect prior to the recall task. While we did not find an increase in negative affect from pre- to post-warning message, it is not uncommon for positive and negative affect to fluctuate independently (Crawford & Henry, 2004). Indeed, in Bridgland et al., we observed increases in negative affect, and no significant changes in positive affect, from pre- to post-warning message. The differences may be explained by the differences in study stimuli. In Bridgland et al., participants were warned prior to viewing a series of potentially distressing photographs. The aversive state provoked by this kind of message may be more related to fear of the unknown and thus better measured by negative affect, which is associated with states such as fear and nervousness (Watson, et al., 1988). In the present study, participants were asked to recall a past experience and thus were not faced with the unknown. However, lower positive affect is associated with feelings of *sadness* (Watson, et al., 1988)—a feeling that might be likely when recalling a negative past event.

Although the warning did not have any immediate effects, differences emerged after the two-week delay, suggesting that receiving the warning message in Session 1 had impact *that was only observable over time*. Our data fit with the idea that the negative anticipatory period became associated with *the act of recalling* the negative memory, therefore affecting delayed but not immediate recall. This possibility seems especially likely because we found evidence that participants who were warned in Session 1 experienced lowered positive affect throughout Session 2—even prior to the Session 2 tasks (e.g., while waiting for the experiment to begin). Thus, perhaps these participants were already anticipating the negativity of the recall task. However, we acknowledge that a limitation of this interpretation is that it was not possible to obtain a true baseline measurement of mood. Therefore, it is possible participants in the warning in Session 1 condition were in a more negative mood by random chance at the beginning of Session 2 due to natural variation (e.g., feeling negative due to other factors unrelated to the experiment).

A possible reason why warnings do not ameliorate negative affect may be because they do not appear to enhance the use of coping strategies to cope with negative events. Despite theorizing that warnings may increase avoidance behaviors, and despite claims that warnings help people to use coping strategies, we found no evidence for this idea. Moreover, no participants decided to exit the study at the point of viewing the warning message—showing the warning did not seem to promote complete situation selection avoidance behaviors. However, twice as many participants we warned in Session 1 (*n* = 16; unwarned, *n* = 8) did not return for Session 2. Perhaps these participants did maintain a higher level of IES symptoms regarding their negative event and thus did not wish to take part in Session 2 and have to recall and answer questions about their negative event again.

The data also suggest that warning messages could be considered a source of misinformation/feedback (e.g., Takarangi & Strange, 2010), and are capable of distorting how people perceive memories after a delay. In addition, our findings make a novel contribution to the nocebo literature by showing that anticipatory information may manifest as distress associated with memories over time. These findings are important because no published research has examined the effects of warning messages beyond a single experimental session.

There are several limitations. First, event impact ratings (measured by the IES) were quite low—meaning the effect of the warning message on these ratings was also small—likely because we asked participants to recall a negative event that occurred within the past two weeks. However, it is worth noting that even over this constrained period, 13.6% of our sample (13.2% unwarned in Session 1 and 14.0% warned in Session 1) reported an event that might be classified as Criterion A (actual or threatened death or serious injury; e.g., sexual assault, physical assault, loss of a loved one). Given that around 90% of people have experienced at least one lifetime traumatic event (Kilpatrick, et al., 2013), it is likely that the effects observed here may be magnified when targeting lifetime traumatic events or populations with clinical levels of PTSD.

Second, many of the effects we observed were small. However, while small effects may not be very consequential in a single episode, they may matter in the long run (Funder & Ozer, 2019). This consideration may be especially important for warnings that are becoming increasingly prevalent in everyday life. Consider one setting: an average adult spends three hours and 30 minutes per day on a mobile device (Molla, 2020), equating to 53 full days a year, viewing thousands of online posts and articles, a proportion of which contain trigger warning messages. Over time, small negative effects caused by warning messages, such as anticipatory anxiety (Bridgland et al., 2019), enhanced event centrality (Jones et al., 2020), and memory distortion, may accumulate and have large consequences. Previous work on warning messages has only used single measurement designs and focused on the short-term effects. Our results highlight that although warnings do not always have immediately observable effects, warnings may change emotional responses over time. Indeed, if we had obtained measurements from a third time point, a month after the initial session, we may have observed further effects. Lastly, it is possible the wording of the warning message itself (i.e., "*negative mood and intrusive mental images*") may have related most strongly to the intrusion and hyperarousal scales of the IES. This feature of the warning may explain why the warning inhibited "healing" over time for the IES but not for other measures. Therefore, it is necessary for future research to examine warnings that emphasize different negative outcomes and use different wording.

Third, because we did not obtain a second measure of mood and state anxiety in the no warning condition, it is possible that the decreases in positive affect we observed in the warned condition from pre to post-warning reflect a general decrease over time—perhaps due to a natural decrease in positive arousal due to sitting in a laboratory room—rather than attributable to the warning message itself. However, because the warning is only 40 seconds in length and participants completed only three demographic questions before the second measure of mood and anxiety, it seems unlikely that participants' positive affect would have deteriorated much in such a short lapse of time. Furthermore, this explanation does not account for the increase in state anxiety also reported by warned participants from pre- to post-warning message—suggesting that the warning message did cause some levels of genuine negative affect. Nevertheless, future research should consider this limitation, perhaps by providing neutral instructions matched to the length of the warning message in the unwarned condition.

Fourth, the Cronbach's alpha for some of the memory phenomenology subscales were low, suggesting poor internal consistency. This pattern is perhaps because we assembled our own set of items from several memory questionnaires—as is customary for research using items from the Autobiographical Memory Questionnaire—and therefore the questionnaire does not have a validated factor structure. In future, it would be beneficial to validate the factor structure of our questions prior to conducting any follow-up experiments.

In summary, this study is the first to examine the effects of warning messages on the recall of personal memories (rather than novel stimuli) with two important findings: first, we found that warning messages seem capable of prolonging aversive aspects of a negative event. Second, if we turn to what we *did not* find, warnings do not seem to diminish the distress associated with recalling a negative memory or increase the reported use of coping strategies. These data have important implications for renewed calls to use trigger warnings to improve mental health by adding to the growing body of evidence that trigger warnings at best may have trivial effects or at worst cause harm. Further, our results have implications for trauma researchers and clinicians who use warnings as part of informed consent procedures. In a sample of 180 ProQuest dissertations that contained one or more of nine trauma related terms (e.g., disaster), over one third of the consent documents suggested participation would be moderately to *severely* distressing (Abu-Rus et al., 2018). Further, recommended practice for exposure therapy is to make patients aware of possible risks (e.g., distress and temporary symptom exacerbation when repeatedly recalling the traumatic memory; Altis, Elwood, & Olatunji, 2014). However, our results suggest that by setting up the expectation of risk, this consent ritual may actually be a source of harm (Loftus & Teitcher, 2018).

5 The effect of trigger warnings on bringing coping strategies to mind

Manuscript under review at *Journal of Behavior Therapy and Experimental Psychiatry* as:

Bridgland, V. M. E., Barnard, J. F., & Takarangi, M. K. T. (2021). Unprepared: Thinking of a trigger warning does not prompt preparation for trauma-related content.

Author Contributions: I developed the study design with the guidance of MKTT. JFB and I collected the data. JFB and I and analyzed and interpreted the data and she reported a subset of the data in her honours thesis. I analyzed and interpreted the remaining data and independently wrote a complete draft of the paper. MKTT provided critical revisions. All authors approved the final version of the manuscript for submission.

Abstract

Trigger warnings have been described as helpful—enabling people to "emotionally prepare" for upcoming trauma-related material via "coping strategies." However, no research has asked people what they *think* they would do when they come across a warning—an essential first step in providing evidence that trigger warnings are helpful. Here, participants from Amazon's Mechanical Turk (n = 260) completed one of two future thinking scenarios; we asked half to think about coming across a warning related to their most stressful/traumatic experience; the others thought about the actual content (but no warning) related to their most stressful/traumatic experience. The warning condition did not produce differences in coping strategies, state anxiety, or phenomenology (e.g., vividness, valence) relative to the content condition. Only one key difference emerged: participants who imagined encountering a warning used fewer positive words, when describing how they would react. One potential explanation for the consistent finding in the literature that trigger warnings fail to ameliorate negative emotional reactions is that these warnings may not help people bring coping

strategies to mind. Although, further empirical work is necessary to fully substantiate this potential interpretation

Introduction

Trigger warnings are alerts about upcoming content that may contain themes related to traumatic experiences (Bridgland, Green, Oulton, & Takarangi, 2019). Advocates claim that warnings help people to emotionally prepare, use coping strategies, or avoid distressing material (DeBonis, 2019; Lockhart, 2015). But recent evidence shows trigger warnings, in their current form, do little to ameliorate emotional reactions (e.g., Bridgland, Green, Oulton, & Takarangi, 2019). Therefore, advocates likely call for trigger warnings because they *believe* warnings will be helpful. Yet, when provided with a warning, they may not know *how* to receive its alleged benefits. One way that warnings might prepare people to face potentially distressing content is to prompt them to bring to mind and then enact helpful coping strategies. Of course, the first step is essential: people must be able to *bring existing coping strategies to mind* before they can use them. Thus, here we sought evidence that warnings prompt people to bring existing coping strategies to mind. We asked one group of participants to report *what they would do* if they came across a trigger warning and another group of participants to report what they would do if they came across content (i.e., with no warning) related to their most stressful/traumatic experience (e.g., in the news, in a lecture etc.). We then measured the coping strategies that participants brought to mind and thought they would use. To align with previous research, we also measured participants' emotional reactions to their imagined scenarios, and to capture our sample's underlying belief in the efficacy of trigger warnings, we asked participants if they *believed* trigger warnings would be helpful in reducing distress.

Prior research has asked people to describe *how* trigger warnings are helpful. Common responses reflect a belief that warnings help people to "prepare" for distressing material (Bentley, 2017; Cares, Franklin, Fisher, & Bostaph, 2017; DeBonis, 2019; George & Hovey, 2019). This belief does not fit with emerging empirical evidence, showing that viewing a trigger warning can increase anticipatory anxiety (e.g., Bridgland, et al., 2019) but has little impact on subsequent emotional reactions to distressing material (e.g., Sanson et al., 2019). Yet, limited research has focused on explaining *why* warnings do not ameliorate emotional reactions. To do so, we must take a closer look at the vague concept of "preparing"—to "prepare" is defined as "mak[ing] (someone) ready or able to do or deal with something" (Oxford Languages, 2021). While there may be many ways to examine the concept of "preparation," one way to operationalize preparing in a trigger warning context is to examine bringing *coping strategies*—a conscious effort to manage the demands of a stressful situation using thoughts and behaviors (Folkman & Moskowitz, 2004)—to mind.

Coping strategies are generally classified along four dimensions: whether they focus on managing thoughts and emotions, versus behavioral actions, and whether they are approach (e.g., focused on the stressor itself) versus avoidance based (e.g., avoiding the stressor; Littleton, Horsley, John, & Nelson, 2007). Of course, people need to be able to bring existing coping strategies to mind first to actually use them—however no research has investigated if trigger warnings are a useful tool in prompting coping strategies to come to mind. The available research on trigger warnings has only considered *behavioral avoidance* of experimental stimuli, finding no preference for film (Gainsburg & Earl, 2018) and newspaper (Bruce & Roberts, 2020) titles with versus without warnings.

Behavioral avoidance is only one potential method of coping when someone encounters a trigger warning; the other is to engage with the content. It might be tempting to align these two courses of action with avoidance-based or approach-based coping. However, approach-based coping requires an *active* effort to directly address a problem causing distress behaviorally (e.g., learning more about the stressor) or cognitively (e.g., reappraising the way a situation is construed to decrease emotional impact; Littleton et al., 2007). Thus, viewing trauma-related content could sometimes constitute *avoidance*, if a person tries to avoid their emotions, reactions, or parts of the material they consider distressing. It is also possible that someone might use complete behavioral avoidance (e.g., leaving a lecture/turning off TV), to enable a different approach strategy later (e.g., learn more about the class material at home).

Taken together, past research shows that people who ask for trigger warnings believe trigger warnings help people to “prepare,” yet trigger warnings do not seem to be effective in reducing negative reactions or promoting avoidance. But no research has investigated *why*. One possibility is that trigger warnings change—or do not change—how someone brings existing coping strategies to mind. Here we randomly assigned participants to a future thinking scenario: where they either encountered a *trigger warning* (warning-only condition), or *content* (content-only condition; between subjects), related to their most stressful/traumatic experience. We did not instruct participants in the warning condition to think about the content following the warning. Since we draw on past experiences to generate hypothetical future experiences (Schacter & Addis, 2007), and previous research has employed mental simulation exercises in order to investigate trauma memory (e.g., Newton & Hobbs, 2015), a future thinking scenario provides an interesting medium to examine how participants would respond to the scenarios in “real life,” without having to present them with traumatic content. Our first key aim was to examine the coping strategies that people bring to mind when they think about a trigger warning versus those they bring to mind when they imagine viewing distressing content/material. More specifically, as a first step to address this aim, we examined the number and type (e.g., approach vs. avoid, reappraisal vs. suppression) of coping strategies participants reported. Assessing the efficacy of these coping strategies was beyond the scope of the present investigation and experimental design. To align with previous research, our second key aim was to examine if imagining encountering a warning (versus content) would help ameliorate negative emotional reactions—operationalized as state anxiety and phenomenological characteristics such as vividness, intensity etc. Our third key aim was to examine to what extent people believed trigger warnings would be helpful in reducing distress. Finally, as an exploratory aim, because trigger warnings were originally intended for use by people suffering from PTSD (Haslam, 2017), we examined differences in our pattern of results for people who are likely PTSD-positive (vs. negative).

Method

The Flinders University Social and Behavioral Research Ethics Committee approved this experiment. We preregistered this experiment (osf.io/cqtzw/) and the data and supplementary material can be found here: osf.io/7n85z/. We made changes to prevent bots/farmers completing the study (i.e., a captcha and English proficiency test), screened existing data (see below), and updated the registration (osf.io/szaw8/) after issues were identified on Amazon's Mechanical Turk during data collection (Bai, 2018). We have reported all measures, conditions, and data exclusions.

Participants

Previous research has not investigated the effects of trigger warnings on coping strategies. Therefore, we estimated sample size based on the weighted effect size (*d* = 0.35) from a meta-analysis of the impact of warnings on state anxiety (Bridgland, et al., 2019). An a priori power analysis for a two-tailed, independent samples t-test (using G*Power; Faul, Erdfelder, Lang, & Buchner, 2007) with an alpha of .05, power of .80, resulted in a target sample size of 260 participants. We recruited 336 participants through Mechanical Turk. Participants were limited to people over the age of 18 who were proficient in English and resided in the United States. Thirty-five were identified as likely 'bot' respondents and excluded. The remaining 301 participants received a payment of $3.00 USD. We excluded a further 10 participants who failed all three embedded attention checks (Berinsky, Margolis & Sances, 2014; Hauser & Schwarz, 2015), four who completed the survey twice, and 27 who did not meet the criteria for a 'useable' response to the future thinking scenario. Participants should have mentioned *at least one* of the following: 1) the place they were imagining being in/seeing the warning or content; 2) that they saw something related to their event (either warning or content); 3) how they felt/what they would have done. The sample were predominately female (58.1%), and Caucasian/White (81.92%; 8.46% African American; 5% Asian; 4.6% other), with a mean age of 36.57 (range: 19-66, *SD* = 10.77).

Materials

Trauma history screen (THS; Carlson et al., 2011; Appendix K). Participants responded Yes/No (and how many times) to a list of 14 High Magnitude Stressor events (sudden events that have been found to cause most people extreme distress; e.g., a really bad transport accident). Participants then indicated if any of the events bothered them emotionally, and, if so, were prompted to describe the event that bothered them the most. If the event did not bother them emotionally, or they had not experienced any of the events, they were asked to describe the most stressful experience of their life. Participants then provided: their age at the time of their most traumatic/stressful event; whether anyone was hurt or killed (Yes/No); whether they felt afraid, helpless or horrified (Yes/No); how long they were bothered by the event (1 = *not at all*, 4 = *a month or more*); and how much the event bothered them emotionally (1 = *not at all*, 5 = *very much*). We told participants they would refer back to their identified event in subsequent survey questions.

Short-form Spielberger State-Trait Anxiety Inventory (STAI-6; Marteau & Bekker, 1992; Appendix D). Participants rated how they felt at that current moment for three anxiety-present items (e.g., "*I am worried*") and three anxiety-absent items (e.g., "*I feel calm*";1 = *not at all*, 4 = *very much*; (present study α = .88-.90).

Posttraumatic Stress Disorder Checklist (PCL-5; Bovin et al., 2016; Appendix L). Participants indicated how bothered they were by a list of symptoms over the past month (e.g., repeated, disturbing dreams of the stressful experience; 0 = *not at all*, to 4 = *extremely*) in relation to their most stressful/traumatic experience. Questions correspond to the DSM-5 symptom criteria for PTSD (present study $\alpha = .95$).

Future thinking scenario and question. Participants were asked to write about the following:[22] *"Imagine you are performing everyday tasks, in a familiar place, with familiar people—for instance, watching a lecture for your degree, reading the news or viewing a news report on television, watching a television show or reading social media posts etc.**—and come across a warning that informs you the content you are about to view might be distressing or triggering to people who have suffered traumatic experiences. Imagine that this warning also explicitly mentions the subject of your own traumatic or most stressful experience (that you reported earlier). /—and come across content that explicitly mentions the subject of your own traumatic or most stressful experience (that you reported earlier).** Using the box below, giving as much detail as possible, please describe what this scenario might be like, step by step, starting from the beginning where you see the **warning/content** (e.g., television, social media, lecture presentation etc.) and what it might **say/be**, to what would happen immediately after (e.g., how you would react and what you would do). Give a step by step account of what you would do in this situation, noting how you would feel at each point."*

Open response coping question. To capture coping strategies without prompting from questionnaires, we asked: "*In the scenario you read and wrote about, what coping strategies or techniques would you use? (e.g., any ways you might try and manage your reactions or respond to the situation). Please describe them*."

Modified Autobiographical Memory Questionnaire (D'Argembeau & Linden, 2006; Appendix M). Participants rated the subjective experience of their imagined event on

[22] Participants in the 'warning-only condition' saw the bolded text before the forward slash, while participants in the 'content-only condition' saw the bolded text after the forward slash.

12 indices: autonoetic consciousness (e.g., feeling as if one is experiencing the event), visual details, other sensory details, spatial context, temporal information, feeling emotions, intensity, valence, personal importance, in words, coherent story, and visual perspective, and vividness (1 = *not at all*, 7 = *completely*). We also included questions relating to anxiousness/worry about the expected outcome of the event, if participants expected a good/bad outcome, and how difficult it would be to cope (1 = *not at all*, 9 = *extremely*; Jing, Madore, & Schacter, 2016).

Coping Response Inventory (CRI; Moos, 1993; Appendix N). The CRI asks people to indicate how often they used approach and avoidance coping for a past stressful situation. We modified the instructions to ask participants how likely they would be to use the strategies in the scenario and used rating scales from the Ways of Coping Questionnaire (0 = *would not use,* 3 = *would use a great deal*; N/A = *Not Applicable*; Folkman & Lazarus, 1985). Despite these changes, scale reliability was similar to the original (approach scales: present study *α* = .67- .77, Moos, 1993 *α* = .64-74, avoidance scales: present study *α* = .64-.78, Moos, 1993 *α* = .58-.72).

Emotion Regulation Questionnaire (ERQ; Gross & John, 2003; Appendix O). We modified the instructions to ask how participants would use emotion regulation strategies in the scenario, rather than generally. Participants rated six items (1 = *strongly agree*, 7 = *strongly disagree*) relating to reappraisal (e.g., "*If I wanted to feel less negative emotion, I would change what I was thinking about*"; *α* = .89) and four relating to suppression (e.g., "*I would control my emotions by not expressing them*"; *α* =.83).

Questions regarding trigger warnings. We asked participants in the warning condition (Yes/No checkbox and an open textbox): (1) "*Do you think that this kind of warning would prevent you from being emotionally affected or triggered later on when viewing the material (versus if a warning had not been issued first)?*", and (2) "*Would this* *reminder of your trauma (in the form of a warning) make you react differently to if you just saw content related to your trauma itself? (i.e., might you be triggered by the warning itself?)*".[23] We asked participants in the content group "*Do you think that a warning before seeing content (like that in the previous scenario) would prevent you from being emotionally affected or triggered later on when viewing the material (versus if a warning had not been issued first)?*". Finally, all participants were asked, "*What do you think would be the best way to help you cope with trauma 'triggers' in everyday life?*".

Procedure

We told participants we were interested in studying feelings and beliefs about different types of traumatic experiences. After consent, participants completed demographic information, rated current anxiety (STAI), and traumatic event exposure (THS). Participants then rated how central their identified event felt to their identity using the Centrality of Events Scale (CES-7-item; Berntsen, & Rubin, 2006) and PTSD related symptomology (PCL-5), in random order, followed by their current anxiety. Next, participants were randomly assigned to complete one of the two future thinking scenarios (warning-only, content-only). Participants then rated their current anxiety, answered the event outcome questions, and rated characteristics of the imagined scenarios (AMQ). Next, participants completed the open response coping question, identified coping strategies (CRI) and emotional reappraisal (ERQ; in random order), completed the CES and the PCL-5[24] for a

[23] A colleague noted the phrasing of question (2) may have been confusing to participants. We therefore checked Y/N answers against text responses. We amended responses so that 'yes' responses included people who generally believed warnings were helpful (e.g., would be less distressing/triggering than seeing content), and 'no' responses were people who generally believed warnings were not helpful (e.g., they would be just as distressed/triggered by seeing a warning as seeing content). 10% of participants changed from a 'no' to a 'yes', and 13.1% of participants changed from a 'yes' to a 'no'. Where text responses were ambiguous or missing, we retained original responses.

[24] These data relate to a secondary interest: appraisals of past emotional experiences are influenced and often based on appraisals of current emotions (e.g., Levine, Prohaska, Burgess, Rice, & Laulhere, 2001). We were therefore interested in exploring if perceptions of event centrality and PTSD symptoms might change from pre to post scenario depending on the future thinking condition, given that research has shown that trigger warnings can change perceptions of event centrality (Jones, Bellet, & McNally, 2019). See: https://osf.io/e2xcq/

second time (in random order), and questions regarding trigger warnings. Finally, participants were asked if they left the survey (if yes, for how long), and if they had any technical problems.

Results

Statistical Overview

Where variables did not meet the assumption of normal distribution, we ran analyses using transformed and untransformed scores. In all analyses, the pattern of results did not differ and therefore, we report untransformed scores. Patterns remain unchanged by Holm-Bonferroni corrections, so we present uncorrected data unless specified. We initially ran analyses using Null-Hypothesis Significance Tests but also report Bayes Factors (BF_{01}), evidence for the null hypothesis [strong: BF_{01} = 10 – 30, substantial: BF_{01} = 3 – 10, anecdotal: BF_{01} = 1 – 3], no evidence [BF_{01} = 1], and evidence for the hypothesis [anecdotal: BF_{01} = .3 – 1, substantial: BF_{01} = .1 - .3, strong: BF_{01} = .03 - .1]; Jeffreys, 1961). The prior is described by a Cauchy distribution centered around zero and with a width parameter of 0.707. This distribution corresponds to a probability of 80% that the effect size lies between -2 and 2.

Coping strategies

We turn to our first key aim: to examine the coping strategies that people bring to mind when they imagine coming across a trigger warning or content related to their most stressful/traumatic experience.

Qualitative Responses. Two researchers coded responses to the future thinking scenario and the open response coping question according to the two broad approach and avoidance categories of coping Littleton et al. (2007) describe, and categories from the CRI (1 = yes, 0 = no; see Table 5.1 and Figure 5.1; see https://osf.io/cjz2a/ for instructions). Responses were coded according to the *active* use of approach or avoidance-based coping wherever mentioned. For instance, a participant who mentioned they would avoid content with a warning message, but would then seek social support, was coded as using both an avoidance and approach strategy. Agreement between coders was good (77.31%-86.15%). Coders met to resolve discrepancies. Where agreement could not be reached, a third coder resolved differences. Responses differed depending on condition, making it impossible to completely blind coders to condition. Thus, we asked a fourth coder—unaware of the study aims and of participant condition—to code the data. This coder had 90% congruence with the original coding and the pattern of findings remained unchanged.

Table 5.1

Examples of qualitative response coding for the future thinking scenario and open response coping question by future thinking scenario

	Future thinking scenario		Open response coping question	
	Trigger warning-only condition	Content-only condition	Trigger warning-only condition	Content-only condition
Evidence of approach coping	"*…I would read the warning but I would still watch the program and see if I could learn something to help with what I am going through*."	*"I would feel curious when they first started talking about it. I would listen to gain a better understanding of the subject."*	*"Talking to friends and family. Remembering the good things."*	*"Training my mind to focus on the present, and to think positively about my current life."*
Evidence of avoidance coping	"*Turn it off / leave lecture...I do not want to be reminded of it*."	*"I would change the channel very quickly and try my best to push the memories out of my head."*	*"I would immediately avoid the situation entirely."*	*"I would use avoidance to deal with the event. I would think about something else so that I didn't feel bad and didn't feel all my memories."*

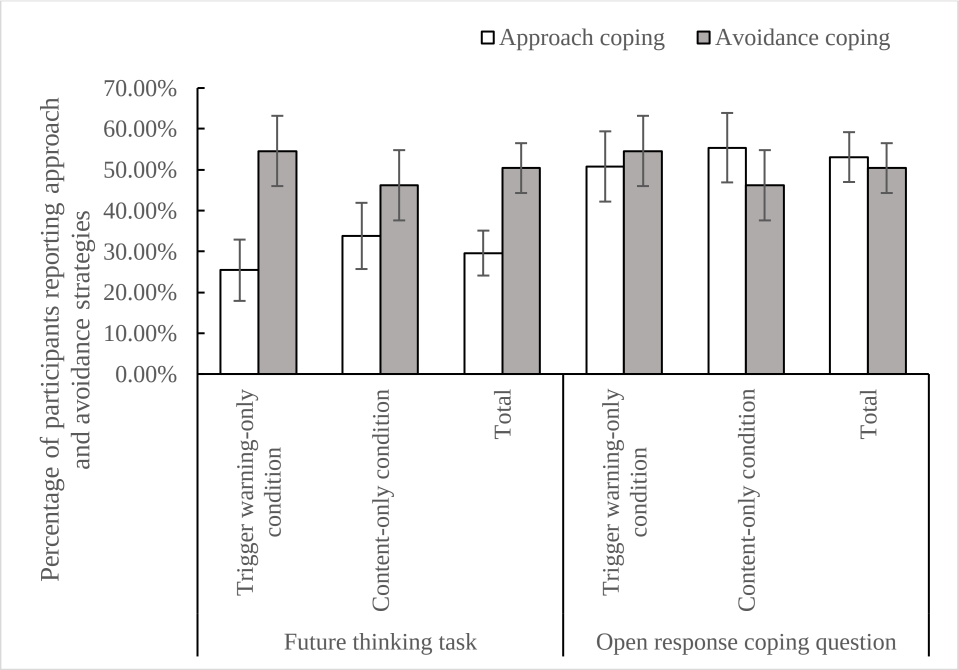


Figure 5.1. Percentages of approach and avoidance coping strategies by future thinking scenario.

We conducted Chi-square tests to compare the proportion of participants in each condition whose responses showed evidence of generating each coping strategy category.

Overall, participants indicated they would be more likely to use an avoidance-based strategy (50.4%) than an approach-based strategy (29.60%). Contrary to claims that trigger warnings help people "prepare," participants who imagined coming across a trigger warning-only brought to mind a similar percentage of approach (future thinking scenario: $\chi^2(1) = 2.23$, $p = .135$, $\Phi = .09$, open response coping question: $\chi^2(1) = .56$, $p = .456$, $\Phi = .05$) and avoidance (Future thinking scenario: $\chi^2(1) = 1.82$, $p = .172$, $\Phi = .08$, open response coping question: $\chi^2(1) = 1.86$, $p = .172$, $\Phi = .08$) coping strategies versus those imagining content-only.

Questionnaire-assessed strategies. We next examined the coping strategies participants said they would enact in the scenarios (CRI scored using the Moos (1993) protocol [https://osf.io/8df4s/] and cognitive regulation strategies using the ERQ).

We ran a series of independent samples *t*-tests comparing scores on the CRI's avoidance and approach coping scales and the ERQ's suppression and reappraisal scales, for participants in the warning-only and content-only conditions. Aligning with the qualitative data, we found no significant differences between the conditions (*Fs* = 0.03-1.03) and substantial evidence for the null hypothesis (BFs_{01} = 4.46-7.25; Figures 2a-2b). Therefore, imagining encountering a trigger warning-only does not seem to prompt someone to select more coping strategies from a given list compared to content-only.

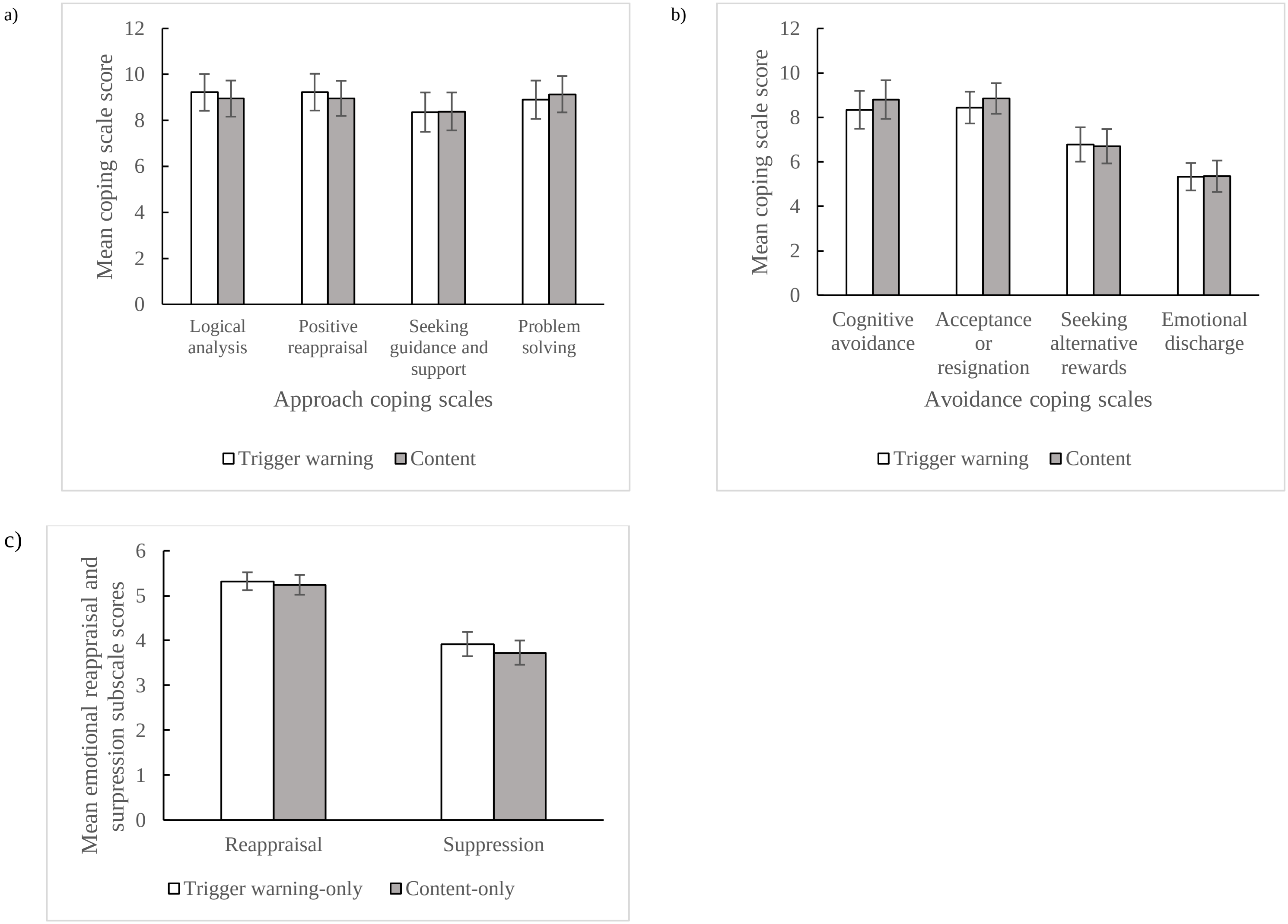


Figure 5.2a. Mean coping strategy scores on approach coping scales (with 95% Confidence Intervals) by future thinking scenario.

Figure 5.2b. Mean coping strategy scores on avoidance coping scales (with 95% Confidence Intervals) by future thinking condition.

Figure 5.2c. Mean emotional reappraisal and suppression subscale scores (with 95% Confidence Intervals) by future thinking condition.

Emotional appraisals

Recall our second aim: to examine participants' emotional reactions to imagining encountering a trigger warning versus trauma-related content. State anxiety increased significantly for all participants from baseline, to directly before, to directly after the future thinking scenario; a large main effect of time, ($F(1.65, 426.05) = 189.54$, $p < .001$, $\eta_p^2 = .424$, $BF_{10} = 7.442e+58$). Importantly however, we found that thinking about encountering a trigger warning-only resulted in similar levels of emotional reactions as imagining trauma-related content-only. That is, there were no significant interactions between time and future thinking condition ($F < 1$, strong evidence for no interaction: BF_{01} =15.84), or main effects of future thinking condition, for ratings of state anxiety ($F < 1$, $BF_{01} = 5.68$; Figure 3). Additionally, participants who imagined seeing a trigger warning-only versus content-only related to their most stressful/traumatic event reported similar phenomenological ratings; our analyses revealed no significant differences between conditions (*ps* = .154-.942; BF_{01} = 2.79-7.33).

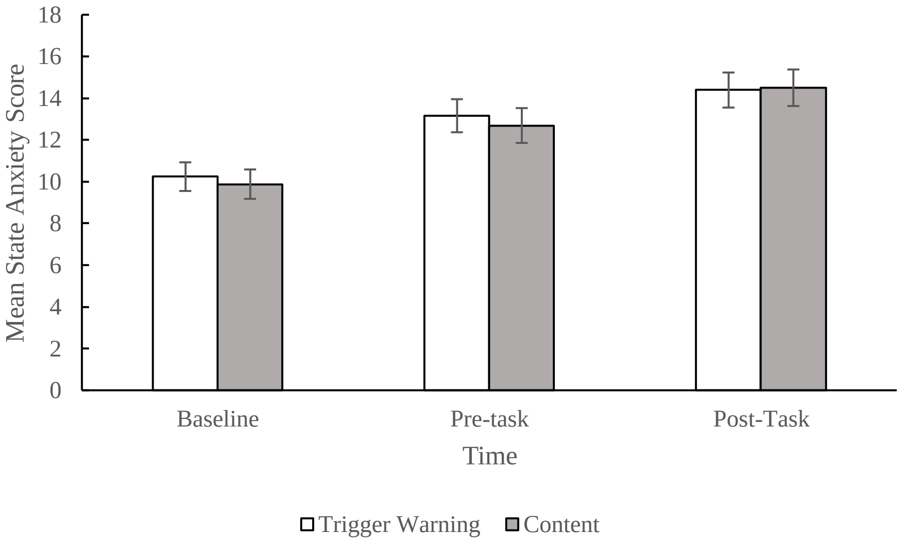


Figure 5.3. Mean state anxiety scores (with 95% Confidence Intervals) by future thinking condition and time.

Text analysis. We analyzed the text from the scenario and the open response coping question using the Linguistic Inquiry and Word Count (Pennebaker, Booth, Boyd, & Francis; 2015 internal dictionary) software to examine emotion-related words (Table 5.2). Because the negative emotion category includes the word stem “warn” and participants in the warning condition were instructed to describe a warning, we removed words containing “warn” prior to analyzing. There were no differences in word count between the warning-only and content-only conditions for the scenario description ($t < 1$, $d = -0.02$; overall $M = 96.51$, $SD = 54.71$), or for the open response coping question ($t < 1$, $d = 0.01$; overall $M = 43.64$, $SD = 33.91$). Only one significant difference remained after corrections for multiple comparisons; participants in the trigger warning-only condition on average used a lower percentage of positive emotion words (out of total words used), when answering the open response coping question (BF_{01} = substantial evidence).

Table 5.2

Summary of independent samples t-tests and Bayes Factors for text analysis for the future thinking scenario and open response coping question text[25]

		Future thinking scenario condition								
		Trigger warning-only		Content-only						
		M (SD)	*n*	*M (SD)*	*n*	*d [95% CI]*	*t*	*df*	*p*	BF_{01}
LIWC categories	*Examples*									
Future thinking scenario text										
Affective processes	Happy, cried	5.54% (3.25%)	130	5.63% (2.65%)	130	-0.03 [-0.27, 0.21]	-0.24	258	>.999	7.15
Positive emotion	Love, nice, sweet	1.49% (1.72%)	130	1.49% (1.48%)	130	0.007 [-0.24, 0.25]	-0.06	258	>.999	7.34
Negative emotion	Hurt, ugly, nasty	3.86% (2.67%)	130	3.91% (2.26%)	130	-0.02 [-0.26, 0.22]	-0.16	258	>.999	7.26
Anxiety	Worried, fearful	1.83% (1.85%)	130	1.30% (1.51%)	130	0.31 [0.07, 0.56]	2.52	247.67	.072	0.37
Anger	Hate, kill, annoyed	0.46% (0.76%)	130	0.59% (1.02%)	130	-0.14 [-0.39, 0.10]	-1.15	238.50	>.999	3.91
Sadness	Crying, grief, sad	0.73% (1.32%)	130	1.01% (1.43%)	130	-0.21 [-0.45, 0.04]	-1.67	258	.576	1.97
Open coping question text										
Affective processes	Happy, cried	6.30% (4.88%)	130	6.52% (4.74%)	130	-0.05 [-0.29, 0.20]	-0.37	258	>.999	6.89
Positive emotion	Love, nice, sweet	2.73% (2.97%)	130	4.14% (4.48%)	130	-0.37 [-0.62, -0.13]	-3.00	224.15	.018	0.11
Negative emotion	Hurt, ugly, nasty	2.92% (3.97%)	130	1.87% (3.02%)	130	0.30 [0.05, 0.54]	2.41	258	.102	0.48
Anxiety	Worried, fearful	1.26% (2.34%)	130	0.95% (2.14%)	130	0.13 [-0.11, 0.38]	1.09	258	>.999	4.19
Anger	Hate, kill, annoyed	0.21% (0.73%)	130	0.19% (0.68%)	130	0.02 [-0.22, 0.26]	0.17	258	>.999	7.25
Sadness	Crying, grief, sad	0.32% (0.94%)	130	0.33% (0.94%)	130	-0.02 [-0.26, 0.23]	-0.13	258	>.999	7.29

[25] Holm-Bonferroni corrections applied for six comparisons.

Questions about the effectiveness of trigger warnings. Recall our third aim: to assess how effective participants *believed* trigger warnings were in reducing distressing reactions. Only 35.8% of participants indicated that they believed a warning would prevent them from having an emotional reaction to upcoming material. This percentage did not differ between the warning-only (35.4%) and content-only conditions (36.2%; $\chi^2(1) = 0.017$, $p = .897$). Thus, imagining a trigger warning did not seem to enhance participants' perceptions that warnings were helpful. Finally, we asked participants in the warning-only condition if they believed that a warning would make them react differently than if they just saw content related to their trauma itself. While most participants said "no" (54.61%)—the warning would not make them react differently (e.g., "*It wouldn't make any difference*."), 45.38% said "yes" ("*I would react less negatively*."). This finding is striking considering that we did not find any evidence that thinking of a trigger warning would help participants to react differently towards trauma related content.

PTSD probability. Finally, to examine if trigger warnings were any more helpful (e.g., in bringing coping strategies to mind or reducing imagined negative reactions) for people with a probable PTSD diagnosis, we reran all of our analyses using PTSD probability as an additional factor.[26] The prevalence of mental health disorders in MTurk populations has been found to match or exceed that of the general population, and clinical measures demonstrate high reliability and validity (Shapiro, Chandler, & Mueller, 2013). Indeed, 96.5% of participants in our sample reported having experienced one (or more) High Magnitude Stressor event and 69.6% reported a Criterion A event (actual or threatened death or injury; Carlson et al., 2011). The most common events were the sudden death of a close family member or friend (60.4%), followed by exposure to a hurricane, flood, earthquake, tornado, or fire (39.2%). Further, 29.2% of the sample

[26] https://osf.io/anj65/

(warning condition = 27.7%, content condition = 30.8%; χ^2 (1) = 0.30, p = .585, φ = .03) were likely PTSD-positive according to the conservative PCL-5 cut-off (> 33; Bovin et al., 2016). Consistent with previous results, no interaction patterns emerged between the future thinking conditions and PTSD probability for our main outcome measures.

Interestingly, people who were PTSD-negative overwhelmingly indicated that a warning would not prevent the emotional impact of viewing trauma-related content (‘No’: 68.5% versus ‘Yes’: 31.5%), while people who were PTSD-positive were more evenly spread between ‘Yes’ (46.0%) and ‘No’ (53.9%) responses (χ^2 (1) = 4.94, p = .026, φ = .14). Similarly, the majority of people who were PTSD-negative indicated they believed that trigger warnings would not help them react differently compared to if they saw content related to their trauma (‘No’: 58.5%, versus ‘Yes’: 41.5%) versus people who were PTSD-positive, who were more evenly spread between responses (‘Yes’: 55.6%, versus ‘No’: 44.4%; though we note this difference was not statistically significant (χ^2 (1) = 2.08, p = .149, φ = .13). These findings indicate that people who are PTSD-positive generally perceive trigger warnings as more helpful than people who are PTSD-negative.

Discussion

Overall we found that imagining encountering a trigger warning-only does not prompt people to bring to mind more, or different kinds of, coping strategies compared to the same hypothetical situation without a warning (i.e., content-only)—including for participants with a probable PTSD diagnosis. Moreover, thinking about encountering a trigger warning or trauma-related content resulted in similar emotional reactions, with one exception: participants who imagined encountering a trigger warning-only (vs. content-only) used fewer positive emotion words when describing what they would do in that scenario. Finally, participants did not generally believe that trigger warnings would help reduce distressing reactions.

Our results may explain the consistent finding that trigger warnings do not ameliorate negative emotional reactions. If trigger warnings do not cause coping strategies to come to mind when people view subsequent material (e.g., reappraisal strategies) it stands to reason those emotional reactions are not improved. One interpretation of these findings—in line with previous findings (e.g., Bridgland et al., 2019; Sanson et al., 2019)—is that trigger warnings are inert. However, we found that imagining encountering a warning to be just as anxiety-provoking as imagining encountering trauma-related content. This result aligns with prior research showing that trigger warnings provoke uncertainty and anxiety (e.g., Bridgland et al., 2019). This uncertainty likely drove participants in the warning condition to use fewer positive emotion words when describing how they felt about the warning scenario.

The effectiveness of trigger warnings largely relies on warnings prompting people to draw on an existing coping strategy. However, if someone has not accessed mental health services then they may not know what coping strategies they could or should use—a conclusion supported by qualitative responses (e.g., *"I don't have a lot of coping techniques. I never was able to afford to see a therapist..."*). As an exploratory analysis, we examined the percentage of reported coping strategies when participants wrote their step-by-step response to the scenario versus the specific open response coping question. When asked to report specifically about coping strategies the number of approach strategies increased significantly ($\chi^2(1) = 29.51$, $p < .001$, $\Phi = .24$), while the percentage of avoidance strategies remained consistent. Therefore, future research could explore if trigger warnings could be more successful if they directly instructed people to bring existing coping strategies to mind. This is not to say that warnings should be more detailed (e.g., listing distressing aspects of content) but rather that they could specifically mention *coping strategies* themselves.

Our research has several limitations. First, although we draw on past experiences to generate hypothetical future experiences (Schacter & Addis, 2007), and intentions (e.g., I plan to exercise) *generally* map onto future behavior (e.g., actually exercising; *r* = 0.53; Sheeran, 2002), they may sometimes be inconsistent with actual behavior—the intention-behavior gap (Sheeran & Webb, 2016). Therefore, although measuring actual behavior was not our aim, hypothetically simulating the future may not capture what actual future behavior would look like. However, if people cannot bring to mind ideas about how warnings may be helpful during a low stress task (i.e., a future thinking task), it seems unlikely that they could bring to mind such strategies in a real-world setting. Indeed, there may not be much time between a warning and the warned-of content (e.g., on a TV show), and the circumstances may be more stressful than our scenario (e.g., in a public place like a lecture theatre).

Second, although participants could and did report they would use avoidance-based strategies in the scenarios, it is possible that we did not capture participants who tend to use avoidance as a primary coping strategy. These participants may have opted out of the survey at an earlier point (e.g., when reading consent information). Therefore, the true frequency with which people use avoidance strategies when they come across a trigger warning may be higher than reported here.

Third, participants may have had difficulty bringing coping strategies to mind during the future thinking task because they had already been reminded of their most stressful/traumatic event when completing the THS. Moreover, given that participants were paid a flat rate for completing the study—as is often the case with online research more broadly—regardless of the nature and length of their responses, it is possible that they were simply not motivated to write about what they would do in the scenario. However, given the THS was presented prior to the future thinking task in both conditions and payment was the same regardless of condition, any influence these factors had should be similar.

Forth, although beyond our aims here, we did not consider the efficacy of participants' reported coping strategies. It is generally accepted that avoidance strategies are maladaptive and that approach strategies are adaptive (Ehlers & Clark, 2000; Littleton et al., 2007). However, recently a more nuanced picture has emerged. Decreasing avoidance is key to exposure therapy (Rauch, Eftekhari, & Ruzek, 2012)—although experimental evidence shows that the use of avoidance does not necessarily reduce treatment efficacy (Blakey et al., 2019) and can assist with fear reduction within the early stages of treatment (Rachman, Radomsky, & Shafran, 2008). Additionally, recent theoretical (Bonanno & Burton, 2013) and experimental (Bonanno, Papa, Lalande, Westphal, & Coifman, 2004) work suggests that a flexible approach to coping—i.e., using a combination of strategies—may actually be the most efficacious.

Fifth, because trigger warnings were originally designed for trauma survivors and people suffering from PTSD, it is possible that our results would differ if we specifically recruited participants with a clinical diagnosis. Furthermore, our design did not test whether there might be a small subset of people with PTSD for whom trigger warnings provide a helpful opportunity to manage their reactions via coping strategies. However, we note that close to a third of our sample could be classified as "probable PTSD" according to the PCL-5.

In sum, our findings may help explain why trigger warnings fail to ameliorate emotional reactions to distressing material. While around half our sample believed trigger warnings would be helpful, we found no evidence to that thinking about a trigger warning, rather than thinking about actual exposure, was not more helpful in bringing more coping strategies to mind.

6 The effect of trigger warnings on avoidance behaviors in an analogue trauma task

Manuscript under review at *Behavior Therapy* as:

Bridgland, V. M. E. & Takarangi, K.T. M. (2021). Something distressing this way comes: The effects of trigger warnings on avoidance behaviors in an analogue trauma task.

Author Contributions: I developed the study design with the guidance of MKTT. I collected the data, and performed the data analysis and interpretation, and drafted the manuscript. MKTT made critical revisions to the manuscript and approved the final version of the manuscript for submission.

Abstract

Avoidance is one of the purported benefits and harms of trigger warnings—alerts that upcoming content may contain traumatic themes. Yet, previous research has focused primarily on emotional responses. Here, we used a trauma analogue design to assess people's avoidance behavior in response to stimuli directly related to an analogue trauma event. University undergraduates (*n* = 199) watched a traumatic film and then viewed film image stills preceded by either a trigger warning or a neutral task instruction. Apart from a minor increase in avoidance when a warning appeared in the first few trials, we found that participants did not overall avoid negative stimuli prefaced with a trigger warning any more than stimuli without a warning. In fact, participants were reluctant overall to avoid distressing images; only 12.56% (*n* = 25) participants used the option to cover such images when given the opportunity to do so. Furthermore, we did not find any indication that trigger warning messages help people to pause and emotionally prepare themselves to view negative content. Our results contribute to the

growing body of literature demonstrating that warnings seem trivially effective in achieving their purported goals.

Introduction

A scene depicting dubious sexual consent in Netflix's regency drama series *Bridgerton* (2020) was the recent cause of online uproar, with one Twitter user noting: "I was considering watching…however I have decided I will not because rape scenes on screen are a trigger for me" (Logan, 2020). Echoing these concerns, news outlets were quick to point out that the "controversial sex scene needs a trigger warning"—an alert that upcoming content may contain themes similar to traumatic experiences that could "trigger" someone to re-experience a traumatic event (Jean-Philippe, 2020), for example via vivid thoughts, feelings or flashbacks about the event (Ehlers, Hackmann, & Michael, 2004). These criticisms are not new, and in 2021 Netflix is under increasing pressure to add trigger warnings to their content (Medhora, 2021). Advocates of this idea—and of trigger warning use in other domains (e.g., academic and online contexts)—claim that such warnings are necessary so that people have the chance to emotionally prepare for, or to completely avoid, the content (George & Hovey, 2020; Strothman, 2021). However, thus far, the extant research has focused primarily on people's emotional reactions to novel content following trigger warnings and less so on how trigger warnings may—or may not—lead to avoidance behaviors. Moreover, no previous research has accounted for the *uniqueness* of trauma triggers—which typically relate to stimuli similar to a traumatic event (Ehlers et al., 2004). Here, we addressed these shortcomings of previous research. Using a trauma analogue design, we assessed people's avoidance behaviors in response to stimuli directly related to an analogue trauma event. We had two main aims: First, to investigate if participants would be more likely to avoid content associated with an analogue trauma film when that content is preceded by a trigger warning message; and second, to assess if participants' avoidance behavior would be related to a) their emotional responses to the stimuli, and b) individual difference factors such as state and trait anxiety, experiential avoidance, perceptions of harm, reported use of avoidance strategies, and perceptions that trigger warnings are beneficial.

Paradoxically, avoidance is both one of the purported harms *and* one of the potential benefits of implementing trigger warnings. Critics (e.g., Lukianoff & Haidt, 2015) argue that trigger warnings contribute to a "counsel of avoidance" culture (McNally, 2019, as quoted in Flaherty, 2019). In support of these concerns, a substantial literature implicates avoidance as a primary maintaining factor in Posttraumatic Stress Disorder (PTSD; Badour, Blonigen, Boden, Feldner, & Bonn-Miller, 2012) as well as a central characteristic of a broad range of mental disorders (Krypotos, Effting, Kindt, & Beckers, 2015). In some exceptions to this rule, avoidance can assist with fear reduction within the early stages of treatment (Rachman, Radomsky, & Shafran, 2008) and increase adherence to concurrent exposure therapy (Levy & Radomsky, 2014). Indeed, warning advocates claim trauma survivors should be able to decide if they want to avoid content that may trigger re-experiencing symptoms, arguing that avoidance aids recovery (Cripps, 2020). However, regardless of long-term harms or benefits, little to no research has addressed the underlying assumption that trigger warnings influence—i.e., by promoting or not—avoidance-type behaviors.

Advocates also claim trigger warnings allow time for people to emotionally "prepare" for, or cope with content. For instance, Strothman (2021) of the Digital Citizens Academy—which focuses on educating the public about mental health—writes that trigger warnings allow people "*to take a step back and pause or pass over the content*." That is, seeing a trigger warning should help someone to bring helpful strategies to mind (e.g., reappraisal, e.g., "*a television show is not real life and therefore poses no real threat*"), thus reducing negative emotional reactions. However, previous research—focused primarily on people's emotional responses towards content following warning messages (e.g., how warning messages make people feel; Bellet, Jones, & McNally, 2018; Boysen, et al., 2021; Bridgland, Green, Oulton, & Takarangi, 2019; Bruce, 2019)—shows that trigger warnings *do not* ameliorate distress. One possibility could be that despite advocates' claims, trigger warnings do not help people "*pause and prepare*" by bringing coping strategies to mind.

In the first empirical investigation to explore how trigger warning messages may or may not change the coping strategies people bring to mind, we asked participants to report what they would do when encountering a trigger warning related to their most stressful/traumatic experience (Bridgland, Barnard, & Takarangi, 2021). These participants reported a similar number of approach-based strategies (e.g., reappraising a situation in order to reduce its emotional impact) and avoidance-based strategies (e.g., leaving the situation) to participants who imagined the same hypothetical situation of encountering trauma-related content, but without a warning. However, due to the intention-behavior gap—i.e., that intentions do not always map onto future behavior (Sheeran & Webb, 2016)—asking people to hypothetically simulate the future may not capture actual behavior. To directly investigate avoidance behaviors, participants would need the opportunity to avoid content following a trigger warning.

Only a handful of studies have explicitly examined whether people avoid material accompanied by warning messages. Gainsburg and Earl (2018) asked participants to select a film to watch from a series of titles and found no difference in how often participants selected titles accompanied by a trigger warning or no warning. But, participants' anticipated anxiety about warned-of content, as well as the belief that trigger warnings are protective from real harm (vs. coddling/overprotective), was associated with increased avoidance of material accompanied by trigger warnings. These results suggest that trigger warnings may encourage avoidance only among certain people.

However, other research has not found increased avoidance among groups who are traditionally associated with the belief that trigger warnings are helpful in reducing distress (i.e., trauma survivors). Bruce and Roberts (2020) found no preference for articles labelled with trigger warnings compared to the same titles without warnings, including among participants who had experienced trauma matching the article (i.e., interpersonal violence). Kimble et al. (2021) found only a small minority of participants (< 6%)—including those with a history of trauma or with probable PTSD—avoided potentially triggering text when provided with a non-triggering alternative option. Similarly, we found that when presented with a single Instagram sensitivity screen (a version of a trigger warning focused on blurring graphic content), only 10-15% of people opt to avoid potentially distressing content, while the remainder opt to uncover and reveal the image (Bridgland, Bellet & Takarangi, 2021). Indeed, Simister, Bridgland, and Takarangi (2021) found that more than 90% of people repeatedly uncover content covered by a sensitivity screen, even after exposure to graphic photo imagery underneath. Importantly, neither study found evidence that users with mental health concerns (e.g., symptoms of depression or PTSD) were any more likely to use the screens to avoid sensitive content. In fact, in one of Bridgland et al.'s studies, participants' desire to view potentially negative content covered by a warning screen was associated with risk markers for psychopathology (e.g., lowered wellbeing). Trigger warnings in this instance may therefore foster a "Forbidden Fruit effect" (Ringold, 2002)—where a restricted behavior becomes more desirable—and encourage morbid curiosity about distressing content (Oosterwijk, 2017). Taken together, early research provides an incomplete account of how warnings may affect avoidance behaviors.

A persisting problem with *all* previous trigger warning research is the *uniqueness* of trauma triggers. That is, trauma triggers typically relate to stimuli with sensory similarities to the traumatic event, unique to a person's individual situation (Ehlers et al., 2004). For instance, while a fictional depiction of sexual assault (e.g., on *Bridgerton*) may trigger one person to have re-experiencing symptoms (i.e., be 'triggered'), the same scene may not elicit any response in another person who has also experienced sexual assault. Therefore, a crucial piece of information is still missing from trigger warning research: when warned about and given the chance to avoid material *that reminds someone of their trauma*, do they actually avoid it?

Of course, ascertaining individual participants' unique trauma triggers and then matching study stimuli to those triggers would be challenging. The "trauma film paradigm," a well-established method to simulate exposure and reactions to psychological trauma, provides a practical alternative (Holmes, Brewin, & Hennessy, 2004; Lazarus, Speisman, Mordkoff, & Davison, 1962; Holmes & Bourne, 2008; James et al., 2016). Typically, non-clinical participants watch a short film depicting a traumatic event and then answer questions about the event and their emotional responses. This experience reliably induces PTSD-like symptoms—such as intrusive memories of the film footage—that are similar to, but occur to a lesser extent and shorter duration, as those for real trauma (Holmes & Bourne, 2008; James et al., 2016). This paradigm therefore offers a suitable method to study people's avoidance of warned of stimuli that *directly reminds* them of a trauma analogue event (i.e., stimuli taken from the film).

Here, participants watched a traumatic film and then viewed 32 still images taken from the film for 5(s) each. On half the trials, a trigger warning message preceded each image; on the remaining trials a neutral task instruction appeared before each image. We instructed participants to view each image for the entire time it was displayed, but also gave them the option of pressing a "stop viewing" button that would take them to a blank screen for the equivalent time if they did not want to view the image. We therefore measured avoidance behaviors in two distinct ways: passive avoidance—operationalized as more time spent avoiding viewing the image by remaining on the instruction screen, and active avoidance—operationalized as the number of times participants chose to cover photos and the subsequent time spent viewing each image. We also examined if avoidance behaviors were related to emotional responses or individual difference factors as suggested by previous research (e.g., Gainsburg & Earl, 2018). To do so, we measured state anxiety, projected PTSD-like symptoms, perceptions of how others might be harmed by the content, trait anxiety, experiential avoidance (the tendency to avoid feelings associated with anxiety), beliefs that general trigger warnings are helpful/were helpful in the study, use of and perceptions that avoidance/approach strategies are beneficial, prior exposure to traumatic events, and also prior personal experience with the topic of the film.

We recruited an undergraduate university population. This sample is particularly appropriate to evaluate the reported efficacy of trigger warning messages for three reasons. First, the debate about the purported benefits and harms of trigger warnings is centralized around the use of trigger warnings on college campuses (e.g., Lukianoff & Haidt, 2015). Second, many universities now mandate the use of trigger warnings as part of mental health initiatives (e.g., Harris, 2016; Palmer, 2017). Third, at least one third of first-year students around the world screen positive to an anxiety, mood, or substance abuse disorder (DSM–IV; Auerbach et al., 2018).

We had two competing predictions regarding avoidance behaviors. On the one hand, if critics' and advocates' claims that trigger warnings lead to avoidance are true, then we would expect participants to engage in more avoidance behaviors—more time spent waiting on the message screen, more instances of covering an image and less time spent viewing an image before it is covered—on trigger warning message trials compared to neutral task instruction trials. On the other hand, given other research has found that trigger warnings do not promote avoidance behaviors (Bruce & Roberts, 2020, Gainsburg & Earl, 2018), it is also possible that we may find little difference in avoidance behaviors between warning and control trials.

We made additional predictions about the emotional consequences of avoidance behaviors. If advocates' claims—that trigger warnings help people to *pause* and emotionally *prepare* (Strothman, 2021) to view distressing content—are true, then we should find that the time spent waiting on the message screen before viewing each image, particularly the trigger warning screen, is negatively associated with extent of negative emotional reactions throughout the study (i.e., state anxiety after the image task, projected future PTSD-like symptoms and harm to others). However, based on critics' claims, as well as other previous empirical work showing that trigger warnings have negligible effects on emotional responses (e.g., Bridgland et al., 2019) and reported coping strategies (Bridgland, Barnard & Takarangi, 2021), we may instead find that time spent on the message screens is not associated with reduced distress.

Finally, based on previous work showing that certain populations (e.g., people with increased anxiety; Gainsburg & Earl, 2018) may be more likely to avoid content marked with a trigger warning, we expected avoidance behaviors would be positively correlated with trait anxiety, experiential avoidance, reported use of and perceptions that avoidance strategies are beneficial overall and within the study (i.e., perceptions that being able to press a key to stop viewing the images was beneficial), and perceptions that trigger warnings are helpful overall/within the study. Further, avoidance behaviors should be negatively associated with participants' reported use of approach strategies and perceptions that approach strategies are beneficial overall and within the study (i.e., perceptions that viewing the images in the image task was beneficial in reducing distress). Finally, we expected these correlations to be stronger when participants saw a trigger warning message versus control message.

Method

The Flinders University Social and Behavioural Research Ethics Committee approved this experiment. We preregistered this experiment (https://osf.io/dmeuq). The data and supplementary material for this experiment can be found here: https://osf.io/gcsf8/. We have reported all measures, conditions, and data exclusions.

Participants

According to an a priori power analysis for a two-tailed, matched pairs t-test (using G*Power; Faul, Erdfelder, Lang, & Buchner, 2007) with an alpha of .05, power of .80, and effect size of d = 0.2 (the smallest effect we would be interested in) a sample size of N = 199 is required.

We recruited 207 participants using the Flinders University participation system via SONA. We excluded four participants because of electronic data collection malfunction and four participants withdrew during the film viewing phase of the study. Our final sample of 199 participants ranged from 18-60 years (M = 22.78, SD = 6.58), and were predominately female (71.9%), and Caucasian/White/European (41.7%), followed by Asian (19.6%), Middle Eastern (3.0%), African (2.01%), 1.0% Indigenous Australian, and 1.0% Polynesian. Some participants specified nationality[27] (Australian 23.6%), nationality and ethnicity (3.5% European-Australian; 1.5% Asian-Australian; 0.5% Asian-European; 1.0% Middle Eastern-Australian; 0.5% South American-Australian) or provided no answer (0.5%).

Materials

Film stimulus. Participants watched an 8-minute negative film widely employed within analogue trauma research (e.g., Green, Strange, Lindsay, & Takarangi, 2016; Takarangi, Strange, & Lindsay, 2014; Woodward, & Beck, 2017) from the 1988 fictional movie *The Accused*. This clip depicts the gang rape of a female in a bar.

Film ratings. After watching the film, participants indicated (a) how distressing they found the film, and (b) how closely they paid attention to it, and (c) how involved they felt on an 10-point Likert scale (1 = *Not at all*, 10 = *Extremely*).

Image stimuli. We generated an image still from every 7(s) of the film, resulting in 64 images. To determine the 32 most distressing images, pilot participants (*N* = 33) first watched the film, and then answered, "How distressed do you feel at the present moment?" (on a scale of 1 = *Very slightly or not at all*, 5 = *Extremely*) in response to each of the 64 images, presented in a random order (resulting 32 most distressing images: *M* = 4.55, *SD* = 0.24).

Image task. All participants saw all 32 images, but we manipulated message within-subjects: half of the image trials were preceded by a trigger warning message ("*Warning: The image you are about to view contains disturbing content that may be distressing*."), and half with a control message ("*Instructions: When you are ready, the next image will be displayed on the next screen*."; counterbalanced). We created four predetermined random sets of image and

[27] Participants were asked to report their ethnicity in an open text box resulting in the range of answers provided—i.e., some people interpreted "ethnicity" as "nationality", and some participants provided both nationality and ethnicity.

message screen pairings (control and warning). Participants saw one image set, presented on E-Prime (Psychology Software Tools, Pittsburgh, PA). Image and message pairings appeared in a random order with one exception: the first two trials were always one warned image and one control image (randomized) from four predetermined subsets of images (i.e., one from each larger set). This setup allowed us to examine participants' first two responses in isolation, in case participants habituated to the effect of the warning over extended trials, or the effect of the warning message on behavioral responses was temporally brief.

Participants could spend an unlimited time on each message screen and up to 5(s) on each image. We instructed participants that they should view each image for the entire time it was displayed, but advised them that they could use a "stop viewing" button at any time before the end of the 5(s) which would take them to a blank screen. Thus, participants had the option to stop viewing an image and to instead view a blank screen for the remainder of the viewing time. Participants could not move through the overall task faster by using the "stop viewing" option.

Short-form Spielberger State-Trait Anxiety Inventory (STAI-6; Marteau & Bekker, 1992; Appendix D). Participants rated how they felt at that current moment for three anxiety-present items (e.g., "*I am worried*") and three anxiety-absent items (e.g., "*I feel calm*";1 = *Not at all*, 4 = *Very much*; (present study α = .75-.83). Scores reflect state anxiety responses and are summed (range 4-20).

Anticipated traumatic stress symptoms (Appendix Q). Participants completed a modified version of the Pretraumatic Stress Reactions Checklist (PreCL; Berntsen & Rubin, 2014). The original PreCL asks participants to rate traumatic stress symptoms that reflect the DSM-5 symptom criteria for PTSD in relation to possible *future events*. Here, we modified the measure to instead ask about possible *future symptoms* that participants anticipated over the next

24 hours (e.g., *Repeated, disturbing and unwanted images related to the film or images, Avoiding imaginings, thoughts or feelings related to the film or images*) as a result of doing the film/image task (0 = *Not at all*, to 4 = *Extremely*. One item from the original scale was excluded because it did not fit the present context (Blaming yourself or someone else for a possible future stressful experience or what has led up to it). Scores are summed (range 0-76). Despite modifications to the instructions, internal reliability in the present study was excellent (*α* = .92).

Spielberger Trait Anxiety Inventory (1983; STAI-form Y; Appendix R). Participants rated how often they generally experience a series of anxiety symptoms (e.g., anxiety present item: "*I worry too much over something that really doesn't matter*" and anxiety absent item: "*I am content; I am a steady person*.";1 = *Almost never*, to 4 = *Almost always*). The scale has excellent internal consistency (present study *α* = .93). Total scores range from 20 to 80, with higher scores indicating greater trait anxiety.

The Acceptance and Actions Questionnaire (AAQ; Hayes et al., 2004; Appendix S). Participants rated how well a series of statements reflecting experiential avoidance—a tendency to avoid particular private experiences (e.g., bodily sensations, emotions, thoughts etc.) typically associated with anxiety (e.g., If I could magically remove all of the painful experiences I've had in my life, I would do so)—generally applies to them (on a scale of 1 = *Never true*, to 7 = *Always true*; present study; *α* = .71). Scores are summed with higher scores indicating greater experiential avoidance (range 9-63).

Coping strategies following a traumatic event (Appendix T). Participants rated how often they used and the beneficial nature of: a) behavioral approach, b) emotional approach, c) behavioral avoidance, and d) emotional avoidance coping strategies, following a stressful/traumatic event (on a scale of 1 = *Never*, to 5 = *Often*).

Questions related to the image task and warning screens. Participants rated the perceived effectiveness of the image rating task: a) Did you find it beneficial to view images related to the film again? (e.g., do you think it made you feel less distressed over time?), b) Did you find it beneficial to be able to stop viewing the images with the "stop viewing" option? (e.g., did it make you feel less distressed?; 1 = *Not at all*, 5 = *Extremely*), c) I needed the warning messages about the images to prevent them from causing distress, d) People should always receive a warning message before viewing images like this (1 = *Strongly disagree*, 5 = *Strongly agree*).

Topic of the film relevance. Participants indicated a) if they had any personal experience with the topic of the film/images (Yes/No), b) how much anxiety the average person would feel viewing the film and images, and c) how much anxiety someone who had a personal experience with the content would feel viewing the film and images (0 = *None*, to 5 = *An extreme amount*).

Single item Criterion A question. We asked participants to think of their most traumatic or stressful event, and to indicate (Y/N), if they were exposed to: death, threatened death, actual or threatened serious injury, or actual or threatened sexual violence, in any of the following way(s): a) Direct exposure, b) Witnessing the trauma c) Learning that a relative or close friend was exposed to a trauma d) Indirect exposure to aversive details of the trauma, usually in the course of professional duties (e.g., first responders, medics; i.e., Criteria A for PTSD in the DSM-5).

Beliefs about trigger warnings as protective (vs. coddling; Gainsburg & Earl, 2018). Participants rated their agreement with two statements, "Trigger warnings that precede distressing content 'coddle' people, hurting them in the long run," and "Trigger warnings that precede distressing content 'protect' people, helping them in the long run" on 7-point scales (1 =

Strongly disagree, to 7 = *Strongly agree*). A two-item composite score was created by reverse scoring the "coddling" item and averaging the two items. Additionally, participants were also coded as believing that trigger warnings are coddling (those one standard deviation below the mean), of average protectiveness (falling between one standard deviation below and above the mean) or more protective (those one standard deviation above the mean; Gainsburg & Earl, 2018).

Procedure

After completing informed consent procedures participants completed demographic questions and a measure of baseline state anxiety (STAI-6). Next, participants watched the film, and completed a second measure of state anxiety, in addition to the film rating questions. Participants then completed the image task. To enhance our cover story and obscure our hypothesis, we falsely told participants these images would relate to a memory test later in the experiment. Participants then completed a third measure of state anxiety, possible future post-traumatic stress symptoms related to the film (Pre-CL), trait anxiety (STAI-Form Y), experiential avoidance (AAQ), questions about coping strategies, questions about the warning messages/image task, personal experiences with the topic of the film, the single item Criterion A trauma question, and the question about their belief in trigger warnings as protective or coddling. We also asked participants if they had seen the film before.[28] Participants were then fully debriefed and granted course credit (*n* = 170) or paid $10AUD (*n* = 29) for their time.

Results

Statistical overview

[28] See https://osf.io/tgk7n/ for descriptive statistics.

We ran analyses using Null-Hypothesis Significance Tests (α = .05) in SPSS Version 25 and JASP for MacOS version 0.13.1. Since each of our analyses relates to a preregistered hypothesis, we have retained original *p*-values and have not corrected for multiple comparisons (Rubin, 2021).

Effectiveness of the analogue trauma task

Before turning to our main analyses, we examined how effective our analogue trauma film and image task was in inducing negative affect, as well as task compliance. Participants rated the film as distressing (*M* = 8.96, *SD* = 1.28), paid close attention (*M* = 8.83, *SD* = 1.25), and felt highly involved (*M* = 7.56, *SD* = 2.14). To examine state anxiety throughout the study, we ran a repeated measures ANOVA on state anxiety scores at baseline, post-film, and post-image task. The overall main effect of time was significant (*F* (1.49, 294.18) = 858.18, *p* <.001, partial eta = .81).[29] State anxiety scores increased significantly from baseline (pre-film; *M* = 10.57, *SD* = 2.84), to post-film (*M* = 19.84, *SD* = 3.37, *p* <.001), and decreased from post-film to post-image task (*M* = 18.67, *SD* = 3.74, *p* <.001).

Avoidance behavior

Reaction time data exclusions. As per our pre-registration, we excluded reaction time data if it fell below 200 ms or over 3SD above the mean—based on each participant's average reaction time for that task. However, likely because the first two trials were the first time participants viewed the warning and control screen messages (randomized as either the first or second trial participants saw), participants' first two trial responses (*M* = 4235.95, *SD* = 1288.14) were on average significantly slower than the remaining trial responses (*M* = 1666.08, *SD* =

[29] Mauchly's Test of Sphericity was violated (<.001), therefore a Greenhouse-Geisser (.743) correction was applied (Field, 2005).

985.69, paired samples t-test = *t* (198) = 27.33, *p* <.001). Likewise, for participants who chose to avoid images with the black screen during viewing time, the average time taken to cover the image was slower when it occurred during the first two trials (*M* = 3277.28, *SD* = 1068.04) than for the remaining trials (*M* = 2214.80, *SD* = 850.91). Only six participants chose to cover an image during the first two trials (M = 3097.00, SD = 1046.82) *and* in the remaining trials (*M* = 1770.13, *SD* = 1049.16, paired samples t-test = *t* (5) = 3.37, *p* = .020). Thus, for the first two trials, we only excluded responses < 200 ms. In total, we excluded 16 message screen responses < 200 ms, and 89 message screen and three cover time responses for falling higher than 3*SD* above the mean for that participant.[30]

We now turn to our main research aim—were participants more likely to avoid stimuli associated with an analogue trauma film if the content was preceded by a trigger warning message versus a control screen? Recall that avoidance behaviors were operationalized as a) time spent avoiding viewing an upcoming image by remaining on the instruction screen, b) number of images in each condition that participants covered, c) time spent viewing images. We now examine each of these avoidance behaviors in turn.

Time spent avoiding viewing the image by remaining on the instruction screen. Recall that trigger warning advocates claim that warnings help people to pause and prepare to cope with content. For this claim to be true, we expected that participants would spend *more* time on the warning screens compared to control screens. Contrary to these claims, across all 32 trials there was no difference in the average time participants spent on the instruction screen when it was a trigger warning compared to when it was a control instruction (paired samples: *t* (198) =

[30] Only 10 participants had more than one data point excluded. See https://osf.io/tgk7n/ for a detailed breakdown.

0.67, *p* = .502, *d* = 0.05, 95% CI [-0.09, 0.19]; Figure 1). Because participants may have come to expect— over the course of the image task—that similar types of images were covered by warning and control screens, we also examined the first two trials only (which were always a warning and a control screen; randomized). Participants tended to spend *less* time on the trigger warning, versus the control screen (paired samples: *t* (198) = -2.11, *p* = .036, *d* = -0.15, [-0.29, -0.01]) within the first two trials. A similar pattern emerged when we compared participants who viewed a warning (*n* = 100) with those who viewed a neutral task instruction (*n* = 99) on the first trial, between-subjects; but this difference was not statistically significant (*t* (197) = -1.61, *p* = .110, *d* = -0.23, [-0.51, 0.05]).

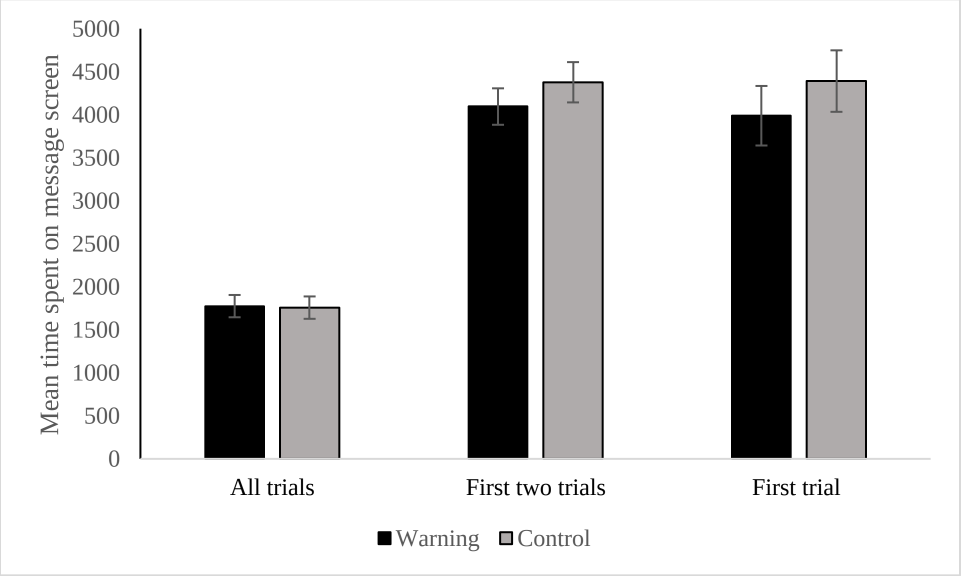


Figure 6.1. Mean time (ms) spent on instruction screens (with 95% Confidence Intervals) by message screen type.

Number of images that participants covered and time spent viewing images before they were covered. Recall that warning advocates and critics both claim that warnings encourage people to avoid potentially distressing stimuli. For this claim to be true, we expected that participants would be more likely to cover, and thus would spend less time looking at, images preceded by warning screens compared to the control screens. However, overall, participants rarely avoided the images by using the cover option: only 239 images were covered out of a total of 6,368 trials (3.75%), only 25 (12.56%) participants used the cover feature at all; of these, five participants covered only one image and 20 participants covered more than one. Therefore, our resulting sample sizes for the following analyses should be interpreted with extreme caution as they are drastically underpowered.

Among participants who covered both warning and control screens (*n* = 20), covering rates were similar for images preceded by a warning screen (*M* = 5.75, *SD* = 4.51) and images preceded by a control screen (*M* = 5.95, *SD* = 4.51, paired samples t-test: *t* (19) = 0.61, *d* = 0.14, 95% CI [-0.31, 0.57]). The average time spent viewing images preceded by a trigger warning (*M* = 2225.32, *SD* = 959.21) was also similar to images preceded by a control (*M* = 2116.70, *SD* = 945.70; paired samples: *t* (19) = -0.68, *p* = .505, *d* = -0.15, [-0.59, 0.29]). Participants covered twice as many images if they were preceded by a trigger warning message (3.01% of total participants, *n* = 6) within the first two trials, versus a neutral task instruction (1.51% of participants, n = 3; χ^2 = 42.20, *p* <.001, *Φ* = .461. Only two participants used the cover feature for *both* of the first two image trials (trigger warning: *M* = 3949.50, *SD* = 122.32, control: *M* = 2766.50, *SD* = 2130.51; paired samples: *t* (1) = -0.74, *p* = .593, *d* = -0.52 [-1.96,1.07]). Only four participants who viewed a warning (average time before covering images: *M* = 3581.50, *SD* = 1029.72) on the first trial covered the image, while no participants who viewed a control screen on the first trial covered the image. Therefore, we could not run an independent samples t-test for these data. However, a one sample binomial test revealed that the degree to which participants avoided images (by choosing to cover the image), when a warning occurred in the first trial (estimate = 0.04, in the form of choosing to cover the image), was significantly greater than .001 (*p* <.001; n = 100; Clopper-Pearson 95% CI [.01, .09]).

Taken together, contrary to the claims of advocates and critics, participants were extraordinarily reluctant to avoid viewing negative study stimuli when given the option to do so throughout the image viewing task, and warnings did not enhance avoidance behavior—apart from a minor increase in avoidance when a warning appeared in the first few trials.

Were avoidance behaviors associated with emotional responses and individual differences?

Recall our secondary aim: to assess if avoidance behaviors were related to emotional responses and to individual difference factors. We correlated our avoidance behaviors (time spent on message screens, number of images participants covered and time spent on images) with state anxiety, film ratings, trait anxiety, experiential avoidance, reported use of and perceived benefits of coping strategy types (approach vs. avoidance), anticipated symptoms, how others' would perceive the film/image content, perceptions that trigger warnings are protective (vs. coddling), and the perceived effectiveness of the image task (see Tables 6.1 and 6.2 and also https://osf.io/tgk7n/ for analyses pertaining to perceptions that trigger warnings are protective vs. coddling). As per our preregistration, we also ran exploratory analyses to investigate whether splitting the data on various individual difference characteristics (i.e., having prior experience with the topic vs. no prior experience; belief that trigger warnings are coddling vs. of average protectiveness or more protective, or having experienced vs. not experienced a Criterion A event), revealed different relationships with our main avoidance variables (see Tables 6.1 and 6.2 and also https://osf.io/tgk7n/ for all analyses pertaining to having experienced vs. not experienced a Criterion A event).

Table 6.1

Welch's t-test and descriptive statistics for participants who said they have personal experience with the topic of the film on key avoidance variables

					95% CI for Cohen's *d*					
Avoidance type	*t*	*df*	*p*	Cohen's *d*	Lower	Upper	Personal experience	*n*	*M*	*SD*
Average time spent on message screens across all trials	0.69	50.40	.496	0.13	-0.22	0.48	Yes	40	1871.74	1135.86
							No	159	1740.14	845.88
Average time spent on control screens	0.91	47.38	.365	0.18	-0.17	0.53	Yes	40	1913.06	1278.27
							No	159	1718.79	820.81
Average time spent on warning screens	0.39	55.50	.697	0.07	-0.28	0.42	Yes	40	1830.41	1017.73
							No	159	1761.48	904.43
Average time spent before covering images across all trials	-1.98	19.64	.062	-0.79	-1.66	0.09	Yes	8	1835.05	677.84
							No	17	2509.63	998.54
Average time spent before covering control images	-1.53	17.82	.144	-0.64	-1.52	0.25	Yes	8	1822.11	761.39
							No	15	2385.34	976.95
Average time spent before covering warning images	-2.63	19.95	.016	-1.04	-1.98	-0.08	Yes	7	1685.98	469.36
							No	15	2573.70	1112.35
Total times images covered across trials	0.61	13.57	.552	0.26	-0.59	1.10	Yes	8	11.25	9.56
							No	17	8.76	9.38
Total times images control images covered	0.52	14.05	.608	0.23	-0.64	1.09	Yes	8	6.00	4.69
							No	15	4.93	4.56
Total times images warning images covered	0.45	11.84	.664	0.20	-0.70	1.10	Yes	7	6.00	4.90
							No	15	5.00	4.91

Table 6.2

Correlations between distress variables and individual differences, and key avoidance variables

	n	Baseline STAI	Post Film STAI	Post Photo Task STAI	Film Distressing	PCL Modified	AAQ	Trait Anxiety	TW Protect	Photo task beneficial	Cover option beneficial	Warnings needed for images like these	Should always receive warnings for similar content	Anxiety average other person	Anxiety other personal experience	Avoid General average	Approach general average	Avoid Beneficial	Approach beneficial
Average time spent on message screens across all trials	199	-.04	-.01	-.02	-.06	.01	<.001	-.01	.08	.06	.06	.18*	.02	-.06	-.07	-.01	.14*	.01	.03
Average time spent on warning screens	199	-.03	.01	<.01	-.04	.01	<.01	.01	.09	.05	.05	.16*	<.01	-.01	-.07	-.02	.13	.02	.04
Average time spent on control screens	199	-.04	-.02	-.03	-.07	.01	<.001	-.02	.06	.06	.06	.19*	.05	-.10	-.06	<.01	.15*	<.01	.01
Total times images covered across trials	25	.25	.12	.21	.18	.31	.11	-.02	<.01	-.31	.19	-.10	.14	.05	.24	-.03	.04	.23	-.33
Total times images warning images covered	22	.31	.03	.19	.17	.24	.01	-.06	-.04	-.21	-.06	-.21	-.03	-.05	.15	-.25	.15	.20	-.29
Total times images control images covered	23	.16	.11	.20	.22	.40	.10	<.001	.04	-.27	.28	-.03	.25	.07	.30	.07	-.07	.21	-.30
Average time spent before covering images across all trials	25	-.33	.08	-.06	-.05	-.33	-.17	-.03	-.03	.47*	-.25	.08	-.02	-.08	<.01	-.06	-.17	-.27	.04
Average time spent before covering warning images	22	-.35	.19	-.08	-.05	-.27	-.14	-.03	-.39	.62**	-.31	.14	-.13	-.12	<.01	-.06	-.10	.34	.97
Average time spent before covering control images	23	-.33	-.16	-.15	.13	-.49*	-.12	-.19	-.23	.34	-.34	-.06	.03	-.11	.08	-.12	-.08	-.23	-.08

Note * correlation significant at the 0.05 level (2-tailed) ** correlation significant at the 0.01 level (2-tailed)

Time spent avoiding viewing the image by remaining on the instruction screen and emotional responses and individual difference factors. The belief that participants needed warning messages about the images to prevent those images from causing distress was positively correlated with overall average time spent on message screens (*n* = 199, *r* =.18, *p* = .013), time spent on warning message screens (*n* = 199, *r* = .16, *p* = .028), and time spent on control message screens (*n* = 199, *r* = .19, *p* = .008). Therefore, it is possible that general beliefs about the usefulness of warning signals (i.e., as ways to prepare for upcoming content) extended to *both* screen types and led these participants to linger longer on the screen preceding each image as a way to emotionally prepare/brace themselves to view those images. However, the time participants spent on message screens—both overall and by screen type—was *not* associated with any reductions in reported distress (i.e., state anxiety, *rs* = -.01-.02, post image task, projected future symptoms, *rs* = .01, or perception of possible harm to others *rs* = -.01-.10, all *ps* > .05). Therefore, it seems unlikely that time spent waiting on each message screen—regardless of whether this screen contained a warning—helped participants to emotionally prepare to reduce the negative impact of each image.

Finally, generally using approach strategies when personally experiencing a stressful/negative event was positively associated with the average time spent across both message screen types (*n* = 199, *r* = .14, *p* = .046) as well as the average time spent on the control message screen (*n* = 199, *r* = .15, *p* = .040), and the average time spent on the warning message screen (*n* = 199, *r* = .13, *p* = .065) analyzed individually, though the relationship with warning screen was not statistically significant. Perhaps participants who are generally more likely to try and "*engage with thoughts and feelings*"[31] lingered longer on the screens to try and think about/reappraise the upcoming content in a different way. However, the effect sizes of these relationships are small, with p-values close to .05 and thus should be interpreted with caution.

Advocates claim that trigger warnings help trauma survivors prepare to cope with upcoming content. For this claim to be true we expected that trauma survivors would spend *more* time on the warning screens compared to people who had not experienced a trauma. However, there was no difference for time spent on the trigger warning screen (experienced a Criterion A event = Yes: *n* = 151, *M* = 1830.95, *SD* = 950.91, No: *n* = 48, *M* = 1600.39, *SD* = 827.97, Welch's *t* (89.73) = 1.62, *p* = .109, *d* = 0.26, 95%CI [-0.07,0.58]). Curiously, the average time spent waiting on the control screen was higher for participants who had experienced a Criterion A event (*n* = 151, *M* = 1821.75, *SD* = 983.87) vs. who had not (*n* = 48, *M* = 1556.81, *SD* = 708.97, Welch's *t* (109.32) = 2.04, *p* = .044, *d* = 0.31, [-0.02, 0.64]).

Number of images that participants covered, time spent viewing images before they were covered, and emotional responses and individual difference factors. Because only 25 participants used the cover image option, the correlations between the number of images participants covered, time spent viewing images before they were covered, and emotional responses and individual difference factors are drastically underpowered. However, for completeness we report them in Tables 6.1 and 6.2 and describe them in detail here: https://osf.io/tgk7n/

[31] e.g., from the questions participants were asked: *"finding personal meaning in the event"*, *"thinking of different ways to deal with the outcomes of the event"*,*"trying to see the good side of the situation"* etc.

Discussion

Advocates (e.g., Cripps, 2020) and critics (e.g., Lukianoff & Haidt, 2015) of trigger warnings both claim that trigger warnings promote avoidance behaviors. However, we found that apart from a minor increase in avoidance when a warning appeared in the first few trials, undergraduates did not passively (by remaining on an instruction screen) or actively (by covering images) avoid negative stimuli prefaced with a trigger warning any more than stimuli without a warning. In fact, participants were reluctant overall to avoid distressing images; only 12.56% (n = 25) participants used the option to cover such images when given the opportunity to do so in the image viewing task. Put differently, when warned about and given the chance to avoid material that reminds them of a negative experience (here operationalized as exposure to a trauma film), participants are extraordinarily reluctant to do so. Furthermore, we did not find any relationships between avoidance behaviors and emotional responses throughout the study (e.g., state anxiety) or with individual difference characteristics (e.g., trait anxiety)—including for participants with prior experience of the topic of the trauma film and/or exposure to a Criterion A trauma.

Our finding that trigger warnings do not seem to *enhance* avoidance behaviors towards distressing stimuli fits with emerging research (Bruce & Roberts, 2020; Bridgland, Barnard, & Takarangi, 2021; Bridgland, Bellet, & Takarangi, 2021; Gainsburg & Earl, 2018; Kimble et al., 2021). Our results build on prior research, by showing this pattern for content directly related to a negative experience. Previous trigger warning research has examined what happens when someone views novel distressing stimuli, with no way of knowing if this material reminds the person of their past experience. Here we addressed this limitation and found that participants *did not* avoid more stimuli related to an analogue trauma event when they saw a trigger warning message versus a control message. In fact, despite reporting that they found the film highly distressing, participants *rarely* avoided images at all in the image viewing phase—only *one* participant used the "cover photo" feature for all 32 images. Our results may also indicate that trigger warnings are linked to avoidance behaviors, but only in a minority of people already prone to avoidance. That is, it seems likely that people who use avoidance as a way to cope with distressing material may be as likely to do so when seeing content presented with a warning as when seeing content without a warning. Nevertheless, converging evidence from multiple experimental paradigms demonstrates trigger warnings do not seem to *enhance* the avoidance of potentially negative material.

However, there are at least two alternative possibilities for the low rate of avoidance behaviors we observed. First, recall that our cover story was that participants were completing a juror decision making task and that we falsely told them that they would complete a memory test for the images. It is possible that the desire to perform well on a memory test drove participants to want to view the images for the full time those images were displayed on screen. However, providing participants with a reason to view the images mimics people's motivation for consuming negative or potentially triggering media in a real-world setting (e.g., lecture material required for a course, entertainment via watching a television show). Therefore, avoidance behaviors across the study may have been low because of a motivation to view the images, but this is likely reflective of real-world conditions where a desire to avoid distressing content competes with other reasons to consume it. Of course, this cover story was also integral to justify the task to participants and prevent participants from responding differently to warning versus control trials based on what they may have guessed about the expectations of the experimenter or study aims. Second, recall that state anxiety decreased significantly from before to after the image viewing task. It is possible that the image viewing task was not as distressing as viewing the film itself and therefore participants felt that they did not need to avoid the images. Along the same lines, since participants were already in a negative mood state (due to watching the film) it is possible that they did not perceive any further emotional harm could come to them by viewing film image stills. This possibility is supported anecdotally by participant comments during debriefing and carries some worrying implications. Trigger warnings are typically for "vulnerable" people such as people with poor mental health (e.g., low-wellbeing, depression etc.; Mosseri, 2019) who are likely in negative mood states. These people may therefore be *even less* likely (versus non-vulnerable people) to use trigger warnings as a tool for avoidance if they believe that what they view is unlikely to have any further dampening effect on their mood. Future research could address the possibility by allowing sufficient time for mood to return to normal before participants complete the photo viewing task or by manipulating mood to examine if it affects avoidance rates.

Aside from complete avoidance, trigger warning advocates also claim that warnings help people *pause* and emotionally *prepare* to cope with material. However, participants did not spend longer on the trigger warning screens versus the control screens. In fact, within the first two trials (which were always a warning and control screen) we found that participants spent more time waiting on the *control* screen. One potential explanation for this finding is our warning screens created a "Forbidden Fruit effect" (Ringold, 2002), encouraging participants' morbid curiosity about distressing content (Oosterwijk, 2017), and therefore leading them to engage with the covered content more quickly.

However, we did find that participants' belief that warning messages about the images would prevent them from causing distress was positively associated with spending more time, on average, on the message screens (both overall and by screen type). Thus, it is possible that general beliefs about the usefulness of warning messages (i.e., as ways to prepare for upcoming content) extended to *both* screen types and led these participants to linger longer on each screen as a way to emotionally prepare/brace themselves to view the next image. Yet there was no evidence that time spent on the message screens was associated with reduced distress throughout the study—which is what we would have expected if participants were spending their time "pausing" on message screens to emotionally prepare themselves to view content. Therefore, even if participants prepare in a physical *pausing* sense, that preparation appears unhelpful.

Overall, one implication of these results—consistent with prior work on the effects of trigger warnings on emotional reactions—could be that trigger warnings are inert or trivial (Sanson et al., 2019). That is, trigger warnings do not do any overt good, but they also do not seem to lead to any obvious harm either. Therefore, if people use warnings because they believe warnings are helpful, they will not experience increased harm as a result. Alternatively, trigger warnings *could* lead to harm if they are the only mental health safeguard or support policy employed in various domains. Indeed, there is a continued push to promote the use of trigger warnings despite mounting evidence that they are ineffective in reducing emotional reactions or promoting avoidance behaviors. For instance, Strothman (2021) states that even though research suggests warnings are not helpful, she *"feel[s] strongly that we need to keep using them*." This continued belief in the illusory of trigger warnings could result in two potential harms. First, for individuals trigger warnings may become a "box-ticking exercise" (Hay, 2019) or a "sticker-fix" (Fagan, 2019). That is, some people may think that adding a trigger warning to their content—whether that be a university lecture or a social media post—might absolve them of making any other efforts to present distressing material in a conscientious way. On a more macro level, the continued beliefs about the benefits of trigger warnings could result in reduced efforts by policy makers or institutions to find efficacious mental health support strategies, because trigger warnings may be considered one such approach already in use.

Our research has several limitations. First, of course the trauma analogue paradigm differs from autobiographical trauma. Yet research examining how participants respond to stimuli that match an autobiographical trauma has also found trigger warnings do not prompt avoidance behaviors (i.e., Bruce & Roberts, 2020; Bridgland, Barnard, & Takarangi, 2021) or ameliorate negative emotional reactions (i.e., Jones, Bellet, & McNally, 2019) in response to the stimuli. Moreover, it is important to triangulate findings (i.e., use multimethod approaches) to reduce the possibility that a set of results can be attributed to specific methodological approaches or data sources (Lin, Werner, & Inzlicht, 2021). That is, our results only contribute one piece to the puzzle; future research is necessary to expand our knowledge on the ways that trigger warnings may or may not affect avoidance behaviors and thus allow us to reach convergent and valid conclusions. For instance, one future direction could be to include a ‘no video’ or ‘neutral video’ control condition. Such a design would test whether experiencing a trauma changed how people respond to trigger warnings about that type of trauma (i.e., video condition), versus people who had not experienced that event (i.e., no/neutral video condition).

Second, because the rates of avoidance in our study were extraordinarily low, our analyses for some avoidance variables (i.e., rates of covering photos and also time spent on photos before they were covered) were underpowered. Therefore, we were not able to comprehensively test some of our hypotheses related to these avoidance variables. However, if the overall rate of avoidance is so low, it seems unlikely that avoidance is a widespread behavior in the general population. Perhaps we should be more concerned with finding out why people seem so eager to approach negative content and the implications of this behavior, rather than on the minority of people who seem to want to avoid this material.

Third, we used a university undergraduate sample—meaning that our results may not generalize to other populations that vary in sociodemographic variables (e.g., age, education, socioeconomic status etc.). Indeed, because trigger warnings are primarily intended for people with clinical levels of symptoms, it is possible that avoidance behaviors might change if we specifically recruited specific clinical populations (e.g., people with a clinical diagnosis of PTSD). However, bearing this limitation in mind, many universities now mandate the use of trigger warnings as part of mental health initiatives (e.g., Harris, 2016; Palmer, 2017) and at least one third of first-year students around the world screen positive to an anxiety, mood, or substance abuse disorder (DSM–IV; Auerbach et al., 2018). Therefore, the use of an undergraduate university population is appropriate when evaluating the reported efficacy of trigger warning messages.

Taken together, contrary to the claims of advocates and the concerns of critics, we did not find evidence that trigger warnings significantly enhance avoidance behaviors in undergraduates—apart from a *minor* increase when a warning appeared in the first few trials.. Furthermore, we did not find any indication that trigger warning messages help people to pause and emotionally prepare themselves to view negative content. Future work should focus on developing strategies that fulfil the purported aims of trigger warning messages.

7 The effects of trigger warnings on avoidance behaviors in an applied context: Investigating Instagram's Sensitive Content Screens

Manuscript under review at *Clinical Psychological Science* as:

Bridgland, V. M. E., Bellet, W. B., & Takarangi, K.T. M. (2021). Curiosity disturbed the cat: Instagram's Sensitive Content Screens do not deter vulnerable users from viewing distressing content.

Author Contributions: I developed the study design with the guidance of MKTT. I collected the data, and performed the data analysis and interpretation, and drafted the manuscript. MKTT and BWB made critical revisions to the manuscript. All authors approved the final version of the manuscript for submission.

Abstract

In an attempt to mitigate the negative impact of graphic online imagery, Instagram has introduced *sensitivity screens*—graphic images are obfuscated with a blur and accompanied by a warning. Sensitivity screens purportedly allow "vulnerable people" with mental health concerns, to *avoid* potentially distressing content. However, no research has assessed whether or not sensitivity screens operate as intended. Here we examined whether people, including "vulnerable users" (operationalized as people with more severe psychopathological symptoms, e.g., of depression), use the sensitivity screens as a tool for avoidance. In two studies we found that the majority of participants (80-85%) indicated a desire (Study 5a) or made a choice (Study 5b) to uncover a screened image. Furthermore, we found no evidence that "vulnerable users" were any more likely to use the screens to avoid sensitive content. Therefore, warning screens appear to be an ineffective way to deter vulnerable users from viewing negative content.

Introduction

In 2017, Instagram’s mental health policies were thrust into the public spotlight when details emerged about the platform’s alleged role in the suicide of 14-year-old Molly Russell. A recent inquest revealed that the social media posts Molly viewed before she took her own life—content relating to anxiety, depression, self-harm, and suicide—were too graphic even for police and lawyers to view for long periods of time. In response to the ongoing investigation of social media’s alleged role in Molly’s death, Instagram has made a number of changes to “support and protect the most vulnerable people” (Mosseri, 2019a). As well as completely removing content related to self-harm, part of Instagram’s mental health initiative involves adding sensitivity screens, in which images are obfuscated with a blur and accompanied by a warning: “Sensitive Content: This photo may contain graphic or violent content.” The primary purpose of these screens is to reduce “surprising or unwanted experiences” and allow people, and in particular “vulnerable people” with mental health concerns, to avoid potentially distressing content. That is, although avoidance is generally considered a maladaptive coping response (Littleton, Horsley, John, and Nelson, 2007), Instagram claims that minimizing exposure to negative content via sensitivity screens helps to preserve mental health (Mosseri, 2019b). However, there is currently no research assessing whether or not sensitivity screens operate as intended. To address this gap in knowledge, we examined whether people, including vulnerable users, use the sensitivity screens to minimize their exposure to negative content. First, we investigated whether sensitivity screens are helpful in deterring people from viewing potentially negative content. Second, we examined how vulnerability variables (operationalized as risk markers for psychopathology such as depression) relate to the success or failure of deterrence.

Traditional trigger warnings—alerts that upcoming material may be offensive or distressing—are mostly limited to simple lines of text (e.g., "This article may contain themes related to sexual abuse") presented before various types of media (e.g., news, social media, film/television, lectures). However, new policies on social media are primarily focused on censoring *visual content* via an image processing technique called a Gaussian Blur, which reduces image noise and detail (see *Fig. 1*). Although here we focus on Instagram, many other platforms, such as Facebook, Twitter, Reddit, and Buzzfeed, use similar sensitivity screens. It is thus surprising that no research has investigated sensitivity screens or the use of trigger warnings in a social media context. However, research on traditional trigger warnings has found that at best, warnings appear to have little effect on people's reactions towards material (Bellet, Jones, Meyersburg, & McNally, 2019; Boysen, Issacs, Tretter, & Markow, 2019; Bridgland et al., 2019; Sanson, Strange & Garry, 2019). At worst, trigger warnings create anticipatory anxiety prior to viewing content (e.g., by increasing anxiety; Bridgland et al., 2019; Gainsburg & Earl, 2018) and in some cases increase perceptions of harm caused by the material (Bellet, Jones, & McNally, 2018). In fact, early research shows trigger warnings may be the most deleterious for the very people they are intended to protect. For example, Jones, Bellet, and McNally (2019) found that trauma survivors reported that their trauma was more central to their identity after reading distressing text passages marked with a trigger warning (versus unwarned). Event centrality—the belief that a traumatic event marks a turning point in one's life story—is associated with PTSD symptoms (Berntsen, & Rubin, 2006), and prospectively predicts more severe PTSD (Boals & Ruggero, 2016). Moreover, Bridgland and Takarangi (2020) found that warning messages prolonged the negative characteristics (e.g., PTSD-like symptoms) associated with recalling a negative memory over time.

While trigger warnings have trivial effects on responses to potentially distressing material at best, the primary purpose of sensitivity screens is to allow people who may have mental health vulnerabilities to *avoid* or minimize exposure to potentially distressing content. Therefore, we must first consider whether there is any evidence that such warning methods actually deter people from approaching potentially negative material. Second, we need to examine whether it is likely that "vulnerable people" (i.e., those with symptoms of mental disorder or risk factors for the same) more specifically will use trigger warnings to minimize their exposure to potentially negative content.

Only a handful of studies have focused on how trigger warnings may affect avoidance behavior, with mixed findings. In Bridgland, Barnard, and Takarangi (2021), participants reported they would avoid content related to a stressful/traumatic experience that was accompanied by a trigger warning to the same degree as content with no warning (Φ = .08). Similarly, in Bruce (2020), members of the general population and trauma survivors showed equal preference for news articles labelled with or without a trigger warning. Finally, Gainsburg and Earl (2018) found that participants were no less likely to select a film title for subsequent viewing when the title was accompanied by a trigger warning (versus no warning). Therefore, early evidence focusing specifically on trigger warnings suggests that sensitivity screens may not deter users from consuming negative content.

However, research on warnings in other domains shows that warnings can produce behavior that is the opposite of what is intended. In short, when people's freedom to engage in an experience is restricted, that experience often becomes more attractive (Ringold, 2002). This phenomenon is known as the "forbidden fruit effect" and there is a substantive supporting literature. For example, viewing more advertisements warning of the dangers of smoking was positively correlated with stronger approval of smoking and intentions to smoke (Wakefield et al., 2006), and viewing a high-threat (vs. low threat) warnings about social media censorship led to stronger feelings of aggression and support of social protests (Ng, Kermani, & Lalonde, 2021). Similar patterns have been observed for warnings on violent television shows (Bushman & Stack, 1996), and video games (Bijvank, Konijn, Bushman, & Roelofsma, 2009). Therefore, it is possible that sensitivity screens make viewing images *more* attractive, making avoidance of negative material unlikely.

A closely related finding known as the "Pandora effect" also suggests that people often *do not avoid* potentially aversive stimuli. In fact, people may be *more* likely to engage with stimuli if the consequences of such engagement are uncertain and negative in nature (Hsee & Ruan, 2016). In one series of experiments, participants were more likely to expose themselves to uncertain negative outcomes (e.g., electric shocks and unpleasant sounds) than to certain neutral or certain negative outcomes (Hsee & Ruan, 2016). Participants were also more likely to uncover a masked image of a disgusting insect—a choice similar in nature to that presented by a sensitivity screen—if the outcome was uncertain (marked with a question mark) rather than when the mask included a label of what the image contained (e.g., "mosquito"; Hsee & Ruan, 2016). Similarly, Oosterwijk (2017) found that participants deliberately chose to view images portraying death, violence and harm over non-negative alternatives. One explanation for these results is that people are driven by *morbid curiosity* to close an information gap and acquire information about the world (Loewenstein, 1994). This drive to acquire information may be particularly strong for negative information because negative information is typically *uniquely* negative (e.g., deviations from social norms) and thus represents a strong gain in information, unlike positive information, which is mostly alike in that it conforms to socially constructed norms of positivity (Kashdan & Silvia, 2009). A second, more parsimonious explanation is that people may be driven by a desire to resolve curiosity and uncertainty and therefore sometimes seek unhelpful negative information that provides no long-term pleasure, benefits, or gains (Hsee & Ruan, 2016). Because sensitivity screens do not provide any information about the kind of content that is blurred, they foster uncertainty. Further, the accompanying warning message informs the viewer that the content will be negative. Thus, it is possible that, due to the "Pandora effect," these screens do not deter users.

Based on previous research, it seems likely that sensitivity screens will not deter users from consuming negative material, and may instead even increase users' attraction to the material. However, several lines of research also suggest that sensitivity screens may be even less likely to deter *vulnerable people* from consuming negative content—the very people Instagram is trying to protect. For example, people with prior lifetime exposure to violence, and fear of future terrorism, are more likely to seek out and watch disturbing content online (Redmond, Jones, Holman, & Silver, 2019). Recent research also suggests that some trauma survivors engage in "self-triggering" behaviors, i.e., seeking reminders of their traumatic experience (e.g., graphic imagery and media; Bellet, Jones, & McNally, 2020). Similarly, people with or at risk of depression (vs. healthy controls) have difficulty disengaging attention from negative material that has captured their attention and are more likely to use emotion-regulation strategies to maintain or increase negative mood states—for instance, by choosing to expose themselves to negative rather than positive imagery (Millgram, Joormann, Huppert, & Tamir, 2015).

There are several theoretical perspectives that may help us to understand why people with mental health vulnerabilities may be attracted, rather than deterred by, warnings. First, vulnerable users may be troubled by the uncertain nature of their experiences and symptoms. Thus, they may be motivated to justify or make meaning of their experiences by seeking information related to such experiences (Hogan & Brashers, 2013). Indeed, the desire to make meaning of a traumatic experience was the best predictor of how often participants self-triggered (Bellet et al., 2020). Second, in line with Zillmann's (1988) Mood Management Theory, we know that people often use media to regulate mood. Although we might expect that people would typically select positive media to repair negative mood, people may instead seek other emotional goals beyond immediate mood repair and engage in "counter-hedonistic" consumption behavior. For instance, clinically depressed people (vs. non-depressed) are more likely to use emotion regulation strategies to maintain or increase their level of sadness rather than to alleviate it (Millgram, Joormann, Huppert, & Tamir, 2015), perhaps because sad moods are familiar to people with depression. Therefore, it is possible that people with a tendency towards negative mood states—perhaps due to depression or low wellbeing—or with a desire to make meaning about one's circumstances, would be more likely to uncover screened images.

Third, although approaching aversive content may seem like the opposite of avoidance behavior, it may constitute experiential avoidance. That is to say, unwillingness to remain in contact with private experiences (e.g., feelings of anxiety due to uncertainty) results in behaviors intended to reduce these experiences (Rains & Tukachinsky, 2014). Indeed, it is well documented that people with a range of mental health concerns (e.g., anxiety disorders and depression) also report higher intolerance of uncertainty—a characteristic relating to negative beliefs about uncertainty and its implications (Carleton, 2012). Thus, sensitivity screens may make people especially sensitive to the anxious state created by "the unknown" and increase their desire to uncover screened content.

Taken together, past research suggests that sensitivity screens may not be effective in deterring users—including vulnerable users—from consuming negative content, or may increase the attractiveness of negative content. However, no research has investigated how people respond to sensitivity screens or trigger warnings in a social media context. The present study investigated how participants interact with sensitivity screens in two ways. In Study 5a, we asked participants how likely they would be to uncover a blurred image if they came across it on Instagram and measured a series of factors covering psychopathology and psychological vulnerability variables (i.e., depression, anxiety and stress, PTSD symptomology, general wellbeing, trauma history, centrality of traumatic event to identity, and treatment seeking behaviors). In Study 5b, we presented participants with a mock Instagram photo viewing task where participants had the option to click to uncover (‘see photo’) a single blurred image or select ‘next photo’ to skip uncovering the image. We also included additional measures of wellbeing in Study 5b in order to further explore the association between uncovering behavior and these variables as well as replicating our main findings.

Based on previous literature and psychological theory, we predicted that sensitivity screens would not be effective in deterring the majority of people from desiring to view or deciding to view negative content. Furthermore, we also predicted that sensitivity screens would be even less effective in deterring vulnerable users (e.g., people with mental health vulnerabilities) from consuming negative content than less vulnerable users.

Study 5a

Method

We preregistered this study (https://osf.io/m6d9g). The data we report here were part of a larger project that also investigated the desire for news filtering systems. Study 5a was approved by the Flinders University Social and Behavioral Research Ethics Committee. The data, supplementary files, and materials can be found at: https://osf.io/rj987/. We have reported all measures, conditions, and data exclusions.

Participants

Participants were recruited online through Amazon's Mechanical Turk (MTurk) and received $2.50 USD. The study was open to respondents above 18 years of age who were located in the United States. Because we only wanted to recruit Instagram users, participants who indicated that they do not use Instagram at the beginning of the survey were screened out.[32] We excluded 13 participants for failing an attention check. For the magnitude of a correlation to be deemed stable, the typical sample size should approach 260 (Schönbrodt & Perugini, 2013; Schönbrodt & Perugini, 2018). Therefore, we used a power-based stopping criterion and collected *N* = 260 participants after exclusions.

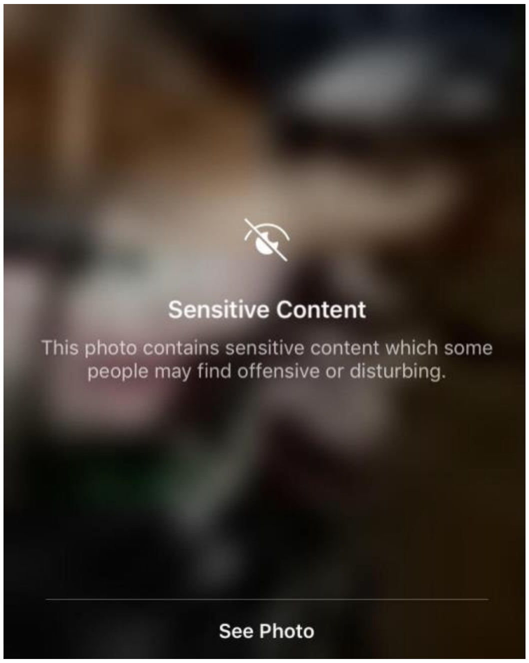
Participants ranged from 20-71 years (*M* = 36.0, *SD* = 10.69) and were more likely to be female (54.2%; 45% male and 0.4% preferred not to specify sex). Our sample was predominantly White/Caucasian (63.8%); others were of African American (14.2%), Asian (7.3%), and Latinx (4.2%), or other (5%; e.g., mixed race/bi-race) descent, while 5.4% of participants specified nationality (e.g., American/USA). The majority of participants (55.8%) reported an income between $45,000-$140,000, and were predominately (58.8%) college graduates.[33]

[32] 155 participants completed both the Instagram screen questions and the news filter questions.
[33] Full demographics: https://osf.io/acpeu/

Measures

Social media/news media use. We asked participants to indicate (from a list) which social media sites they use on a regular basis. We also asked participants to indicate: how many days of the last 7 days (Never, 1 Day, 2 Days...Everyday), and for how many hours each day (I don't use, less than half an hour, 1 hour, 2-3 hours, 4-5 hours, more than 6), they used social media.

Instagram sensitivity screens. Participants were presented with one example of a real Instagram sensitivity screen (from a pool of six examples) taken from the site (Figure 1) and were told 'Imagine you are scrolling (i.e., browsing) through Instagram posts and come across the following image.' Participants were then asked (a) 'Would you click to uncover this image?' (1 = *definitely no,* 6 = *definitely yes*), (b) 'What factors would affect whether you would uncover the image?' (open box response), (c) 'Have you seen these screens on Instagram?' (*Yes*/*No*). If participants answered *Yes,* they were asked: 'When you have seen the screens, do you typically click to uncover and see the image?' (1 = *Never*, 6 = *Always*). Finally, participants were asked (d) 'Would you turn off the sensitivity screen feature (i.e., meaning that all photos would not be screened when browsing through Instagram) if you had the option to do so?' (*Yes*/*No*).



Sensitive Content

This photo contains sensitive content which some people may find offensive or disturbing.

Figure 1. Example of a real Instagram sensitive content screen used in Study 5a[34]

Depression Anxiety Stress Scales-21 (DASS-21; Lovibond, & Lovibond, 1995; Appendix U). The DASS-21 is a self-report instrument measuring the severity of depression (present study; α = .95), anxiety (α = 0.88) and stress (α = 0.91) in the past week. The scales demonstrate convergent validity with other well-validated measures of depression and anxiety (Antony, Bieling, Cox, Enns, & Swinson, 1998).

The Scales of General Well-Being short form (SGWB-14; Longo, Coyne, & Joseph, 2018; Appendix V). The SGBW-14 is a brief assessment measuring 14 dimensions of well-being (present study; α = .96). The scales demonstrate convergent validity with other validated measures tapping various aspects of well-being (Longo, et al., 2018).

[34] Study 1 used the sensitive screen warning worded as pictured here. Instagram subsequently changed the warning text to “Sensitive Content: This photo may contain graphic or violent content,” which we used in Study 2. However, in a separate experiment (see https://osf.io/2fdr7 for details) we found no difference between the two warning types on uncovering behavior. Thus, this change in wording is unlikely to have had a meaningful effect on our results.

Trauma History Screen (THS; Carlson, Smith, Palmieri, Dalenberg, Ruzek & Kimerling, 2011; Appendix k). The THS is a brief questionnaire that measures exposure to high magnitude stressor (HMS) events (sudden events that cause extreme distress in most people exposed) and events associated with posttraumatic distress. The THS asks participants to respond "YES" or "NO" to a list of 14 stressful events (e.g., A really bad car, boat, train, or airplane accident). If a participant answers "yes", they are asked to indicate how many times that event has happened. Participants are then asked to indicate if any of the events bothered them emotionally, and, if so, they were asked to describe (in one or two sentences) the event that bothered them the most. If they responded no, or had not experienced any of the events, they were asked to identify and describe (in one or two sentences) the most stressful experience of their life. Participants were told they would refer back to their identified event later in subsequent survey questions and tasks. All participants were then asked to provide: their age at the time of the event; whether anyone was hurt or killed (*yes/no*); whether they felt afraid, helpless or horrified (*yes/no*); how long they were bothered by it (1 = *not at all*; 4 = *a month or more*); and how much it bothered them emotionally (1 = *not at all*; 5 = *very much*).[35] The THS has been validated for use in both clinical and non-clinical populations and has excellent psychometric properties; high reliability (*r* = .93 for HMS in clinical samples and *r* = .74-.87 for non-clinical samples) and correlates strongly (*r* = .73-.76) with more detailed trauma exposure measures (i.e., the Traumatic Life Events Questionnaire; Carlson et al., 2011).

Posttraumatic Stress Disorder Checklist (PCL-5; Bovin et al., 2016; Appendix L). The PCL-5 is a self-report measure that corresponds to the DSM-5 symptom criteria for PTSD. Participants were asked to indicate in relation to their most stressful/traumatic event—identified

[35] These data are not reported here. See https://osf.io/rj987/

on the THS—(on a scale of 0 = *not at all* to 4 = *extremely*) how bothered they were by a list of symptoms over the past month (e.g., repeated, disturbing dreams of the stressful experience). The PCL-5 has excellent psychometric properties (present study; α = .96), test-retest reliability (r = .84) and convergent and discriminant validity (See Bovin et al., 2016).

Centrality of Events Scale, 7-item version (CES-7; Berntsen, & Rubin, 2006; Appendix H). The CES-7 measures the centrality of a negative event to a person's identity and life story. Participants were asked to think of the most stressful/traumatic event we asked them to identify and answer a series of questions (e.g., 'I feel that this event has become part of my identity'; on a scale from 1 = *totally disagree* to 5 = *totally agree;* present study; α = .93). The scale correlates highly with the full 20-item version (r = .96) as well as displaying a robust association with PTSD symptom severity (r = .37; Berntsen, & Rubin, 2006).

The Self-Triggering Questionnaire (STQ; Bellet, et al., 2020; Appendix W). We piped back participants' most stressful/traumatic event text response from the THS and asked if they have ever self-triggered with reminders of this event (*Yes/No*). If participants answered "Yes", we asked them to indicate the frequency of these behaviors, their motives for self-triggering, and their methods of self-triggering. If participants answered "No", we asked if they had ever self-triggered in regard to any other stressful/traumatic event, (*Yes/No*) and if they answered "Yes", we asked them to describe this event, and to indicate the frequency, motives, and methods for these self-triggering behaviors. We combined these categories of respondents together to form two final categories: those who had self-triggered in reference to either their most stressful/traumatic event or other stressful/traumatic event, and those who had not self-triggered at all.[36]

[36] Analyses involving self-triggering frequency, methods, and motives are not reported here. See https://osf.io/acpeu/

Treatment-seeking behaviors (Appendix X). This questionnaire comprises items from the past help seeking section of the General Help-Seeking Questionnaire (GHSQ, items 2-4; Wilson, Deane, Ciarochi & Rickwood, 2005) and the Actual Help-Seeking Questionnaire (AHSQ, item 5; Rickwood & Braithwaite, 1994) and Eisenberg, Downs, Golberstein, and Zivin (2009). Participants were asked to indicate; if they have taken any medication, have seen a health professional, or sought help from a source other than a professional—in the last 6 months—to help with a personal problem.

Procedure

Participants were required to pass a Qualtrics V2 Captcha as well as correctly answer 8/10 English proficiency questions to enter the survey. We told participants the study was investigating engagement, personality, and life experience. Participants answered demographic questions, indicated which social media sites they use, and completed the sensitive screen task. Next, participants indicated the frequency of their social media use, the Trauma History Screen (THS), PTSD symptomology (PCL), the centrality of their most stressful/traumatic event they identified during the THS (CES), and questions on self-triggering. Participants then answered questions about depression, anxiety, and stress symptoms (DASS), wellbeing (SGWB-14), and individual difference characteristics[37] in a randomized order. Finally, participants were asked about their beliefs about trigger warnings, whether they left the task for any period of time (if they answered "Yes", they were then asked when and for how long they had left), and whether they had any technical issues. Participants were then fully debriefed.

[37] These data were secondary to our main research aims in this study (i.e., which focused on how "vulnerable users" interact with sensitivity screens) and are not reported here. See https://osf.io/mjrq8/

Results and Discussion

Statistical overview

We ran analyses using Null-Hypothesis Significance Tests ($\alpha = .05$) in SPSS Version 25 and JASP for MacOS version 0.13.1. Because “vulnerability” is not a unitary construct, and comprises many different types of psychopathologies and thinking styles (e.g., depression, anxiety, PTSD, etc.), each one of our dependent variables relates to its own hypothesis. In such a case, it is not necessary to correct for multiple comparisons across dependent variables because they are not from the same family of tests/hypothesis (Rubin, 2021). Such an approach constitutes an “individual testing” approach, in which controlling for the family-wise error rate is contraindicated (Rubin, 2021).

Participant characteristics

Because sensitive content screens are intended for vulnerable populations, we examined our sample for prevalence of traumatic event exposure, possible PTSD, and Depression, Anxiety and Stress severity. Overall, 87.7% of participants reported experiencing one or more HMS events, and 68.1% of participants reported a Criterion A event (actual or threatened death or injury; Carlson et al., 2011). The most common events reported were the sudden death of a close family member or friend (61.9%), followed by exposure to a hurricane, flood, earthquake, tornado, or fire (38.8%), a really bad car, boat, train, or airplane accident (31.9%). Further, 24.6% of the sample met criteria for probable PTSD according to the conservative cut-off (sum score > 33; Bovin et al., 2016). For Depression, 51.5% of our participants were in the normal range; 25.7% mild-moderate; 22.7% severe-extremely severe; Anxiety: 53.8% normal; 23.9% mild-moderate; 22.4% severe-extremely severe; Stress; 61.2% normal; 22.3% mild-moderate; 16.6% severe-extremely severe (Depression, Anxiety and Stress Scale [DASS-21] manual cutoffs). Most participants (85.8%) reported that they used social media every day in the last 7 days (followed by 5 days = 5%, 6 days = 4.2%, 3 days = 1.2%, 2 days or less = 0.8%) for an hour or more per day (2-3 hours per day = 35%, 1 hour = 27.3%, more than 6 hours = 15.8%, 4-5 hours = 12.7%, less than half an hour = 9.2%).

Desire to uncover sensitive content screens and prior experience with sensitive screens on Instagram.

We had asked participants if they would click to uncover a sensitive content screen (1 = *definitely n*o, 6 = *definitely yes*); on average, participants indicated a clear desire to uncover (*M* = 4.56, *SD* = 1.52). We also dichotomized participants' answers as '*no*' (responses 1 through 3) or '*yes*' (responses 4 through 6); the majority (80%) of participants fell into the '*yes*' or 'uncover' category.

Aside from asking participants about hypothetically encountering a sensitivity screen, we also asked them about encounters and interactions with sensitivity screens in real life. Over half our participants (53.8%) indicated that they had previously seen a sensitive content screen on Instagram. Participants who said they have seen the screens on Instagram reported that they almost always (*M* = 4.41, *SD* = 1.49; on a scale of 1 = *Never*, 6 = *Always*) uncover a screened image if they come across one. Finally, 51.5% of participants said they would like to be able to turn off the sensitive screen feature (so that all photos were not screened when browsing) if they had the option to do so.

Thus, sensitive screens do not appear effective in deterring the majority of people from approaching potentially negative content. Next, we explored participants' qualitative responses to help us understand why. We coded participants' text responses to the question 'What factors would affect whether you would uncover the image?' (Table 7.1) using the thematic analysis technique described by Braun and Clarke (2006): data are coded and labeled according to overarching themes identified across the dataset. Over one third of participants (35.8%) indicated that they simply wanted to see the image/picture, and of these, 75.3% (26.9% of our total sample) specifically mentioned they would uncover the image because of reasons related to curiosity or related concepts such as intrigue. Over one third (36.2%) of participants indicated they would decide whether to uncover based on the context of the photo, such as who posted the photo or what the caption/description of the image was. We did not include contextual features such as captions, comments, or the posting account because we wanted to know how people react to sensitivity screens independent of these factors. But future research should manipulate these contextual factors to determine how they may reduce or increase the desire to view sensitive content. Other popular reasons for uncovering/keeping the image covered included the type of content participants believed may be under the sensitive screen (14.6%; e.g., nudity or gore), participants' physical surroundings (9.6%), such their location (e.g., at work) and who was present (e.g., children), their current mood (6.9%), and whether they thought the content was something they would not want to see (5.4%).

Table 7.1

Coded text responses to the question 'What factors would affect whether you would uncover the image?' for Study 5a and 5b

Study 5a		Study 5b			
Category		Percentage (*n*)	Category		Percentage (*n*)
Simply "Would want to see picture" or for more specific reason:		35.8% (93)	Simply "Wanted to see the picture" or for more specific reason:		72.5% (190)
	Curiosity/intrigue (specific mention)	26.9% (70)		Curiosity/intrigue (specific mention)	46.2% (121)
	Depend on interest level in the image/at the time	6.2% (16)		Interested in seeing image	4.2% (11)
	Want to see *why* an image is covered	3.10% (8)		Want to see *why* the image is covered	3.8% (10)
Context provided (e.g., posting account/comments/caption)		36.2% (94)	Did not want to see something negative		12.6% (33)
Type of content expected would influence choice (e.g., nudity, gore, violence etc.)		14.6% (38)	Personality traits (e.g., cite general tendency to cope/not cope with sensitive content)		10.7% (28)
Physical location/other people present		9.6% (25)	Type of content expected would influence choice (e.g., nudity, gore, violence etc.)		8.4% (22)
Mood		6.9% (18)	Uncertainty		2.7% (7)
If they believe it would be something they did not want to see/ something negative		5.4% (14)	Did not expect negative content on Instagram/in the study		2.3% (6)
Typically would uncover/would always uncover		2.7% (7)	Typically would uncover/would always uncover		1.9% (5)
Typically would not uncover/would never uncover		1.5% (4)	Context provided (e.g., posting account/comments/caption)		1.5% (4)
Internet security concerns		1.2% (3)	Mood		1.5% (4)
"No factor would prevent me"/"none"		1.9% (5)	If they could visually guess what the image was		1.4% (4)
If they could visually guess what the image was		1.2% (3)	Not interested		0.8% (2)
Trust in the warning that it is for one's own good		0.8% (2)	Physical location/other people present		0.8% (2)
Personality traits (e.g., cite general tendency to cope/not cope with sensitive content)		0.8% (2)	Misc (categories with <2 people)/unclassifiable		5.0% (13)
Misc (categories with <2 people)/unclassifiable		4.23% (11)			

Taken together, these data suggest that the primary motivations for deciding to view images are curiosity and the context in which the image is presented.

Is the desire to uncover a sensitive content screen associated with psychological vulnerabilities?

We next turned to our interest in whether particular psychological vulnerabilities are associated with the desire to uncover sensitive images. We correlated participants' reported desire to uncover the Instagram sensitivity screen as measured on the 6-point scale with our continuous measurements of these variables (Table 7.2). We also ran a series of chi-square analyses on the desire to uncover as a dichotomous variable and our categorical dependent variables (Table 7.3). In terms of participant demographics, we found that age was negatively associated with the desire to uncover, while a higher percentage of males (biological sex), compared to females and people who indicated they would prefer not to say their sex (*n* = 1), were more likely to fall into the "Yes"/uncover classification. In terms of vulnerability factors, we found that the depression, stress, and total score on the DASS, the Criterion D (negative cognition/mood) and Criterion E (hyperarousal) subscales of the PCL-5, were positively associated with the desire to uncover the sensitivity screen, while wellbeing was negatively associated. We also found that people who indicated they self-trigger ("Yes") compared to those who do not ("No"), were more likely to fall into the "Yes"/uncover classification.

Table 7.2

Correlations between the desire to uncover and continuous variables

		r	
		Study 5a	Study 5b
Age		-.16**	.004
Social media use (general)		.05	.06
Instagram use		-	.04
DASS	Stress	.12*	.02
	Anxiety	.11	-.002
	Depression	.13*	.03
	Total	.13*	.02
SGWB-14		-.17**	-.06
WHO-5		-	-.04
PCL-5	Criterion B Intrusions	.07	-.05
	Criterion C Avoidance	.06	-.07
	Criterion D Negative Cognition/Mood	.12*	-.03
	Criterion E Hyperarousal	.14*	.004
	Total	.11	-.03
CES		-.01	-

Note: * *p* <.05, ** *p* <.01

Table 7.3

Desire to uncover or keep covered by key[38] *categorical dependent variables*

		Keep covered (*n*)	Uncover (*n*)	$\chi^2(df)$	*p*	φ	Next photo	See photo	$\chi^2(df)$	*p*	φ
Biological sex	Male	12.2% (14)	87.8% (101)	8.33 (2)	.016	N/A	10.2% (10)	89.8% (88)	3.10 (1)	.078	.11
	Female	26.4% (38)	73.61% (106)				18.3% (30)	81.7% (134)			
	Prefer not to say	0.0% (0)	100% (1)				-	-			
PTSD probability	Yes	15.6% (10)	84.4% (54)	1.02 (1)	.314	.06	14.1% (29)	85.9% (177)	1.05(1)	.305	-.06
	No	21.4% (42)	78.6% (154)				19.6% (11)	80.4% (45)			
Criterion A	Yes	16.9% (30)	83.0% (147)	3.22 (1)	.073	.11	13.5% (23)	86.5% (148)	1.27 (1)	.262	.07
	No	26.6% (22)	73.4% (61)				18.7% (17)	81.3% (74)			
Self-trigger	Yes	11.1% (10)	88.9% (80)	6.80 (1)	.009	.16	72.5% (29)	27.5% (11)	1.57(1)	.211	.08
	No	24.7% (42)	75.3% (128)				62.16% (138)	37.84% (84)			
Used medication in last 6 months	Yes	18.6% (11)	81.4% (48)	0.09 (1)	.767	.02	-	-	-	-	-
	No	20.4% (41)	79.6% (160)				-	-	-	-	-
Saw a mental health professional in last 6 months	Yes	20.7% (11)	79.2% (42)	0.02 (1)	.878	.01	-	-	-	-	-
	No	19.8% (41)	80.2% (166)				-	-	-	-	-
Sought other help	Yes	17.6% (13)	82.4% (61)	0.38 (1)	.536	.04	-	-	-	-	-
	No	21.0% (39)	79.0% (147)				-	-	-	-	-

To examine the characteristics that *best* predict uncovering behavior, we ran a binary logistic regression with our significant vulnerability characteristics as covariates and our dichotomized uncovering variable as the dependent variable. First, we checked for evidence of multicollinearity. We ran standard correlations with our vulnerability factor variables to check if any variables correlated more than .70 with one another (as per our pre-registration). No variables correlated more than .70. We also ran a standard linear regression, using the

[38] For analyses of other categorical variables not mentioned here (i.e,. gender, household income, highest level of education) see https://osf.io/acpeu/

dichotomous Instagram uncover variable as our dependent variable, and our vulnerability predictors, to further check multicollinearity parameters. No variables had a tolerance value of less than .1, a VIF value of more than 10, or high variance proportions on the same eigenvalue (Field, 2005), indicating no issue of multicollinearity amongst our predictors. In our main analysis we entered all of our significant vulnerability predictors (DASS total, Wellbeing, PCL total, and Self-triggering (Y/N) in a single step (Table 7.4). We found that our model significantly predicted the desire to uncover (or not uncover) the sensitive screen, χ^2 (4) = 15.24, *p* = .004 (R^2: Hosmer & Lemeshow = .06, Cox & Snell = .06, Nagelkerke = .09). Wellbeing and the tendency to self-trigger with a reminder of their most stressful/traumatic event were statistically significant predictors in the model. This pattern shows that as well-being decreased, the odds of indicating a desire to uncover the sensitive screen increased, and also that people with a tendency to self-trigger (vs. no) were more likely to indicate a desire to uncover the image (and thus fall into the uncover category).

Table 7.4

Logistic regression results for predicting uncovering desire for vulnerability characteristics

Vulnerability characteristics					
		95% CI for exp *b*			
	B (SE)	Lower	exp *b*	Upper	*p*
Included					
Constant	4.00 (1.16)		54.39		<.001
DASS Total	-.01 (0.02)	0.96	0.99	1.03	.621
Wellbeing	-0.04 (0.02)	0.93	0.96	1.00	.028
PCL Total	0.01 (0.01)	0.98	1.01	1.03	.605
Self-triggering (Y/N)	-0.94 (0.40)	0.18	0.39	0.85	.018
Included					
Constant	1.68 (1.53)		5.75		.274

Taken together, our Study 5a findings demonstrate that sensitivity screens do not seem to be effective in deterring the majority of people from desiring to view negative content; the primary motivations for desiring to view images are curiosity and the context in which the image is presented. Furthermore, we found that various psychological vulnerability factors are associated with the desire to approach sensitive content. Therefore, it is likely that sensitivity screens are even less effective in encouraging avoidance behaviors for vulnerable users (e.g., people with mental health vulnerabilities; especially people with lower wellbeing) than non-vulnerable users.

Study 5b

In Study 5b we aimed to replicate and extend the findings of Study 5a. In Study 5a, we measured the intent to uncover sensitivity screens, which we thought might reflect a broad pattern of approach behavior (e.g., what do people typically do at any given time they encounter a sensitivity screen). However, although intentions *generally* map onto future behavior ($r = .53$; Sheeran, 2002), intentions may be inconsistent with actual behavior—the intention-behavior gap (Sheeran & Webb, 2016). Therefore, in Study 5a we presented participants with a mock Instagram photo viewing task where they had the option to click to uncover (‘see photo’) a single blurred image or select ‘next photo’ to skip uncovering the image. This change in procedure allowed us to examine if participants’ hypothetical responses in Study 5a mapped onto a behavioral task that more closely matches Instagram. As well as replicating our main findings, we included additional measures of wellbeing. There are many ways of defining and therefore measuring wellbeing (Dodge, Daly, Huyton, & Sanders, 2012). While the SGWB-14 focuses on 14 aspects of wellbeing (happiness, vitality, calmness, optimism, involvement, self-awareness, self-acceptance, self-worth, competence, development, purpose, significance, self-congruence and connection), the WHO-5 focuses on wellbeing as a single construct: positive wellbeing as a signifier of mental health and absence of mental illness (e.g., depression; Krieger, 2014).

Based on the findings of Study 5a, we predicted that a majority of participants (~80%) would click to uncover the sensitive photo screen. We further predicted that lower levels of wellbeing, and higher levels of posttraumatic stress disorder symptoms, depression, anxiety, and stress symptoms, would be associated with a higher probability of uncovering the sensitive content screen. We also predicted that participants who indicated that they self-trigger with reminders of their most stressful/traumatic event would be more likely to uncover the sensitive screen. Finally, because self-triggering is associated with PTSD severity (Bellet. et al., 2020), we predicted that the relationship between PTSD symptomology and uncovering behavior would be moderated by the self-triggering behavior. That is, we expected that PTSD severity would be more strongly associated with the decision to uncover for those who endorsed self-triggering, versus those who did not.

Method

We preregistered this study (https://osf.io/8n7er). Study 5b was approved by the Flinders University Social and Behavioural Research Ethics Committee. The data, supplementary files, and materials can be found at: https://osf.io/rj987/. We have reported all measures, conditions, and data exclusions.

Participants

Participants were recruited online through MTurk. Participants received a payment of $2.00USD. As in Study 5a, the study was open to respondents above 18 years of age who were located in the United States and participants who indicated that they did not use Instagram at the beginning of the survey were screened out. We excluded one participant who failed all three embedded attention checks (Berinsky, Margolis & Sances, 2014; Hauser & Schwarz, 2015), and eight participants who indicated that they chose to uncover the photo because they believed the behavior was part of task requirements (e.g., a pre-registered requirement because we are seeking to understand uncovering behaviors as those behaviors typically occur on Instagram). In total we collected *N* = 262 participants after exclusions.

Participants ranged from 19-70 years (*M* = 35.68, *SD* = 9.61) and were more often female (62.6%; 37.5% = male). Our sample was predominantly White/Caucasian (65.6%); others were of African American (11.1%), Latinx (8%), and Asian (5.3%) or other (5%; e.g., mixed race/bi-race) descent, while 5% of participants specified nationality (e.g., American/USA). The majority of participants (58.8%) reported an income between $45,000-$140,000, and were predominately (61.8%) college graduates.

Measures

Mock Instagram task. Participants viewed a set of 5 neutral and 5 positive NAPS photos (Marchewka, Żurawski, Jednoróg, & Grabowska, 2014)—randomly selected from one of 14 sets of 10 images—in a random order, with a 'next photo' button to go to the next image. Each image was presented inside an Instagram frame to make it appear as it would on the website (Figure 2). Participants then viewed a single sensitive screen image (a NAPS photo modified to look like an image with a sensitive content overlay)—randomized from a pool of 20 possible images. They had the option to 'see photo'/'uncover photo'[39] or just go to 'next photo'. Participants did not actually see a negative photo—the photo task ended here. Participants were then asked: (a) "Why did you or did you not uncover the screened image?" (open box), (b) 'Have

[39] Participants were randomized to see a 'see photo' or 'uncover photo' button—however rates of selecting 'Next Photo' did not significantly differ per button type (χ^2 (1) = 0.47, *p* = .492), thus, we collapsed our analyses across button type.

you seen these screens on Instagram?' (Yes/No), If Yes: 'When you have seen the screens, do you typically click to uncover and see the image?' (1 = Never, 6 = Always), (c) 'Would you turn off the sensitivity screen feature (i.e., meaning that all photos would not be screened when browsing through Instagram) if you had the option to do so?' (Yes/No).

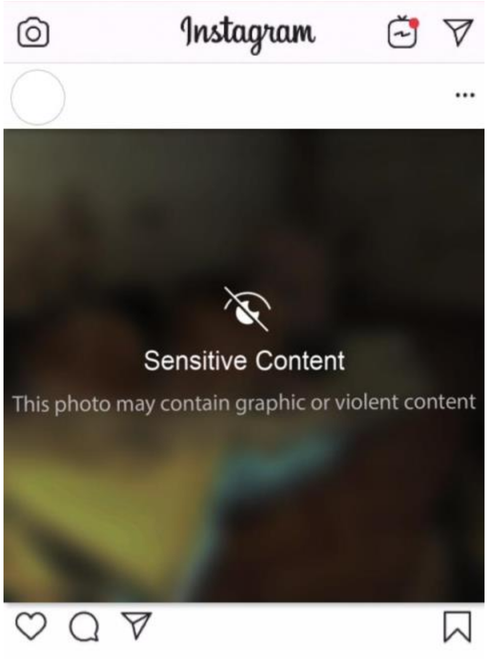


Figure 2. Example of a NAPS photo modified to look like an image with a sensitive content overlay used in Study 5b.

As in Study 5a, participants completed measures of social media use and Instagram specifically, depression, anxiety and stress symptoms (using the DASS-21. Study 2: Depression, α = .93; Anxiety, α = .89; Stress, α = .89), wellbeing (using the SGWB-14: α = .94), the Trauma History Screen, and posttraumatic stress disorder symptoms (using the PCL-5: α = .96). We also measured self-triggering by piping back participants' most stressful/traumatic event text response from the THS and asked if they had ever self-triggered with reminders of this event (Yes/No; i.e., from the STQ). In addition, participants completed the 5-item World Health Organization Well-Being Index (WHO-5; Bech, Gudex, & Johansen, 1996; Appendix Y). Participants rated how five statements (e.g., “I have felt calm and relaxed”) applied to them over the last two weeks (0 = at no time, 5 = all of the time). Total scores (0-25) are multiplied by four to provide a percentage score (0 = worst possible quality of life, 100 = best possible quality of life; α = .91).

Procedure

Participants had to pass a Qualtrics V2 Captcha and correctly answer 8/10 English proficiency questions to enter the survey. After asking them about their social media usage, we allowed only Instagram users to enter the survey. As in Study 5a, we told participants the study was investigating media engagement, personality, and negative personal experiences. Participants filled out demographic questions and then answered questions about Instagram use and items designed to reduce suspicion about the true nature of our study: participants rated how often they usually view a list of topics on Instagram (e.g., Fashion, Food, Design, Travel etc.). Next, participants completed the mock Instagram task and related questions about sensitivity screens, followed by the Trauma History Screen (THS), PTSD symptomology (PCL), the single self-triggering question and coping questionnaires[40] in a randomized order. Participants then answered questions about depression, anxiety, and stress symptoms (DASS), wellbeing (the SGWB-14 and the WHO-5), and individual difference characteristics[41] in a randomized order. Participants were then asked to indicate if they behaved as they normally would on Instagram (“Yes”/” No”), and if “Yes” they were asked to explain how they behaved differently and why, if they left the task for any period of time (if “Yes” when and for how long), and if they had any technical issues. Participants were then fully debriefed.

[40] These data were secondary to our main research aims in this study (i.e., which focused on how “vulnerable users” interact with sensitivity screens) and are not reported here. See https://osf.io/mjrq8/

[41] See https://osf.io/mjrq8/

Results

Statistical overview

We ran analyses using Null-Hypothesis Significance Tests (α = .05) in SPSS Version 25 and JASP for MacOS version 0.13.1. Where data were missing, we used subscale level mean substitution. One person missed three items on the SGWB-14, and one person missed one item on the BACQ Avoidance subscale.

Participant characteristics

We first examined our sample for prevalence of traumatic event exposure and possible PTSD, Depression, Anxiety and Stress. Overall, 85.9% of participants reported experiencing one or more HMS events, and 65.3% of participants reported a Criterion A event. The most common events reported were the sudden death of a close family member or friend (52.7%), followed by exposure to a hurricane, flood, earthquake, tornado, or fire (44.7%), a really bad car, boat, train, or airplane accident (28.2%). Further, 21.4% of the sample met criteria for a likey PTSD diagnosis according to the conservative cut-off on the PCL-5 (> 33; Bovin et al., 2016). For Depression 54.6% of our participants were in the normal range; 29.7% mild-moderate; 18.7% severe-extremely severe; Anxiety; 59.2% normal; 19.8% mild-moderate; 21% severe-extremely severe; Stress; 61.8% normal; 26% mild-moderate; 12.3% severe-extremely severe. The majority of participants (87.8%) reported that they used social media every day in the last 7 days (followed by 5 days = 3.8%, 6 days = 3.1%, %, 2 days = 2.7%, 4 days = 1.5%, 1 day = 0.1%, and 3 days = 0.4%) for an hour or more per day (2-3 hours per day = 39.7%, 1 hour = 21.8%, more than 6 hours = 17.6%, 4-5 hours = 11.8%, less than half an hour = 9.2%). Most participants (51.1%) had used Instagram every day in the last 7 days (followed by 2 days = 11.1%, 3 days =

9.9%, 4 days = 9.5%, 5 days = 8.8%, 1 day = 6.1%, 6 days = 2.7%, and did not use in last 7 days = 0.8%).

Decision to uncover sensitive content screens and prior experience with sensitive screens on Instagram

Recall that participants viewed a sensitive content screen from Instagram and had the option to uncover and view the photo, or avoid the photo by selecting the ‘next photo’ button. Consistent with Study 5a, the majority of participants (84.7%) fell into the ‘uncover’ category. Again, as in Study 5a, we also asked our participants about encounters and interactions with sensitivity screens in real life. Over half our participants (64.5%) indicated that they had previously seen a sensitive content screen on Instagram. Participants who said they have seen the screens on Instagram reported that they almost always (*M* = 4.21, *SD* = 1.60; on a scale of 1 = *Never*, 6 = *Always*) uncovered a screened image when they came across one. Finally, 43.9% of participants said they would like to be able to turn off the sensitive screen feature (so that all photos were not screened when browsing) if they had the option to do so.

Like Study 5a, we coded participants’ text responses to the question “Why did you or did you not uncover the screened image”’ (Table 7.1) using the thematic analysis technique described by Braun and Clarke (2006). A majority of participants simply stated that they uncovered the screened image because they wanted to see the photo (72.5%), and 63.7% of these participants (around half of our total sample = 46.2%) also specifically indicated that they would uncover the image because of reasons related to curiosity or related concepts. The next most common response was to say they did not uncover because they did not want to see something negative (12.6%), or that they would uncover/keep covered based on a general tendency/personality trait to cope with or not cope with distressing content (10.7%). The type of content that may be behind the screen (e.g., nudity or gore) was mentioned by 8.4% of participants. Surprisingly, while 36.2% of participants in Study 5a mentioned contextual factors that may accompany a photo in real life (e.g., posting account, caption, comments etc.), only 1.5% of participants mentioned it in Study 5b.

Taken together, Study 5b confirms Study 5a: sensitive content screens do *not* appear to deter the majority of people from wanting to view potentially distressing images, and extends this finding from the desire to uncover to an actual behavioral task. While curiosity remains a popular reason for approaching muted content, participants in Study 5b were more likely to cite not wishing to see negative content, and personality traits, and less likely to mention the context of the image as a factor in decision making.

Is the decision to uncover a sensitive content screen associated with vulnerabilities or individual characteristics?

Unlike Study 5a, we did not find any significant associations between psychological vulnerabilities and the decision to uncover the screened image in Study 5b (Tables 7.2-7.4). One potential statistical explanation for this pattern of results is the relatively high base rate of people choosing to uncover (*n* = 220) relative to people avoiding (*n* = 40), which may have led to variance heteroskedasticity. However, Levene’s Test for Equality of Variances—comparing people who uncovered and those who did not in Study 5b—did not reveal any significant violations of homogeneity for any of our continuous dependent variables. We discuss further potential explanations below.

Discussion

Instagram claims that sensitivity screens allow “vulnerable users”—such as people with mental health concerns—to minimize unwanted negative experiences. However, in two studies we found that the majority of participants (80-85%) indicated a desire (Study 5a) or made a choice (Study 5b) to uncover a screened image. Furthermore, we found no evidence that “vulnerable users” (i.e., people with more severe psychopathological symptoms) were any more likely to use the screens to minimize exposure to sensitive content. In fact, in Study 5a we found that the desire to uncover a muted image was associated with a number of vulnerability factors, including depression, wellbeing, and PTSD symptoms. Although we did not replicate this pattern in Study 5b when we directly measured uncovering behavior, we also did not find that vulnerable people were any more likely to use the screens as a tool for avoidance. Taken together, our results show that despite the claims made by Instagram, sensitivity screens do not appear to be effective in deterring the majority of people or “vulnerable users” from consuming negative content.

Our findings fit with other recent research (Bridgland, Barnard, & Takarangi, 2021; Bruce, 2020) demonstrating that trigger warnings may not be an effective way to limit people’s exposure to negative material, and with the broader finding that people often willingly expose themselves to negative content (e.g., Oosterwijk, 2017) despite potential negative sequelae (e.g., being distressed by the content). We also found preliminary evidence that sensitive content screens—and therefore possibly trigger warnings more generally—may *enhance* curiosity about potentially negative content. This result aligns with research on the “forbidden fruit effect”: when something is forbidden or restricted, it becomes more attractive and curiosity towards it increases (e.g., Ringold, 2002). Our results also fit with the “the Pandora effect,” which shows that people are especially willing to engage with stimuli if an outcome is uncertain and negative. Skipping a covered image would have maximized certainty and emotional homeostasis; yet, when given this opportunity in Study 5b, only 15% of participants took it.

We also note that our results from Study 5b could also be explained by boredom induced novelty seeking. Boredom creates an emotional state that causes people to seek novel counter-hedonic experiences. For instance, participants given a high-boredom task (neutral photo viewing) are more likely to choose a negative than neutral set to view next (Bench & Lench, 2019). Therefore, perhaps participants in Study 5b chose to uncover the negative image because it represented a novel negative experience following five neutral and five positive photos. But because participants in Study 5b only viewed 10 photos (a task lasting from around 30 secs to 1 minute), it seems unlikely boredom would be a pertinent factor. We also note that in real world conditions, sensitivity screens are far less common than normal images, such that Instagram users may also become bored and inclined to approach screened photos. Future research should consider testing whether boredom explains participants' willingness to expose themselves to negative stimuli.

While our findings from Study 5a and 5b demonstrate a general tendency to uncover potentially negative images, we found mixed support for the ideas we posited about vulnerable groups being more susceptible to this behavior. In Study 5a, we found that certain vulnerability characteristics (e.g., poorer ratings of general wellbeing and higher ratings of depression) were associated with a greater desire to uncover screened content. These findings fit with data showing individuals with depression are more likely than those without depression to use emotion-regulation strategies to maintain or increase negative mood (Millgram, et al., 2015). Therefore, it is possible that for some people, difficulties with mental health may arise as much from one's emotion regulation *goals* as from an inability to regulate emotions (Millgram et al., 2015). If such is the case, then practices such as sensitivity screens and trigger warnings may reinforce rather than allay goal-related emotion-regulation difficulties by flagging negative content, and thereby making it easier to find. Additionally, we also found that the tendency to self-trigger was associated with the desire to uncover the screened content. Self-triggering primarily occurs in an effort to make meaning out of traumatic experiences. In this case, participants may have been motivated to uncover the image in order to ascertain meaning from doing so.

However, in Study 5b, when we asked participants to choose between uncovering a screened image or skipping the image in a behavioral task (rather than a hypothetical question as in Study 5a) we failed to replicate these associations. One possibility for this discrepancy is that vulnerability characteristics are simply not associated with the *behavioral* choice to approach or avoid muted content. However, we also did not find any evidence that the 15% of people who skipped (and therefore avoided the potentially distressing content) were people from vulnerable subpopulations. Therefore, our results demonstrate that at best, when first presented with a sensitive content screen, most vulnerable and non-vulnerable users are not deterred from approaching distressing content.

Why else may we have found differences between Study 5a and Study 5b? One possibility is that the intention to uncover a photo may be inconsistent with actually doing so (the intention-behavior gap; Sheeran & Webb, 2016). However, the fact that we found that actual frequency of uncovering behavior (84.7%) was around the same as the hypothetical desire to uncover the screened photo once we dichotomized responses (80%; χ^2 (1) = 2.01, p = .156) shows that the intention-behavior gap may not be a satisfactory answer here—unless of course, the types of participants who expressed the desire to uncover a muted photo in Study 5a were different from the types of participants who actually uncovered the photo in Study 5b. To get at this possibility, we compared participants in Study 5a and Study 5b on our key vulnerability factors of interest (i.e., variables that were significantly associated with uncovering behavior in Study 5a). We found no significant differences (*ps* = .061-.922, *ds* =0.01-0.16, ϕs = 0.02-0.07).[42] Another possibility involving individual differences is that our dichotomous variable in Study 5b was less sensitive to our vulnerability factors than our ordinal variable in Study 5a.

A second explanation lies within participants' qualitative responses about the decision to approach or avoid muted content. Specifically, participants in Study 5a seemed to place a higher importance on contextualizing the Instagram post and considering elements such as posting account, captions, comments etc. on the post as an important factor when deciding if they would uncover the photo. In Study 5a, these reasons were listed over one third of the time, whereas they were minimally mentioned by participants in Study 5b, who instead placed a high importance on feelings of curiosity. It is possible that this difference occurred because Study 5a asked a hypothetical question and therefore participants may have been more likely to contextualize the sensitivity screen in their own imagination (e.g., the type of account which may have posted it), or think about past experiences with sensitivity screens when making their decision. For instance, someone with a past trauma may have imagined what they might do if they saw a photo caption related or not related to that trauma, and selected a 3 or a 4 on the scale to indicate the fact that they may not *always* approach or may not *always* avoid content. Indeed, as stated previously, it is likely that by measuring intent in Study 5a we captured people's broader pattern of approach behavior (across different scenarios). In contrast, when we presented a sensitivity screen to participants in a mock Instagram task using an Instagram frame with a blank posting account, caption, and comments, participants may have based their decision to uncover the image

[42] We found one difference in our individual difference variables. Participants in Study 1 (*M* = 18.82, *SD* = 5.80) scored slightly higher on the deprivation sensitivity subscale of the 5 Dimensional Curiosity Scale Revised (Kashdan, Disabato, Goodman, & McKnight, 2020) than participants in Study 2 (*M* = 17.52, *SD* = 5.31; *d* = 0.23). Reported here: https://osf.io/mjrq8/

based on that image alone. That is, participants had to accept a lack of contextual information when they made their choice to uncover the photo. Future studies should manipulate contextual factors such as the posting account, captions, and comments to see if these features influence the desire to uncover the screened images. Furthermore, it may also be necessary to investigate whether alternative warning messages on the sensitivity screen would influence uncovering behavior. For instance, it is possible that curiosity and uncovering behavior would be reduced if the wording of the current warning system (i.e., “graphic and violent”) was replaced with something less extreme/sensational (e.g., “negative content”).

A third explanation is that that vulnerable populations *are* less (as found in Study 5a) or more deterred by sensitivity screens but that these effects are small. Our existing sample size (*n* = 260) is based on the finding that small correlations (*r* = .10) typically stabilize at 260 people (at 80% power; Schnbrodt & Perugini, 2013). Therefore, we believed this sample size was adequate. For 95% power for a small effect, a sample of roughly double this size (470 participants) would have been required (Schnbrodt & Perugini, 2013). However, this sample size was not feasible due to resource constraints (Lakens, 2021).

A fourth explanation involves the time we collected data. Study 5a was collected in December of 2019 prior to COVID-19 becoming a global pandemic, while Study 5b was collected in December of 2020. Given that the COVID-19 pandemic has ravaged all areas of human life, including exacerbating mental health issues (e.g., Bridgland et al., 2020), it is possible that the pandemic’s impact had some unmeasurable impact on the way our mental health variables interacted with uncovering behavior.

Our study has several limitations. First, while we found that sensitivity screens seem ineffective at deterring the majority of people from exposing themselves to potentially harmful imagery, we did not measure what happens once someone actually goes on to face the graphic content. One could argue that seeing a sensitivity screen and then viewing a graphic image may be less distressing than coming across a graphic image unaware. However, previous work on the effects of trigger warnings shows that this claim is unlikely to be true; rather, trigger warnings seem to be ineffective in alleviating emotional reactions towards negative material (Bellet, et al., 2020; Boysen et al., 2019; Bridgland et al., 2019; Sanson, et al., 2019). Moreover, because sensitivity screens seem to foster curiosity and intrigue, it is possible they also enhance other cognitive processes such as attention, encoding, and memory for negative or graphic images (versus unscreened). Regardless of the exact mechanism, an essential next step in this area of research is to assess how sensitivity screens affect, or do not affect, emotional reactions to negative images.

Second, our open text responses showed that contextual elements (such as the posting account name, captions, and comments) are likely an important factor in uncovering behavior. Since we did not include these elements, we cannot generalize our results to these contexts. However, we note that at present, there is no standardized form of captioning required for photos with a sensitive screen on Instagram. It is not uncommon for photos with sensitive screens to be posted with ambiguous or no clear captions/context.

Third, sensitivity screens—and trigger warnings—have historically been primarily intended for people with mental health vulnerabilities (e.g., PTSD, exposure to trauma, etc.). Therefore, it is possible that our results would have been different had we specifically recruited and powered our sample for specific clinical populations (e.g., people with a clinical diagnosis of PTSD). However, bearing this limitation in mind, we note that MTurk has been identified as an excellent source for studying clinical and subclinical populations (Shapiro, Chandler, & Mueller, 2013).

Fourth, while we focused primarily on trait-level effects (because of Instagram's claims about vulnerable users, as opposed to users in a vulnerable state of mind), we did not investigate state-level effects (e.g., mood and anxiety) on uncovering behaviors. It is plausible that users in different affective states may differentially chose to engage with content, or that different affective states may interact with trait-level characteristics. For instance, someone diagnosed with depression who is also in a particularly negative mood at the time that they are using a social media platform (vs. a positive mood) may be more likely to uncover and view screened content. However, prior work suggests that preference for negative content in depressed (vs. non-depressed) people persists after controlling for current emotions/mood (Milgram et al., 2015). Future research should investigate how trait- and state-level factors interact in influencing uncovering behavior.

Bearing limitations in mind, our findings from Study 5a and 5b significantly add to the field of applied clinical research on the behavioral effects of trigger warnings. This research is in its infancy and there are currently only two published papers that have examined the effect of trigger warnings on avoidance behaviors. However, neither of these papers focused on approach or avoidance behaviors as a main aim. Additionally, no research has examined the rates of approach versus avoidance behaviors for visual content censoring systems. To date, the effectiveness of content censoring systems remains untested, even though they are widely employed across the internet—including on Instagram, Twitter, Reddit, Buzzfeed. Furthermore, no research has examined if vulnerability factors (e.g., wellbeing, depressive symptoms etc.) relate to the rates of approaching or avoiding content accompanied by trigger warning messages.

Therefore, we believe our key finding that sensitivity screens do not deter vulnerable people from viewing negative content offers a valuable contribution to the field of clinical science. Overall, our results demonstrate that sensitive content screens may be ineffective at deterring vulnerable and non-vulnerable users from approaching potentially graphic content. Our data suggest that alternative, empirically grounded methods for flagging potentially negative content on social media may be necessary.

8 General Discussion

My thesis aimed to bridge gaps left by the first wave of trigger warning research to help determine *how* and *why* these warnings may or may not change emotion and behavior. My final chapter serves to draw together the findings from my five thesis chapters in the context of previous research, theory, and claims (i.e., trigger warning advocates vs. critics). Finally, I will also discuss the real-world implications of my findings and the key methodological limitations of my research.

8.1 Trigger warnings and emotional reactions: Summary of findings and theoretical implications

The first main aim of my thesis was to bridge critical gaps in knowledge about the emotional effects of trigger warnings. That is, what happens when someone sees a trigger warning and how do warnings change emotional reactions towards content?

Antecedent focused trigger warning effects: The emotional and cognitive effects of seeing a trigger warning

My research is the first to examine what happen to someone's emotions and cognitions when they simply see a trigger warning message. That is, what are the emotional and cognitive effects of trigger warnings *prior* to viewing material? In Study 1 (Chapter 3) and Study 2 (Chapter 4) I found that state anxiety and negative affect increased from before to after viewing a trigger warning message, a pattern that indicates trigger warnings cause noxious anticipatory periods prior to viewing material. Additionally, in Studies 1b and 1c (Chapter 3) this effect remained when I removed the more extreme elements of the trigger warning message (i.e., mentions of torture and maltreatment) participants saw—attesting to the robustness of this effect. In Study 3 (Chapter 5), I found that imagining encountering a warning (e.g., on television) was just as anxiety-provoking as imagining encountering trauma-related content itself (e.g., a television show depicting a traumatic event). Additionally, aside from direct negative emotional reactions, in Study 1 (Chapter 3) I also confirmed that trigger warnings cause people to develop negative expectations about the contents of upcoming material. Therefore, overall, I found that viewing a trigger warning message can create anticipatory anxiety and foster negative expectations about upcoming material.

My findings therefore support the idea that trigger warnings lead people to develop negative expectancies about content. This process is the first step in *response expectancy theory* (Kirsch, 1985), where people anticipate responses to environmental cues (e.g., distress about potentially negative content), leading them to internally generate those anticipated responses. Negative expectancies are also the first step in the development of *nocebo effects*—when expecting something negative exacerbates negative reactions (Benedetti et al., 2007). These results also offer evidence that trigger warnings might act as *an emotional prime* for upcoming content, because they activate a negative mood and mindset (Collins & Loftus, 1975; Lee, Oyserman, & Bond, 2010).

However, these results could also offer support for the idea that that trigger warnings lead people to *brace for the worst*. Indeed, when people brace for the worst—i.e., expect something negative to happen to them in the near future—they can experience a marked decline in optimism in favor of pessimism (Shepperd, et al., 1996), prolonged negative affect (Neubauer et al., 2018), and physiological effects such as increased blood pressure (Spacapan & Cohen, 1983). While the act of bracing leads to these immediate negative effects, its ultimate aim is to help people take control of their emotional state and *prepare* for the worst when anticipating how they may react to upcoming content. Indeed, some previous studies show support for the idea that bracing can lead to short term emotional benefits in the immediate aftermath of an expected (vs. unexpected) negative outcome (e.g., Shepperd & McNulty, 2002; Sweeny & Dillard, 2013). One could argue then that the noxious anticipatory period people experience when seeing a trigger warning reflects emotional resources being devoted to some kind of antecedent-focused emotion regulation strategy or mental preparation, such as changing their interpretation of the upcoming content (Gross, 2002). For instance, you might feel anxious because you have been told something distressing is upcoming, but you might also focus on bringing strategies to mind to help mitigate distress.

Unfortunately, however, findings from Studies 3 and 4 (Chapters 5 and 6) suggest that the negative anticipatory period experienced when someone sees a trigger warning is very unlikely to reflect any form of helpful "emotional preparation." In Study 3 (Chapter 5), when I asked participants to tell me about what they would do when they came across a trigger warning that mentioned the topic of their most stressful/traumatic event, only a minority of participants (25.4%) mentioned some form of approach coping strategy (e.g., re-appraisal such as reminding themselves to focus on non-emotional aspects of the situation like factual information; Shiota & Levenson, 2009). This result suggests that most people *do not* bring mental preparation strategies to mind when they first see a trigger warning, prior to viewing material. Furthermore, when I compared these participants to those who thought about coming across content related to their most stressful/traumatic event, the number of participants who generated approach strategies was similar, suggesting that specifically thinking of a trigger warning does not act as a magic tool that reminds people to draw on a set of coping skills. These conclusions were confirmed in Study 4 (Chapter 6) when I used how long participants spent waiting on warning versus neutral instruction screens—prior to viewing distressing images—as a proxy for preparation opportunity. Participants should have spent longer on the trigger warning message screens (vs. control screens) if warnings cause people to *pause and prepare* to cope with upcoming distressing material. That is, someone thinking of a way to emotionally reappraise upcoming material should spend a longer time lingering on the screen compared to someone just reading an instruction to press the "next image" button. Yet, across the whole image viewing task, participants spent a similar amount of time waiting on warning versus control screens. Furthermore, when examining the first two trials of the image task—which were always a warning and control screen (randomized)—participants spent *less* time on the trigger warning screen.

As I argued in Study 3 (Chapter 5), these findings likely reflect the fact that most people may not even know what coping strategies are, or how they could use them (e.g., how to reappraise emotional content), a conclusion supported by participants' qualitative responses (e.g., *"I don't have a lot of coping techniques. I never was able to afford to see a therapist..."*). Alternatively, perhaps people do know about helpful strategies, but trigger warnings do not remind people that they are encountering a situation where coping strategies should be used. Whatever the scenario, trigger warnings seem to be ineffective. Interestingly, in Study 3 (Chapter 5), when I prompted participants to specifically think about the *coping strategies* they would use if they came across a trigger warning, rather than just asking what *they would do* when they saw a trigger warning, the number of reported approach strategies participants reported increased significantly. Therefore, it is possible trigger warnings could be more successful in achieving their proposed aims if they specifically and directly instructed people to bring coping strategies to mind.

One potential criticism of this interpretation might be that not everyone feels the need to bring coping strategies to mind when they see a trigger warning. Put differently, trigger warnings were originally used to warn people about content related to traumatic experiences; the term "trigger" originates from research showing that survivors of trauma can re-experience strong emotional responses to trauma if they are *triggered* by stimuli similar to the event. Therefore, perhaps people who have experienced traumatic events, or have personal experience with the topic of the trigger warning, may use them differently when they first see them, versus people who have not been exposed to trauma matching the warned content.

In Study 3 (Chapter 5) I explored this possibility by comparing participants who were likely PTSD-positive versus those likely negative according to the conservative PCL-5 cut-off (> 33; Bovin et al., 2016). I found no group differences in the types of coping strategies participants brought to mind. Furthermore, recall that in Study 4 (Chapter 6) I did not find that participants spent more time pausing on the warning versus control screens. Critically, these findings run against the argument that people with prior experience with a negative event will use trigger warnings to "pause and prepare," since all participants had some experience with the potentially distressing stimuli because they had all seen the trauma film. Additionally, I also found no difference in the "preparation" time spent waiting on the warning screens for participants who had experienced a Criterion A event (i.e., actual or threatened death or serious injury) or events specific to the topic of the film compared to those who had not had these experiences. Therefore, it seems unlikely that participants who are likely PTSD positive (vs. negative), have experienced a Criterion A trauma, or have personal experience with the topic of the trigger warning (in real life or within a trauma analogue paradigm), use warnings to pause and prepare to cope with content.

Taken together, seeing a trigger warning causes a noxious anticipatory period characterized by anxiety and negative affect. This aversive waiting period does not appear to reflect a conscious (e.g., when people are asked directly) or unconscious (e.g., based on how long someone spends preparing for negative content) effort to bring strategies to mind to emotionally prepare oneself to cope with content.

Response focused trigger warning effects: The emotional consequences of viewing a trigger warning on reactions to stimuli.

Thus far, I have talked at length about what happens when someone merely sees a trigger warning. However, what happens after someone sees a warning and subsequently views material? Overall, previous research has found that trigger warnings seem to have trivial effects on people's immediate emotional reactions to distressing novel stimuli including text passages (Bellet et al. 2018; Bellet et al. 2019; Jones et al., 2020; Gainsburg & Earl, 2018) lecture material (Boysen et al., 2021), and trauma films (Sanson et al., 2019). Importantly, my thesis helps to answer questions left by this first wave of research.

Do trigger warnings skew perceptions about neutral or ambiguous material? First, in Study 1 (Chapter 3), I investigated the effects of trigger warnings on emotional reactions to neutral and ambiguous material (i.e., images that could be interpreted as positive or negative). This investigation was important because trigger warnings are provided for a wide range of topics—from relatively neutral subjects such as pregnancy, to overtly distressing subjects such as suicide—and research on *priming effects* shows that priming a concept (e.g., negative adjectives such as "mean") can cause people to interpret ambiguous information in line with the primed expectations (e.g., rating an unknown person in a photograph higher on a number of negative traits; Ferguson, Bargh, & Nayak, 2005). By warning or not warning participants that a series of neutral (accompanied by a neutral news headline), negative (accompanied by a negative news headline) and completely ambiguous photos (no headline; could be interpreted as positive or negative), would be distressing, I was able to examine if trigger warnings encourage interpretations (Waldman, 2016) or skew perceptions (Filipovic, 2014) about content. In line with prior research on overtly distressing stimuli, I found that despite the fact that warnings seem to *prime* negative expectations (discussed above), they had trivial effects on emotional reactions to neutral and ambiguous content.

Why don't trigger warnings alleviate distress? Second, as discussed in detail earlier, my research is the first to confirm that trigger warnings do not seem to alleviate negative emotional reactions, perhaps because they are ineffective in helping people to pause and prepare or bring coping strategies to mind. Indeed, in Study 4 (Chapter 6), I found no association between the time participants spent waiting on the trigger warning screens (where they could have been emotionally preparing to view the next distressing image) and their ratings of distress throughout the study. In Study 2 (Chapter 4), I found no evidence that seeing a trigger warning led participants to use more strategies to cope with the distress associated with recalling a negative event over a two-week period, versus participants who saw no warning. Furthermore, in Study 3 (Chapter 5) I found that participants used fewer positive emotion words (e.g., love, sweet, nice) when describing how they would cope when thinking about encountering a trigger warning versus content related to their most stressful/traumatic event. Therefore, trigger warnings lead to negative anticipatory periods with no observable emotional pay-offs.

Why don't trigger warnings exacerbate distress? Third, my research helps to answer another remaining trigger warning conundrum: it seems clear that trigger warnings lead to the development of a noxious anticipatory period and negative expectancies, but why don't trigger warnings consistently lead to exacerbating negative responses across studies—as predicted by response expectancy or nocebo effects? There are several possibilities.

Expectations might not always match with the actual emotional experience of viewing the content: "Nothing is so frightening as what's behind the closed door." First, expectancy contrast effects offer one explanation for why trigger warnings do not lead to consistent increases in immediate distress reactions when people view warned of material. Curiously, in Study 1 (Chapter 3) when I investigated the effects of trigger warnings on emotional reactions to neutral and ambiguous material, I found that participants reported feeling *slightly less* arousal towards images when they were warned versus unwarned—although this effect was small. This finding likely reflects the fact that people's expectations about the distressing nature of the images did not match the nature of those images, potentially resulting in mild feelings of relief. In fact, in Study 1 (Chapter 3), I asked participants to rate the experience of the photo stimuli in light of what they expected at the beginning of the study, which confirmed that warned (vs. unwarned) participants found the photos "somewhat more positive than expected." Furthermore, in Study 1e (Chapter 3), I found that directly before viewing the photos, warned participants expected the photos to be more negative than unwarned participants did. After photo viewing, warned participants rated the photos as significantly less negative than their initial rating, while unwarned participants actually rated the photos as significantly more negative than their initial rating.

These findings can be explained by priming: new information is only likely to be assimilated into a primed category if the category is considered moderately extreme (e.g., a moderately hostile person such as "a boxer"; Herr, 1986). However, if the prime is an exemplar of an *extreme* category (e.g., Hitler, Stalin, etc.) participants contrast ambiguous stimuli to this category (e.g., rate ambiguous behavior as *less* hostile). The Affective Expectation Model (Wilson et al., 1989) provides a similar explanation. According to this model, emotional experiences may be assimilated or contrasted with prior expectations based on people's recognition of an expectancy-experience discrepancy. If people do not detect the discrepancy (i.e., that the expectations are at odds with the experience) they will likely assimilate the stimulus with the prior expectation. If the discrepancy is detected or is made obvious, people will contrast their reactions in a direction away from the prior expectation. Thus, because the warning was emotionally distressing, the ambiguous photos may have been contrasted with—rather than assimilated into—a negative category, resulting in the photos feeling more positive than expected.

But, how does a priming contrast or expectancy violation effect help to explain why previous research has *not* found that trigger warnings exacerbate reactions to more explicitly distressing stimuli. Or, put differently, why do the negative expectancies primed by warnings not result in assimilation of negative material every time? Here, I refer to Stephen King (1981), describing why horror writers should avoid showing the monster in order to create suspense and dread:

> You approach the door in the old, deserted house, and you hear something scratching at it. *The audience holds its breath along with the protagonist as she/he (more often she) approaches that door. The protagonist throws it open and there is a ten-foot-tall bug. The audience screams, but this particular scream has an oddly relieved sound to it. "A bug ten feet tall is pretty horrible," the audience thinks, "but I can deal with a ten-foot-tall bug. I was afraid it might be a hundred feet tall."* (pp. 116-117).

It is therefore possible that some people are relieved when they actually come to view the material the trigger warning was alerting them about—despite the negative nature of the material. That is, the material in the aforementioned trigger warning studies (i.e., text passages, films, lecture material etc.) is undoubtably negative and makes people feel negative, but perhaps *not as* negative as some people expected when seeing the trigger warning. Of course, it is likely that any contrast effects that occur when someone views negative material will be smaller than those I observed in Study 1 (Chapter 3) when using material that highly contrasted (i.e., neutral and ambiguous material) with what the warning suggested. However, thus far, my research in Study 1 (Chapter 3) is the *only* research to employ a pre and post expectancy measure, so there is no way of knowing if the distressing stimuli used in previous studies matched or exceeded the negative expectancy created by the trigger warning messages.

However, why don't we observe a consistent relief reaction when someone is warned versus when they are not warned? Put differently, why don't people experience a consistent contrast effect and find negative study stimuli more positive? One explanation is that expectancy violation does not occur for everyone. For some people, their expectancy may match their emotional experience with the material (resulting in assimilation), but for others it may not (resulting in a contrast effect). This possibility may explain the results of several previous studies. Recall that Bellet et al. (2018) found people who believed that words could cause harm experienced more anxiety when reading distressing text passages when they were warned versus unwarned. Similarly, Gainsburg and Earl (2018) found that participants who believed that warnings were protective (vs. coddling) experienced more distress when viewing content marked with a warning—while there was no difference between these two groups of participants when distressing material was preceded by a control message. Finally, Jones et al. (2020) found that participants with *higher PTSD symptoms* had increased anxiety when viewing content accompanied by a trigger warning (vs. no warning). Participants who held the belief that words can cause harm, that warnings protect people from a credible harm, or who had higher PTSD symptoms may have experienced levels of distress when viewing the material that *matched or exceeded* their expectations about the task when they read the trigger warning message. However, consistent with other research, none of these researchers found that trigger warnings led to increases in anxiety responses towards the content when they examined the sample as a whole. This points to the possibility that, based on individual characteristics, some participants may find that the levels of expected distress prompted by the trigger warning match their experience with the stimuli, and experience a nocebo/assimilation effect. Other participants may experience a relief/contrast reaction because they were expecting to see something worse, based on the trigger warning, but actually find the content bearable. Still other participants may experience *no* nocebo or relief effects—perhaps they did not expect the content to be negative and also were not bothered by the content. This expectancy washout effect helps to explain why no trigger warning research has observed clear nocebo effects, or relief effects, for warned versus unwarned participants.

Does this idea mean that warnings might work for some people and provide positive emotional relief in the face of distressing content? This seems unlikely. It seems more likely that relief reactions, if they do happen to people in the first place, are not a very effective way to alleviate negative affect. Indeed, in Study 1 (Chapter 3), even when I used highly contrasting stimuli, the relief effect caused by the trigger warning did not extend to the way that participants rated how emotionally distressing the images were or how costly or beneficial they perceived the entire study experience. Moreover, people who do not experience relief reactions (i.e., people who have their expectancies confirmed) may be experiencing nocebo effects—leading to obvious emotional harm. Future research should include measures of participants' expectations and examine if certain types of people may be more or less susceptive to expectancy violation versus confirmation effects.

Negative effects may only emerge over time. A second explanation for why trigger warnings do not lead to immediate increases in distressing reactions is that warnings may not always have immediately observable effects, but instead change emotional responses over time. Thus far, my research in Study 2 (Chapter 4) is the only investigation of the effects of trigger warning messages over a delay. Participants were warned or not warned about the distressing nature of a memory recall task that required them to report a negative event from the past two weeks. Participants returned after two weeks to recall the same memory again and those that had been warned in the first session (vs. unwarned) reported a smaller reduction in PTSD-like symptoms (e.g., "I had trouble staying asleep") over the delay period. This pattern may reflect a memory amplification effect (Southwick, Morgan, Nicolaou, & Charney, 1997). That is, it is possible that warned participants thought about or were reminded more about their negative memory over the two-week delay, resulting in a smaller decrease in PTSD-like symptoms.

These findings are important because previous work on trigger warnings has only used single measurement designs and focused on the short-term reactions immediately following stimuli. It is possible that over time, small negative effects caused by warning messages, such as anticipatory anxiety (Study 1 and 2; Chapter 3 and 4), accumulate and have more potent emotional consequences (Funder & Ozer, 2019). This consideration may be especially important for warnings that are becoming increasingly prevalent in everyday life across both formal (e.g., educational) and casual (e.g., entertainment) contexts. This possibility thus presents an important avenue for future research efforts. First, it would be useful to conduct a field or diary study to track the frequency of encountering trigger warnings in daily life. Second, more trigger warning research should measure participants' emotional responses to warned material over more than one experimental session.

Negative effects may occur for different types of appraisals rather than immediate emotional reactions. A third possibility is that trigger warnings do little to change immediate emotional reactions, but do change other types of reactions and appraisals. For instance, in Study 2 (Chapter 4), I found that PTSD-like symptoms about a negative event—such as event impact characteristics like feeling jumpy and easily startled—subsided less over a two-week delay for participants who were warned in the first session (vs. unwarned participants). This investigation is important because it is the only of its kind to investigate how warnings might change how people appraise the PTSD-like symptoms associated with a negative event over time. Critically, these findings also help us to understand the efficacy of trigger warnings as they were originally defined—to mitigate the "triggering" process by alerting viewers that upcoming content may spark the recall of traumatic memories, specifically, not just that provocative or sensitive material may be encountered (Haslam, 2017). Indeed, many websites promoting the use of trigger warnings claim that "triggers are more distressing if they come as a surprise" (Cuncic, 2020; Sullivan, 2019; Good Therapy, 2018), or similarly, that "vivid memories of trauma are more distressing if they happen without any warning" (The Innocent Lives Foundation, 2020). However, I found no evidence for these claims.

The idea that trigger warnings might change other appraisals aside from immediate emotional reactions also fits with Jones et al. (2020), who found that participants with a history of trauma reported that their traumatic event was more central to their identity when they were exposed to trigger warnings (vs. no warnings). Event centrality—the belief that a traumatic event marks a turning point in one's life story—is associated with PTSD symptoms (Berntsen, & Rubin, 2006), and prospectively predicts more severe PTSD (Boals & Ruggero, 2016). Therefore, it seems possible that the negative expectancies caused by trigger warning messages do lead to nocebo effects, but that these effects relate more closely to the way that someone appraises a negative life event, rather than immediate distress reactions.

Trigger warnings and emotional reactions: Conclusion

Taken together, upon viewing a trigger warning people experience an anxious anticipatory period that does not seem to reflect mental preparation to cope with negative content. Indeed, overall, my thesis and the work of others unanimously suggests that trigger warnings *do not* mitigate distressing reactions. Rather, it is more likely that trigger warnings lead to harm and my thesis suggests three possibilities for when and how this harm is likely to occur. First, it is possible that trigger warnings have the potential to exacerbate distressing reactions to experiences when expectations match with those experiences. Second, the negative effects of trigger warnings may only emerge over time. Third and finally, the negative effects of trigger warnings may not occur for immediate emotional reactions, but rather for other kinds of appraisals more closely linked with negative memories, such as PTSD symptoms.

Trigger warnings and avoidance: Summary of findings and theoretical implications

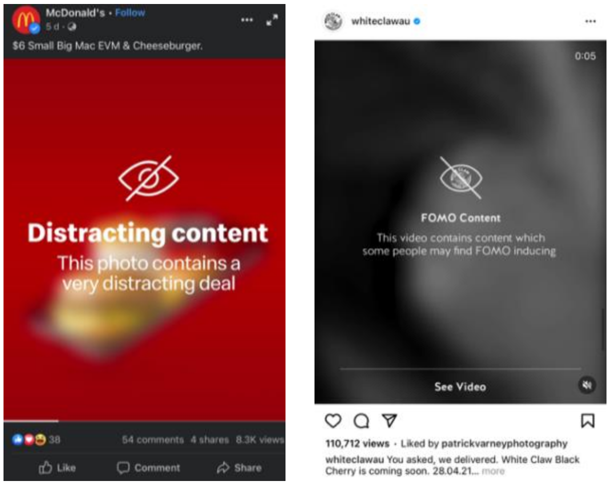
Aside from emotional reactions, the second aim of my thesis was to investigate how warnings may or may not change avoidance behaviors. That is, do people use trigger warnings as a signal to avoid potentially distressing content? Previous research on trigger warnings and avoidance was minimal. My thesis has helped to expand our understanding of trigger warnings and avoidance in three key ways.

A narrow definition of avoidance coping. First, the three previous studies on trigger warnings and avoidance focused on a very narrow definition of avoidance coping—the complete behavioral avoidance of stimuli. Participants were given the choice to pick news headlines (Bruce & Roberts, 2020), film titles (Gainsburg & Earl, 2018), or essay readings (Kimble et al., 2021) with or without trigger warnings. This type of strategy is known more specifically as *situation selection* and occurs when someone is given the opportunity to approach or avoid a specific situation (Gross, 2002). For instance, Situation 1 might be to approach a film title accompanied by a trigger warning while Situation 2 may be to avoid that title by choosing a title unaccompanied by a warning. However, approach coping also comprises a wider range of potential behavioral, emotional, and cognitive responses (Littleton et al., 2007). Indeed, someone could initially select a situation (i.e., choose to approach something accompanied by a warning) but then decide to avert their gaze, skim their eyes over particularly distressing parts or completely quit viewing the material after they realize the true nature of the stimuli. Emotional and cognitive avoidance might occur if someone tries to suppress their thoughts and feelings about the material during and after exposure. Therefore, in Study 3 (Chapters 5) and 4 (Chapter 6) I investigated a more complete picture of the types of avoidance coping strategies that someone might use when faced with a trigger warning which was missing in prior investigations.

In Study 3 (Chapter 5) I asked one group of participants to tell me what they would do if they came across a trigger warning, and another group what they would do if they came across direct content, related to their most stressful/traumatic event. This method enabled me to examine a more nuanced picture of avoidance coping because participants could report a range of potential behaviors. For instance, someone could completely avoid warned content (situation selection), while someone else might view content marked by a trigger warning but try and stop their thoughts and emotions (emotional/cognitive avoidance). I found that participants were just as likely to mention they would use an avoidance-based coping strategy whether they were thinking of a seeing a trigger warning or of content directly related to their trauma. These findings provide a more complete picture of the wide range of coping strategies that people may draw upon when encountering distressing stimuli, and suggest that people would be just as likely to summon an avoidance-based strategy—such as turning off the television or trying to stop their thoughts about the topic—when they see something related to their traumatic event (e.g., a show depicting a car crash) as they would when they see a trigger warning.

However, although measuring actual behavior was not my aim in Study 3 (Chapter 5), hypothetically simulating the future may not capture what actual future behavior would look like. Therefore, in Study 4 (Chapter 6) I conducted a conceptual replication of Study 3 (Chapter 5). In Study 4 (Chapter 6), rather than asking participants to recall their most traumatic/stressful event, participants watched a trauma film. Additionally, rather than asking participants to imagine seeing a trigger warning and reporting what they would do, I measured participants' actual behavioral reactions towards distressing content that was preceded by warning or control messages. Participants completed a photo viewing task where they saw image stills from the film for 5(s) each but were told they could avoid viewing the photos by pressing the space bar on the keyboard that would bring up a black screen for the viewing duration. Aside from complete avoidance—i.e., if the participant pressed the space bar immediately after the message screen—this design also enabled me to examine if warnings change how long people spend viewing stimuli after they are exposed. Replicating the findings of Study 3 (Chapter 5), I found that participants did not avoid more distressing stimuli related to the film if it was preceded by a trigger warning versus when directly encountering the stimuli without a warning (i.e., with a control screen). In sum, trigger warnings do not seem to change or promote a wide range of avoidance coping strategies.

A narrow use of experimental stimuli. Second, both Gainsburg and Earl (2018) and Bruce and Roberts (2020) examined avoidance behaviors using a specific paradigm—how often participants selected *titles* (film and news articles) accompanied by trigger warnings. They found no difference between the selection of titles accompanied versus unaccompanied by warnings. However, this design meant that participants were essentially *warned* about the contents of the article via the information conveyed in the title itself, which may have meant that the trigger warning had little additional effect (e.g., a title from Gainsburg & Earl, 2018: "*When racial profiling leads to policy brutality: An investigation*"). Trigger warnings employed in real life settings often contain only vague information about the content. In Study 5 (Chapter 7), I explored one applied context where vague and nondescriptive trigger warnings are used. Sensitive content screens on Instagram warn users of negative content but do not provide any information about that content (i.e., "*Sensitive content: This photo/video may contain violent or graphic content*")—although I acknowledge it is possible that users might get hints about the content of the images from user generated captions or comments. Rather than completely removing all graphic and negative content from the app, Instagram censors potentially distressing visual material via a Gaussian Blur, which reduces image noise and detail, and by adding a warning message. The purpose of these screens is to help reduce "surprising or unwanted experiences" and help people to *avoid* potentially distressing content. However, prior to my investigation no published research had examined how likely it would be for someone to come across one of these screens and choose to avoid, rather than approach, the content underneath. In two experiments, I found that 80-85% of participants indicated an intention, or actually made a choice, to uncover and view an image covered by a sensitive content screen in the absence of any other information about the image. These low rates of avoidance fit with emerging research (Kimble et al., 2021) showing that participants seem to be extraordinarily reluctant to avoid distressing study stimuli. For instance, in Kimble et al. (2021) when given the option to avoid reading “triggering” text, less than 6% of participants took the option. Indeed, in Study 3 (Chapter 5), overall rates of avoidance throughout the study were incredibly low—only 12.56% of participants used the cover feature at all. These findings likely reflect the “Pandora effect,” which suggests that people have a general tendency to approach rather than avoid stimuli that has been marked as aversive and uncertain (Hsee & Ruan, 2016; Oosterwijk, 2017). Furthermore, these results also raise the possibility that trigger warnings foster a “forbidden fruit effect”—where warnings actually increase rather than decrease attraction to potentially negative material. In fact, the possibility that trigger warnings might increase rather than decrease attraction has already been used by advertisers to draw attention towards unhealthy food products such as fast food and alcohol (Figure 8.1). Taken together, trigger warnings in their current form do not appear to be effective in promoting avoidance behaviors in the majority of people.



McDonald's · Follow
5 d ·
$6 Small Big Mac EVM & Cheeseburger.

Distracting content
This photo contains a very distracting deal

whiteclawau

FOMO Content
This video contains content which some people may find FOMO inducing

110,712 views · Liked by patrickvarneyphotography
whiteclawau You asked, we delivered. White Claw Black Cherry is coming soon. 28.04.21... more

Figure 8.1 Examples of trigger warning "sensitive content screens" used by advertisers.

A narrow exploration of vulnerable populations. Third, previous research (Bruce & Roberts, 2020; Kimble et al., 2021) has exclusively examined how trauma survivors approach or avoid content marked with trigger warnings. However, other research (e.g., Bellet et al., 2020; Redmond et al., 2019) suggests that other "vulnerable populations" (e.g., people with depression; Milgram et al., 2015, or who self-trigger; Bellet et al., 2020) may be *attracted* to negative material. Therefore, trigger warnings may be especially ineffective in deterring these populations from consuming negative content. No previous research had specifically examined how people who have lowered mental wellbeing or who identify as engaging in self-triggering behaviors approach or avoid content marked with warnings. I therefore explored these possibilities in Study 5 (Chapter 7). In Study 5a I found that amongst other characteristics, poorer ratings of general wellbeing and higher ratings of depression) were associated with a greater desire to uncover screened content. Moreover, the tendency to self-trigger was associated with the desire to uncover the screened content. In Study 5b, when we asked participants to choose between uncovering a screened image or skipping the image in a behavioral task (rather than a hypothetical question as in Study 5a) I failed to replicate these associations. However, I also did not find any evidence that the 15.3% of people who skipped (and therefore avoided the potentially distressing content) were people from vulnerable subpopulations. Therefore, Instagram's claims that sensitive screens protect (via increased avoidance of negative material) the most vulnerable users of the platform appear unfounded. Nevertheless, Study 5a in Chapter 7 provides preliminary evidence that vulnerable populations might be more attracted to negative content marked with warning messages under some circumstances. Future research should endeavor to expand this exploration and unearth the factors that might enhance people's attraction to negative material.

Trigger warnings and avoidance: Conclusion

Taken together, trigger warnings do not appear to enhance a range of avoidance coping strategies versus when approaching distressing content unwarned. Furthermore, trigger warnings do not seem to be an effective method of deterring the majority of people from engaging with distressing stimuli.

8.2 Theoretical implications

Emotion regulation is unlikely to be a spontaneous process

Approach coping strategies such as emotion regulation are often nonconscious and highly context sensitive processes (Gross, 1999). Previous studies have shown that explicitly guiding people to use emotion regulation techniques such as emotional reappraisal (i.e., changing how you evaluate a situation) can successfully help people to feel less distressed by negative stimuli—similar to stimuli that typically follow trigger warnings in real life such as graphic films (e.g., Shiota & Levenson, 2012; Troy, Shallcross, Brunner, Friedman, & Jones, 2018; Wolgast, Lundh, & Viborg, 2011). However, my research suggests that it is unlikely that people spontaneously draw upon reappraisal strategies consciously (i.e., when I asked people to think about what they would do in Study 3; Chapter 5) *or* non-consciously (i.e., when measuring how long people spent waiting on warning screens in Study 4; Chapter 6) when they come across a trigger warning. Therefore, my research adds to the emotion regulation literature by showing that it is likely that people need explicit instructions about how to use emotional reappraisal when encountering a potentially negative situation. That is, simply drawing someone's attention to an impending undesirable mood state that might be caused by viewing negative material (via a warning) is *not* sufficient to trigger helpful emotion regulation processes. These findings may extend to other emotional situations where reappraisal may be useful—for instance, when receiving medical test results. You may have been *warned* by your doctor that the results might be bad news, and therefore emotional reappraisal could help you to reduce the emotional impact. However, without explicit directions about *how* to emotionally reappraise the situation, it seems unlikely that you will spontaneously draw upon reappraisal.

Bracing for the worst does not seem to be effective for situations involving response expectancies

Previous studies on bracing for the worst have focused almost exclusively on outcome expectancies—that is expectancies about outcomes related to external stimuli or events (Kirsch, 1895). These outcomes include test performance scores (Gilovich, Kerr, & Medvec, 1993; McKenna & Myers, 1997; Savitsky, Medvec, Charlton, & Gilovich, 1998; Sweeny, Shepperd, & Carroll, 2009), medical test results (Taylor & Shepperd, 1998; Shepperd & McNulty, 2002; Sweeny & Dillard, 2013), financial outcomes (Shepperd, Findley-Klein, Kwavnick, Walker, & Perez, 2000; Mellers, Schwartz, Ho, Katty, & Ritov, 1997; van Dijk & van der Pligt, 1997), and sports games (McGraw, Mellers, & Ritov, 2004). Typically, these studies find that negative outcomes are generally less aversive when they are expected versus come as a surprise. My research expands our knowledge of bracing for the worst by exploring response expectancies—expectancies about internal non-volitional experiences (Kirsch, 1895). That is, rather than investigating how people who expect a positive or negative outcome about an external event (e.g., getting a good versus bad grade on a test) react when they receive a factual outcome (e.g., actually receive a bad grade), I examined how expecting to have a bad emotional response (i.e., towards negative material) changes the way that someone feels when they encounter that material (Studies 1 and 2; Chapters 3 and 4). My findings suggest that bracing for the worst is ineffective when it comes to response expectancies. That is, when the outcome is a negative emotional one, rather than a negative factual outcome about the state of the world, bracing does not seem to lessen the blow. Therefore, when considering if bracing for the worst will be an effective strategy to use when encountering a potentially negative situation, a person should consider whether the situation involves an outcome-based expectancy or a response expectancy.

Extreme negative expectancies might flip the nocebo effect

My research expands our knowledge about expectancy effects and more specifically about the Affective Expectation Model (Wilson et al., 1989) by demonstrating that assimilation and contrast effects are also possible when people have negative expectations about material. Previous research in this area has focused primarily on positive expectations about material—specifically, on expectations about how *funny* cartoons and film clips would be (Geers & Lassiter, 1999; Geers & Lassiter, 2005; Wilson et al., 1989). However, my research in Study 1 (Chapter 3) suggests that contrast and assimilation effects are also possible when participants *negative* expectations about material do not match to what they actually experience. These findings have important implications for response expectancy and nocebo theories, because they suggest that if negative expectations (e.g., about pain) are *much* worse and do not match with what someone actually experiences, there is a chance that negative outcomes are *reduced* rather than *exacerbated*. Future research should explore this possibility in related areas such as nocebo pain responses.

The forbidden fruit and the Pandora Effect might depend on individual difference factors

Thus far, research on the forbidden fruit effect and Pandora Effect has focused on how warnings enhance the desire to consume negative material (e.g., Bijvank, Konijn, Bushman, & Roelofsma, 2009; Bushman & Stack, 1996; Hsee & Ruan, 2016; Oosterwijk, 2017). Along these lines, other research has found that certain vulnerable populations—such as people with depression (e.g., Millgram et al., 2015) or who self-trigger (Bellet et al., 2020)—may also be attracted towards negative material. However, my research is the first to draw these two areas of research together to show how they might potentially interact (Study 4 and 5; Chapter 6 and 7). Specifically, my research contributes to theory by demonstrating that the Forbidden fruit and Pandora effect might be enhanced in certain vulnerable populations (e.g., people with lowered wellbeing). Future research should continue to explore how individual difference factors may play a role in the forbidden fruit and Pandora Effects.

8.3 Applied implications

Claims of advocates and critics.

Emotional effects of trigger warnings. When people discuss why trigger warnings are helpful, there is a chronic, widespread, and culturally ingrained, notion that they help people "*prepare*"—noted in survey data (Bentley, 2017; Cares et al. 2017; DeBonis, 2019; George & Hovey, 2019), media articles (Cripps, 2020; Gust, 2016; Manne, 2015; McNeil, 2015) and mental health resources (Cunic, 2020; Good Therapy, 2018; Innocent Lives Foundation, 2020;

Sullivan, 2019). These claims likely originate from the commonly held assertion that negative outcomes are worse if they are unexpected than expected. However, in my thesis research I found no evidence to support this claim. In Study 3 (Chapter 5) when I asked people what they would actually do when they came across a trigger warning, shifting the question from "*what* do trigger warnings do?" to "*how* do they do this?" most people were not able to provide an answer. In Study 4 (Chapter 6), I also found no evidence to suggest that trigger warnings remind people to *pause and prepare* (Strothman, 2021) to cope with distressing content. Furthermore, my research converges with other evidence showing trigger warnings do not mitigate emotional distress when people are facing potentially distressing material.

Alternatively, critics claim that trigger warnings might *exacerbate* negative emotional reactions, by instilling fears and apprehension about upcoming content that would not have existed in the absence of the warning (e.g., Lesh, 2016; Lukainoff & Haidt, 2015). My thesis offers partial support for these claims. Trigger warnings do lead to negative expectancies and make people feel anxious in the lead up to consuming material, but the negative effects of warnings on reactions to the warned of material appear to occur under specific circumstances. Specifically, negative effects may only occur when expectations match with experiences, may only emerge over time, and seem unlikely to occur for immediate emotional reactions, but rather for other kinds of emotional appraisals such as PTSD symptomology.

In sum, at best, trigger warnings in their current popular form seem to do little to alleviate distress and at worst have harmful emotional effects.

Trigger warnings and avoidance. Both advocates (e.g., Medhora, 2021) and critics (e.g., Lukianoff & Haidt, 2015) of trigger warnings claim that warnings enhance the avoidance of potentially distressing material. However, I found no evidence for this claim (Studies 3-5;

Chapters 5-7). In fact, warnings seem to be an ineffective method to deter people from consuming negative content. Further research is required to unpack the claims surrounding the potential benefits and harms of avoidance—but whatever the case, trigger warnings do not seem to enhance or decrease avoidance behavior.

If trigger warnings don't work as intended, why do people think they do?

If laboratory studies fail to find that trigger warnings mitigate negative emotional reactions, one might assume that after a few experiences with trigger warnings in real life, advocates would learn that they do little to help people emotionally. For instance, people may have continued experiences where they see trigger warnings, view content, and still experience distress. So then why do people still so firmly believe that warnings alleviate distress?

Illusion of control. A possible avenue for future research might be to investigate the relationship between a belief in the efficacy of trigger warnings and the illusion of control—when someone perceives a chance event as controllable (Langer, 1975). There is a random chance of experiencing distress when coming across negative material for any one person depending on any number of individual characteristics, whether the material is introduced with a trigger warning or not. However, warning advocates may *believe* that seeing a trigger warning gives them more control over this outcome, despite the fact that no evidence supports this claim. Indeed, past research shows that people believe they have an illusion of control over completely chance events (e.g., outcomes when gambling; Goffman, 1967; Henslin, 1967), so it is not a stretch to assume that people might fail to predict complex emotional reactions. In fact, importantly, research on affective forecasting shows that people are often ineffective at predicting future affective consequences (e.g., how long someone expects to feel negative after their favorite sports team loses) and also do not learn from previous forecasting errors (Meyvis, Ratner, Levav, 2010).

Confirmation bias. Another possible explanation is that confirmation bias occurs when advocates recall experiences with and without trigger warnings. Previous research shows that memory often distorts to match with pre-existing beliefs (Frost et al., 2015). For instance, people can more accurately recall the content of articles if it is consistent with prior belief systems (e.g., views on gun control; Frost et al., 2015). Advocates believe that trigger warnings lead to emotional benefits, so when they view distressing content preceded by a warning message and feel distressed, they may falsely attribute the distress to the content *only* and not to the failure of the trigger warning. Conversely, when they come across content without a warning, they may attribute their distress to the *absence* of a warning message. Then, when retroactively recalling these two contrasting scenarios, advocates may falsely conclude that they have been less distressed in the past when they have come across content accompanied by trigger warnings versus unaccompanied content.

Disconnect between rights and benefits. "*Informed consent is supposed to be a good thing, isn't it? Motherhood, apple pie, and informed consent*" wrote Loftus and Fries (2008, p. 217). This quote reflects the now-culturally engrained notion that being fully informed about potentially negative outcomes is always a positive thing. These ideas emerged as a push back against historical cases of harm caused by medical professionals and researchers who did not fully inform patients and participants about the risk associated with medical procedures or experiments (Fries & Loftus, 1979). As a result, it is generally believed that people have a right to information regarding outcomes that might affect their mind and body. But, as Loftus and Fries (2008) point out, there is a disconnect between rights and benefits. Similar to the emerging research about trigger warnings showing that warnings have little benefit, numerous studies provide evidence that giving participants information about the risks of medical procedures exacerbates rather than alleviates negative outcomes (see Benedetti et al., 2007 for review). Interestingly, Bruce and Roberts (2020) found that students' desire for trigger warnings to be added to class content was motivated by their belief in institutional betrayal—the idea that their institution has not proven trustworthy in valuing student safety and wellbeing. That is, students who want trigger warnings likely feel that warnings help return decisions related to wellbeing to their own hands, rather than leaving the decision to institutions that they do not trust. These ideas link closely with the idea that individuals have the right to receive fully informed consent to protect themselves from medical institutions that may do them harm. However, as pointed out, these *rights* do not equate to *benefits*.

Trigger warnings and official policy

Informed consent procedures

My findings have important implications for the current recommendations about conveying participation risks in psychological research: warning messages do not prevent distress and in some cases may lead to harm. Although I am not suggesting that warnings should be removed from the consent process altogether, I would urge that researchers "pay some attention to the harm that may be caused by the ritual itself" (Loftus & Fries, 2008, p.g., 217). Perhaps information about how nocebo effects work should be included as part of informed consent procedures, or statements of potential harms should be reframed in terms of the actual risk posed (e.g., that participation is no riskier than everyday life). Indeed, one preliminary test of this idea found that informing participants about nocebo effects as part of the informed consent process helped to reduce the negative side effects of an experimental drug (Loftus & Fries, 2008).

Who are they protecting? The individual or the institution? A final remaining question is to ask: *who* are trigger warnings really designed to protect? The person/institution issuing the trigger warning? Or the recipients of trigger warnings? It could be argued that people use trigger warnings not only to protect people who see them, but also to protect themselves from criticism, social pressure/attack, and even legal action. On an individual level, someone may choose to issue a trigger warning (e.g., on a social media post containing distressing information) because if they don't, they fear they will be criticized. For instance, they may be accused of not caring for the needs, experiences and emotions of others. Indeed, a common reason why faculty and academic staff state that they use trigger warnings is to communicate a message that *they* care and support students (Boysen et al., 2016). Another potential reason is that people may use trigger warnings as a form of "virtue signaling" or "moral grandstanding"—publicly communicating an opinion to demonstrate that you are morally respectable or a good person (Tosi & Warmke, 2016). As one reddit user writes on the subject: "*It just makes you look like you care, without actually doing anything to help"* (u/someoneman, 2017). Institutions and companies may also decide to issue warning messages for similar reasons. Netflix for instance retroactively added trigger warnings to the series *13 Reasons Why* after public outcry (Saint Louis, 2017). Since then, similar calls have been made to add warnings to shows such as *Bridgeton* (Jean-Philippe, 2020) and *The Crown* (Cripps, 2020) and a recent petition argues for mandated trigger warnings on all potentially distressing content on Netflix (Medhora, 2021). Trigger warnings in this context may be used as evidence that *Netflix* cares about its users and their emotional needs and to prevent public backlash. Aside from public criticism, institutions and companies may also use trigger warnings to avoid legal repercussions. Instagram, Facebook, TikTok and Twitter faced potential fines and complete banning of use in the UK in 2020 when new laws were introduced regarding the failure to regulate harmful content (Browne, 2020). Since then, the sites have introduced measures to either remove, or add trigger warnings to, graphic content. Thus, sometimes trigger warnings appear to be used as a tool to protect the individual (or institution) issuing the warning from harm, rather than the person on the receiving end.

Trigger warnings: Leaving people with the illusion of help but no real benefits? A real danger in individuals and institutions relying on the illusory benefits of trigger warnings is that serious harm may occur if they are used as a primary mental health safeguard. A continued push to promote the use of trigger warnings persists in online discourse despite mounting evidence that they are ineffective. For instance, Strothman (2021) states that even though research suggests warnings are not helpful, she *"feel[s] strongly that we need to keep using them*." Similarly, Dowling (2021) acknowledges that previous research has found that trigger warnings do not have obvious benefits, but concludes that "*[b]y prefacing content with a warning, people are given the option to protect their mental health.*" This continued belief in the illusory of trigger warnings could result in two potential harms. First, for individuals trigger warnings may become a "box-ticking exercise" (Hay, 2019) or a "sticker-fix" (Fagan, 2019). That is, some people may think that adding a trigger warning to their content—whether that be a university lecture or a social media post—might absolve them of making any other efforts to present distressing material in a conscientious way. As Jones points out in Hay (2019): "*A professor who is otherwise clumsy in their words and lazy in their methods of teaching could say, 'oh, I gave a trigger warning and students still ran out of the room crying or made a scene;*

I don't understand it.'" On a more macro level, continued beliefs about the benefits of trigger warnings could result in reduced efforts by policy makers or institutions to find efficacious mental health support strategies, because trigger warnings may be considered one such approach already in use. Furthermore, fewer resources could be funneled into research concerning the effects of trigger warnings or how to build better warning systems if people continue to hold the general belief that warnings are helpful. We must first recognize and accept that trigger warnings are ineffective, if we are to arrive at viable alternatives. The next wave of research should focus on developing effective methods that achieve the commendable aims of trigger warnings—these possibilities are discussed below.

8.4 Limitations and future directions

The long-term effects of trigger warnings

First, a limitation of my thesis and work on trigger warnings as a whole is that we do not yet have an understanding about how encounters with trigger warnings may vary over time. Some people argue that they interact with content marked with a trigger warning differently depending on their mood or how they are feeling on any particular day. For instance, a trauma survivor may not feel like viewing content related to their traumatic experience if they have been experiencing more severe PTSD symptoms but may approach content if they have felt like they have been coping better. As a trauma survivor quoted in Medhora (2021) states: "*it's not about tuning that content out of your life; it's about picking the days when you can handle it and when you can't.*" A potential way to investigate this issue would be to conduct a diary study where participants record encounters with trigger warnings on a day-to-day basis and record mood and PTSD symptoms. Similarly, a longitudinal design could be employed where participants return to the lab on multiple occasions to interact with material with trigger warnings to see if fluctuations in mood and symptoms change how they behave and react. Alternatively, an experimental approach could involve manipulating mood to examine how mood may change how someone approaches or avoids, or emotionally reacts to, warned material.

Second, we also do not know about the potentially negative accumulative effects of trigger warnings over extended time periods. While my research in Study 2 (Chapter 4) is the first to examine the effects of trigger warnings beyond a single experimental session, further research should examine trigger warning exposure over a longer period of time. Indeed, while small effects may not be very consequential in a single episode (e.g., anxiety caused by viewing a warning), they may matter in the long run (Funder & Ozer, 2019). Consider one setting: an average adult spends three hours and 30 minutes per day on a mobile device (Molla, 2020), equating to 53 full days a year, viewing thousands of online posts and articles, a proportion of which contain trigger warning messages. Over time, small negative effects caused by warning messages online, such as anticipatory anxiety (Bridgland et al., 2019), enhanced event centrality (Jones et al., 2020), and memory distortion, may accumulate and have significant consequences. Therefore, investigating the accumulative effects of warnings may be an important next step.

Types of trigger warnings

As discussed in my introduction, trigger warnings have evolved significantly in form, function, and domain of use. While I explored a range of trigger warning types in my studies—including standard trigger warnings that warn of upcoming distressing material (Studies 1, 2 and 4; Chapters 3, 4, and 6), warnings specific to the recall of distressing memories (Study 2; Chapter 4), and visual content specific warnings as used on Instagram (Study 5; Chapter 7)—there are still many other variations that should be investigated. For instance, a distinction could be made between content warnings that only state the content of the upcoming material (e.g., "This program contains depictions of sexual assault"), versus a trigger warning that also describes the potential emotional harm that may be experienced (e.g., "This program contains depictions of sexual assault that some people may find distressing or triggering). Additionally, some authors have argued that the term "trigger warning" itself may lead to harm because it relates to violent weaponry and is therefore threatening (Doney, 2019; Stringer, 2016). Future research could therefore investigate if the alternative suggested term "content forecast" (Doney, 2019; Stringer, 2016) reduces the anxiety people feel, when they see an alert about upcoming content, versus the term "trigger warning." Further, future research should also try and develop trigger warnings that actually help to *alleviate* distress. For instance, if trigger warnings used some of the language found in emotion regulation instructions (e.g., re-appraisal such as reminding yourself to focus on non-emotional aspects of the situation such as factual informational; Shiota & Levenson, 2009).

Clinical populations

Another limitation of my work and the work of others is that no research has specifically recruited participants with a diagnosis of PTSD as validated by a clinical interview or clinical diagnosis (e.g., The Clinician-Administered PTSD Scale for DSM-5: CAPS-5; Weathers et al., 2013). Current trigger warning research, including my own, uses questionnaires (e.g., the PCL-5) to assess the *probability* that a participant would qualify for a clinical PTSD diagnosis. Although the PCL-5 shows good convergent validity with the measures like the CAPS-5 (Bovin et al., 2016), it is possible that my results would have been different had I specifically recruited and powered my samples for participants who had a clinical diagnosis of PTSD.

Relatedly, but on a different note, although extant research has focused on participants with a history of trauma and probable PTSD, little research has focused on exploring the reactions of other clinical populations. While I explored self-triggering (Study 5; Chapter 7), depression (Study 5), trait anxiety (Studies 4 and 5; Chapters 6 and 7), and wellbeing (Study 5), there are other noteworthy populations that regularly use trigger warnings who should also be investigated. In particular, people regularly use trigger warnings to flag content related to eating disorders to help people avoid this content and "trigger" disordered eating behaviors (e.g., Cripps, 2020). Yet, anecdotal reports suggest that these populations may actually use trigger warnings to find content that motivates them to lose more weight (Hack, 2017). However, no research has specifically investigated how trigger warnings might work amongst these populations.

8.5 Conclusion

My thesis aimed to bridge gaps left by the first wave of trigger warning research. Overall, I found that when someone sees a trigger warning, it causes an anxious anticipatory period that does not seem to reflect emotional preparation to mitigate distressing reactions. In fact, my thesis and the work of others unanimously suggests that trigger warnings *do not* mitigate distressing reactions. Trigger warnings may actually lead to emotional harm, and my thesis suggests when and how this harm may occur. Trigger warnings also do not seem to be an effective method of deterring the majority of people from engaging with distressing stimuli. Although there is a lot of questions remaining that warrant further investigation, my findings suggest that trigger warnings should not be relied upon as a beneficial mental health tool.

References

Abu-Rus, A., Bussell, N., Olsen, D. C., Davis-Ku, M. A., Alissa L., & Arzoumanian, M. A. (2018). Informed consent content in research with survivors of psychological trauma. *Ethics & Behavior*, *29*(8), 595–606. https://doi.org/10.1080/10508422.2018.1551802

Adler, O., & Pansky, A. A “Rosy View” of the Past: Positive Memory Biases. (2020). In Aue, T. & Okon-Singer, H. (Eds.), *Cognitive Biases in Health and Psychiatric Disorders.* Elsevier.

Altis, K.L., Elwood L.S., Olatunji B.O. (2014) Ethical Issues and Ethical Therapy Associated with Anxiety Disorders. In: Lee, G., Illes J., Ohl F. (Eds.) *Ethical Issues in Behavioral Neuroscience.* Current Topics in Behavioral Neurosciences, (vol. 19, pp. 265–278). Springer, Berlin, Heidelberg. https://doi.org/10.1007/7854_2014_340

Antony, M. M., Bieling, P. J., Cox, B. J., Enns, M. W., & Swinson, R. P. (1998). Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample. *Psychological Assessment*, *10*(2), 176–181. https://doi.org/10.1037/1040-3590.10.2.176

Auerbach, R. P., Mortier, P., Bruffaerts, R., Alonso, J., Benjet, C., Cuijpers, P., Demyttenaere, K., Ebert, D. D., Green, J. G., Hasking, P., Murray, E., Nock, M. K., Pinder-Amaker, S., Sampson, N. A., Stein, D. J., Vilagut, G., Zaslavsky, A. M., Kessler, R. C., & WHO WMH-ICS Collaborators (2018). WHO World Mental Health Surveys International College Student Project: Prevalence and Distribution of Mental Disorders. *Journal of Abnormal Psychology*. *127*(7), 623–638. https://doi.org/10.1037/abn0000362.

Badour, C. L., Blonigen, D. M., Boden, M. T., Feldner, M. T., & Bonn-Miller, M. O. (2012). A longitudinal test of the bi-directional relations between avoidance coping and PTSD severity during and after PTSD treatment. *Behaviour Research and Therapy*, *50*(10), 610–616. https://doi.org/10.1016/j.brat.2012.06.006

Bai, H. (2018, August 8). Evidence that a large amount of low quality responses on MTurk can be detected with repeated GPS coordinates. *MaxHubai.com.* Retrieved from http://www.maxhuibai.com/1/post/2018/08/evidence-that-responses-from-repeating-gps-are-random.html

Bartels, D. J., van Laarhoven, A. I., van de Kerkhof, P. C., & Evers, A. W. (2016). Placebo and nocebo effects on itch: effects, mechanisms, and predictors. *European journal of pain*, *20*(1), 8–13. https://doi.org/10.1002/ejp.750

Bartsch, A., & Mares, M. (2014). Making Sense of Violence: Perceived Meaningfulness as a Predictor of Audience Interest in Violent Media Content. *Journal of Communication*, *64*(5), 956–976. http://dx.doi.org/10.1111/jcom.12112

Baumgartner, S., & Wirth, W. (2012). Affective Priming During the Processing of News Articles. *Media Psychology*, *15*(1), 1–18. https://doi.org/10.1080/15213269.2011.648535

Bedo, S. (2020, September 8). Kids left traumatised after shock suicide video. Retrieved from https://www.news.com.au/technology/online/social/ronnie-mcnutt-suicide-video-leaves-kids-traumatised-after-platforms-struggle-to-remove-it/news-story/1dea8b68b23fcbdc5134e4725a607bd5

Bellet, B. W., Jones, P. J., & McNally, R. J. (2020). Self-Triggering? An Exploration of Individuals Who Seek Reminders of Trauma. *Clinical Psychological Science*, *8*(4), 739–755. https://doi.org/10.1177/2167702620917459

Bellet, B. W., Jones, P. J., Meyersburg, C. A., Brenneman, M. M., Morehead, K. E., & McNally, R. J. (2020). Trigger warnings and resilience in college students: A preregistered replication and extension. *Journal of Experimental Psychology: Applied*, *26*(4), 717–723. https://doi.org/10.1037/xap0000270

Bellet, B.W., Jones, P.W., & McNally, R.J. (2018). Trigger warning: Empirical evidence ahead. *Journal of Behavior Therapy and Experimental Psychiatry*, *61,* 134–141. https://doi.org/10.1016/j.jbtep.2018.07.002

Benedetti, F., Lanotte, M., Lopiano, L., & Colloca, L. (2007). When words are painful: Unraveling the mechanisms of the nocebo effect. *Neuroscience*, *147*(2), 260–271. https://doi.org/10.1016/j.neuroscience.2007.02.020

Benjet, C., Bromet, E., Karam, E. G., Kessler, R. C., McLaughlin, K. A., Ruscio, A. M., Shahly, V., Stein, D. J., Petukhova, M., Hill, E., Alonso, J., Atwoli, L., Bunting, B., Bruffaerts, R., Caldas-de-Almeida, J. M., de Girolamo, G., Florescu, S., Gureje, O., Huang, Y., Lepine, J. P., … Koenen, K. C. (2016). The epidemiology of traumatic event exposure worldwide: results from the World Mental Health Survey Consortium. *Psychological medicine*, *46*(2), 327–343. https://doi.org/10.1017/S0033291715001981

Bentley, M. (2017). Trigger warnings and the student experience. *Politics*, *37*(4), 470–485. https://doi.org/10.1177/0263395716684526

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2013). Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys. *American Journal of Political Science, 58*(3), 739–753. https://doi.org/10.1111/ajps.12081

Berinsky, A. J., Margolis, M. F., Sances, M. W., & Warshaw, C. (2019). Using screeners to measure respondent attention on self-administered surveys: Which items and how many?

Political Science Research and Methods, *9*(2), 430–437. https://doi.org/10.1017/psrm.2019.53

Berntsen, D. (1996). Involuntary autobiographical memories. *Applied Cognitive Psychology*, *10*, 435–454. https://doi.org/10.1002/(SICI)1099-0720(199610)10:5<435::AID-ACP408>3.0.CO;2-L

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*(2), 219–231. https://doi.org/10.1016/j.brat.2005.01.009

Berntsen, D., & Rubin, D. C. (2014). Pretraumatic Stress Reactions in Soldiers Deployed to Afghanistan. *Clinical Psychological Science*, *3*(5), 663–674. https://doi.org/10.1177/2167702614551766

Bettencourt, B. A., & Manning, M. (2016). Negatively valenced expectancy violation predicts emotionality: A longitudinal analysis. *Emotion*, *16*(6), 787–791. https://doi.org/10.1037/emo0000152

Bijvank, M. N., Konijn, E. A., Bushman, B. J., & Roelofsma, P. H. M. P. (2009). Age and violent-content labels make video games forbidden fruits for youth. *Pediatrics*, *123*(3), 870–876. https://doi.org/10.1542/peds.2008-0601

Boals, A., & Ruggero, C. (2015). Event centrality prospectively predicts PTSD symptoms. *Anxiety, Stress, & Coping*, *29*(5), 533–541. https://doi.org/10.1080/10615806.2015.1080822

Boals, A., Hathaway, L. M., & Rubin, D. C. (2011). The Therapeutic Effects of Completing Autobiographical Memory Questionnaires for Positive and Negative Events: An

Experimental Approach. *Cognitive Therapy and Research*, *35*(6), 544–549. https://doi.org/10.1007/s10608-011-9412-9

Bonanno, G. A., Pat-Horenczyk, R., & Noll, J. (2011). Coping flexibility and trauma: The Perceived Ability to Cope With Trauma (PACT) scale. *Psychological Trauma: Theory, Research, Practice, and Policy*, *3*(2), 117–129. https://doi.org/10.1037/a0020921

Bovin, M. J., Marx, B. P., Weathers, F. W., Gallagher, M. W., Rodriguez, P., Schnurr, P. P., & Keane, T. M. (2016). Psychometric properties of the PTSD Checklist for Diagnostic and Statistical Manual of Mental Disorders–Fifth Edition (PCL-5) in veterans. *Psychological Assessment*, *28*(11), 1379–1391. https://doi.org/10.1037/pas0000254

Bower, G. H. (1987). Commentary on mood and memory. *Behaviour Research and Therapy*, *25*(6), 443–455. https://doi.org/10.1016/0005-7967(87)90052-0

Boysen, G. A., Isaacs, R. A., Tretter, L., & Markowski, S. (2021). Trigger warning efficacy: The impact of warnings on affect, attitudes, and learning. *Scholarship of Teaching and Learning in Psychology*. *7*(1), 39–52. https://doi.org/10.1037/stl0000150

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*(4), 334–339. https://doi.org/10.1177/0098628316662766

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

Brainerd, C. J., Stein, L. M., Silveira, R. A., Rohenkohl, G., & Reyna, V. F. (2008). How Does Negative Emotion Cause False Memories? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1159170

Brashers, D. E., & Hogan, T. P. (2013). The appraisal and management of uncertainty: Implications for information-retrieval systems. *Information Processing & Management*, *49*(6), 1241–1249. https://doi.org/10.1016/j.ipm.2013.06.002

Bridgland, V. M. E., & Takarangi M. K. T. (2021). Danger! Negative memories ahead: The effect of warnings on reactions to and recall of negative memories. *Memory*, *29*(3), 319–329. https://doi.org/10.1080/09658211.2021.1892147

Bridgland, V. M. E., Barnard, & Takarangi, M. K. T. (2021). Unprepared: Thinking of a trigger warning does not prompt preparation for trauma-related content. Under review. See https://osf.io/7n85z/ for information.

Bridgland, V. M. E., Bellet, W. B., & Takarangi, K.T. M. (2021). Curiosity disturbed the cat: Instagram's Sensitive Content Screens do not deter vulnerable users from viewing distressing content. Under review. See https://osf.io/rj987/ for information.

Bridgland, V. M. E., Green, D. M., Oulton, J. M., & Takarangi, M. K. T. (2019). Expecting the worst: Investigating the effects of trigger warnings on reactions to ambiguously themed photos. *Journal of Experimental Psychology: Applied*, *25*(4), 602–617. https://doi.org/10.1037/xap0000215

Bridgland, V. M. E., Moeck, E. K., Green, D. M., Swain, T. L., Nayda, D., Matson, L. A., … Takarangi, M. K. T. (2020). Why the COVID-19 pandemic is a traumatic stressor. *PLOS ONE*, https://doi.org/10.1101/2020.09.22.307637

Browne, R. (2020, December 15). Social media giants face big fines and blocked sites under new UK rules on harmful content. *CNBC*. Retrieved from https://www.cnbc.com/2020/12/15/uk-online-harms-bill-tech-giants-face-big-fines-and-blocked-sites.html

Bruce, M. & Roberts D. (2020). Trigger warnings for abuse impact reading comprehension in students with histories of abuse. *College Student Journal. 54*(2), 157–168.

Bruce, M. & Roberts D. (2020). Trigger warnings in context: The role of institutional betrayal in the trigger warning debate. *College Student Journal*, *54*(4), 484–490.

Buhrmester, M., Kwang, T., & Gosling, S. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?. *Perspectives on Psychological Science*, *6*(1), 3-5. http://doi.org/10.1177/1745691610393980

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

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

Cares A. C., Franklin C. A., Fisher B. S., & Bostaph, L. G. (2019) "They Were There for People Who Needed Them": Student Attitudes Toward the Use of Trigger Warnings in Victimology Classrooms, *Journal of Criminal Justice Education*, *30*(1), 22–45, https://doi.org/10.1080/10511253.2018.1433221

Carleton, R. N. (2012). The intolerance of uncertainty construct in the context of anxiety disorders: theoretical and practical perspectives. *Expert Review of Neurotherapeutics*, *12*(8), 937–947. https://doi.org/10.1586/ern.12.82

Carlson, E. B., Smith, S. R., Palmieri, P. A., Dalenberg, C., Ruzek, J. I., Kimerling, R., … Spain, D. A. (2011). Development and validation of a brief self-report measure of trauma exposure: The Trauma History Screen. *Psychological Assessment, 23*(2), 463–477. https://doi.org/10.1037/a0022294

Carter-Visscher, R., Naugle, A., Bell, K., & Suvak, M. (2007). Ethics of Asking Trauma-Related Questions and Exposing Participants to Arousal-Inducing Stimuli. *Journal Of Trauma & Dissociation*, *8*(3), 27–55. https://doi.org/10.1300/j229v08n03_03

Carter, A. M. (2015). Teaching with Trauma: Disability Pedagogy, Feminism, and the Trigger Warnings Debate. *Disability Studies Quarterly*, *35*(2). https://doi.org/10.18061/dsq.v35i2.4652

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers In Human Behavior*, *29*(6), 2156–2160. https://doi.org/10.1016/j.chb.2013.05.009

Collins, A., & Loftus, E. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295x.82.6.407

Copoc, P. (2021). Trigger warnings: A quantitative study on the stigmatization of individuals with a mental illness and university students' help-seeking intentions. *SURG Journal*, *13*(1). https://doi.org/10.21083/surg.v13i1.6314

Crawford, J. R., & Henry, J. D. (2004). The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, *43*(3), 245–265. https://doi.org/10.1348/0144665031752934

Cripps, C. (2020, November 16). The Crown was right to warn viewers of Princess Diana's bulimia – it gave me the chance to prepare myself. *The Independent*. Retrieved from https://www.independent.co.uk/arts-entertainment/netflix/the-crown/princess-diana-bulimia-the-crown-b1721340.html

Cromer, L., Freyd, J., Binder, A., DePrince, A., & Becker-Blease, K. (2006). What's the risk in asking? Participant reaction to trauma history questions compared with reaction to other personal questions. *Ethics & Behavior*, *16*(4), 347–362. https://doi.org/10.1207/s15327019eb1604_5

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

Cunic, A. (2020, December 3). What Does It Mean to Be Triggered?. *Verywell mind*. Retrieved from https://www.verywellmind.com/what-does-it-mean-to-be-triggered-4175432

D. Sullivan, K. (2019, February 1). Five Ways to Take Control of Your Psychological Triggers. *I care for your brain with Dr. Sullivan*. Retrieved from https://www.icfyb.com/four-ways-to-take-control-of-your-psychological-triggers/

D'Argembeau, A., & Linden, M.V. (2006). Individual differences in the phenomenology of mental time travel: The effect of vivid visual imagery and emotion regulation strategies. *Consciousness and Cognition*, *15*, 342–350. https://doi.org/10.1016/j.concog.2005.09.001

D'Argembeau, A., Comblain, C., & Van Der Linden, M. (2003). Phenomenal characteristics of autobiographical memories for positive, negative, and neutral events. *Applied Cognitive Psychology*, *17*, 281–294. https://doi.org/10.1002/acp.856

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

DeBonis, K. (2019). Trigger Warnings in Medical Education, *Academic Medicine, 94*(6), 749. https://doi.org/10.1097/ACM.0000000000002681

Dodge, R., Daly, A., Huyton, J., & Sanders, L. (2012). The challenge of defining wellbeing. *International Journal of Wellbeing, 2*(3), 222–235. https://doi.org/10.5502/ijw.v2i3.4

Doney, M. (2019). Trigger warnings, trauma, and Teaching. *Academia*. Retrieved from https://www.academia.edu/38787683/Trigger_warnings_trauma_and_teaching

Dowling, K. (2021, April 5). It’s time to start using content warnings for course material. *The Peak*. Retrieved from https://the-peak.ca/2021/04/its-time-to-start-using-content-warnings-for-course-material/

Dugdale, J. R., Eklund, R. C., & Gordon, S. (2002). Expected and Unexpected Stressors in Major International Competition: Appraisal, Coping, and Performance. *The Sport Psychologist*, *16*(1), 20–33. https://doi.org/10.1123/tsp.16.1.20

Edwards, K., Kearns, M., Calhoun, K., & Gidycz, C. (2009). College Women's Reactions to Sexual Assault Research Participation: Is it Distressing?. *Psychology Of Women Quarterly*, *33*(2), 225–234. https://doi.org/10.1111/j.1471-6402.2009.01492.x

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

Ehlers, A., Hackmann, A., & Michael, T. (2004). Intrusive re-experiencing in post-traumatic stress disorder: Phenomenology, theory, and therapy. *Memory*, *12*(4), 403–415. https://doi.org/1010.1080/09658210444000025

Eisenberg, D., Downs, M. F., Golberstein, E., & Zivin, K. (2009). Stigma and Help Seeking for Mental Health Among College Students. *Medical Care Research and Review*, *66*(5), 522–541. https://doi.org/10.1177/1077558709335173

Essig, L. (2014, March 10). Trigger Warnings Trigger Me. *The Chronicle*. Retrieved from https://www.chronicle.com/blogs/conversation/2014/03/10/trigger-warnings-trigger-me/

Fagan, A. (2019, April 17). Do Trigger Warnings Actually Work?. *Psychology Today*. Retrieved from https://www.psychologytoday.com/au/blog/brainstorm/201904/do-trigger-warnings-actually-work

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/brm.41.4.1149

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191. https://doi.org/10.3758/BF03193146

Feingold, P. C., & Knapp, M. L. (1977). Anti-Drug Abuse Commercials. *Journal of Communication*, *27*(1), 20–28. https://doi.org/10.1111/j.1460-2466.1977.tb01792.x

Ferguson, M., Bargh, J., & Nayak, D. (2005). After-affects: How automatic evaluations influence the interpretation of subsequent, unrelated stimuli. *Journal of Experimental Social Psychology, 41*(2), 182–191. http://dx.doi.org/10.1016/j.jesp.2004.05.008

Field, A. (2005). *Discovering Statistics Using SPSS*. 2nd Edition, London: Sage.

Filipovic, J. (2016, March 6). We've gone too far with 'trigger warnings'. *The Guardian*. Retrieved from https://www.theguardian.com/commentisfree/2014/mar/05/trigger-warnings-can-be-counterproductive

Finkel, J. (2017, August 17). Art school under fire for bowing to transgender student complaints. *The Art Newspaper*. Retrieved from https://www.theartnewspaper.com/news/art-school-under-fire-for-bowing-to-transgender-student-complaints

Finset, A., Steine, S., Haugli, L., Steen, E., & Laerum, E. (2002). The Brief Approach/Avoidance Coping Questionnaire: Development and validation. *Psychology, Health & Medicine*, *7*(1), 75–85. https://doi.org/10.1080/13548500120101577

Flaherty, C. (2014, April 14). Oberlin backs down on 'trigger warnings' for professors who teach sensitive material. *Inside Higher Ed*. Retrieved from https://www.insidehighered.com/news/2014/04/14/oberlin-backs-down-trigger-warnings-professors-who-teach-sensitive-material

Flaherty, C. (2019, March 21). New study says trigger warnings are useless. Does that mean they should be abandoned?. *Inside Higher Education*. Retrieved from https://www.insidehighered.com/news/2019/03/21/new-study-says-trigger-warnings-are-useless-does-mean-they-should-be-abandoned

Flood, A. (2014, May 20). US students request 'trigger warnings' on literature. *The Guardian*. Retrieved from https://www.theguardian.com/books/2014/may/19/us-students-request-trigger-warnings-in-literature

Folkman, S. & Lazarus, R. S. (1985). If it changes it must be a process: Study of emotion and coping during three stages of a college examination. *Journal of Personality and Social Psychology, 48*, 150–170. https://doi.org/10.1037/0022-3514.48.1.150

Folkman, S., & Moskowitz, J. T. (2004). Coping: Pitfalls and Promise. *Annual Review of Psychology, 55*(1), 745–774. https://doi.org/10.1146/annurev.psych.55.090902.141456

Folkman, S., Lazarus, R. S., Dunkel-Schetter, C., DeLongis, A., & Gruen, R. (1986). The dynamics of a stressful encounter: Cognitive appraisal, coping and encounter outcomes. *Journal of Personality and Social Psychology, 50*(5), 992–1003. https://doi.org/10.1037/0022-3514.50.5.992

Friday, A. (August 31, 2016). What Our Culture Gets Wrong About Trigger Warnings. *Medium.* Retrieved 15 May 2020, from https://medium.com/the-establishment/what-our-culture-gets-wrong-about-trigger-warnings-52b868437fb0

Frost, P., Casey, B., Griffin, K., Raymundo, L., Farrell, C., & Carrigan, R. (2015). The Influence of Confirmation Bias on Memory and Source Monitoring. *The Journal of General Psychology*, *142*(4), 238–252. https://doi.org/10.1080/00221309.2015.1084987

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Gainsburg, I., & Earl, A. (2018). Trigger warnings as an interpersonal emotion-regulation tool: Avoidance, attention, and affect depend on beliefs. *Journal of Experimental Social Psychology, 79*, 252–263. https://doi.org/10.1016/j.jesp.2018.08.006

Geers, A. L., & Lassiter, G. D. (1999). Affective Expectations and Information Gain: Evidence for Assimilation and Contrast Effects in Affective Experience. *Journal of Experimental Social Psychology, 35*(4), 394–413. https://doi.org/10.1006/jesp.1999.1377

Geers, A. L., & Lassiter, G. D. (2005). Affective Assimilation and Contrast: Effects of Expectations and Prior Stimulus Exposure. *Basic and Applied Social Psychology*, *27*(2), 143–154. https://doi.org/10.1207/s15324834basp2702_5

George, E. & Hovey, A. (2020). Deciphering the trigger warning debate: a qualitative analysis of online comments, *Teaching in Higher Education*, *25*(7), 825–841, https://doi.org/10.1080/13562517.2019.1603142

Gilovich, T., Kerr, M., & Medvec, V. H. (1993). Effect of temporal perspective on subjective confidence. *Journal of Personality and Social Psychology*, *64*(4), 552–560. https://doi.org/10.1037/0022-3514.64.4.552

Goffman, E. (1967). Interaction ritual. New York: Anchor.

Goldhaber, G. M. & deTurck M. A. (1988), Effectiveness of Warning Signs: Gender and Familiarity Effects, *Journal of Products Liability*, *11*(3), 271–284.

Goldhber, G. M., & deTurck, M. A. (1989). A Developmental Analysis of Warning Signs: The Case of Familiarity and Gender. *Proceedings of the Human Factors Society Annual Meeting, 33*(15), 1019–1023. https://doi.org/10.1177/154193128903301525

Golub, S., Gilbert, D., & Wilson, T. (2009). Anticipating one's troubles: The costs and benefits of negative expectations. *Emotion*, 9(2), 277–281. https://doi.org/10.1037/a0014716

Good Therapy (2018, May 2). What is a trigger?. *Good Therapy*. Retrieved from https://www.goodtherapy.org/blog/psychpedia/trigger

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/s0092-6566(03)00046-1

Green, D. M., Strange, D., Lindsay, D. S., & Takarangi, M. K. T. (2016). Trauma-related versus positive involuntary thoughts with and without meta-awareness. *Consciousness and Cognition, 46*, 163–172. https://doi.org/10.1016/j.concog.2016.09.019

Gross, J. J. (1999). Emotion Regulation: Past, Present, Future. *Cognition & Emotion*, *13*(5), 551–573. https://doi.org/10.1080/026999399379186

Gross, J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, *39*(3), 281–291. https://doi.org/10.1017/s0048577201393198

Gross, J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology, 85*, 348–362. http://doi.org/10.1037/0022-3514.85.2.348

Gross, J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology, 85*, 348–362. http://doi.org/10.1037/0022-3514.85.2.348

Gust, O. (2016, June 14). I use trigger warnings - but I'm not mollycoddling my students. *The Guardian.* Retrieved from https://www.theguardian.com/higher-education-network/2016/jun/14/i-use-trigger-warnings-but-im-not-mollycoddling-my-students

Hack (2017, July 20). How trigger warnings could harm some people living with an eating disorder. *Hack*. Retrieved from https://www.abc.net.au/triplej/programs/hack/how-trigger-warnings-could-harm-people-living-with-an-eating-di/8728784

Hackmann, A., Ehlers, A., Speckens, A., Clark, D. M. (2004). Characteristics and content of intrusive memories in PTSD and their changes with treatment. *Journal of Traumatic Stress*, *17,* 231–240. https://doi.org/10.1023/B:JOTS.0000029266.88369.fd

Harper, C. (2018, July 29). It's Official-Trigger Warnings Might Actually Be Harmful. *Medium*. Retrieved from https://medium.com/@CraigHarper19/its-official-trigger-warnings-might-actually-be-harmful-3e8acaae098b

Harris, S. (2016, August 31). Think Trigger Warnings Are Never Mandatory on Campus? Think Again. *FIRE*. Retrieved from https://www.thefire.org/think-trigger-warnings-are-never-mandatory-on-campus-think-again/

Haslam, N. (2017, May 19). A short history of trigger warnings. *Psychlopaedia*. Retrieved from https://psychlopaedia.org/society/republished/whats-the-difference-between-traumatic-fear-and-moral-anger-trigger-warnings-wont-tell-you/

Hauser, D. J., & Schwarz, N. (2015). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z

Hauser, D. J., & Schwarz, N. (2015). It’s a Trap! Instructional Manipulation Checks Prompt Systematic Thinking on “Tricky” Tasks. *SAGE Open, 5*(2), 215824401558461. https://doi.org/10.1177/2158244015584617

Hay, M. (2019, May 10). Do Trigger Warnings Actually Work?. *Vice*. Retrieved from https://www.vice.com/en/article/wj9ba4/do-trigger-warnings-actually-work

Hayes, S. C., Strosahl, K., Wilson, K. G., Bissett, R. T., Pistorello, J., Toarmino, D., … McCurry, S. M. (2004). Measuring experiential avoidance: A preliminary test of a working model. *The Psychological Record*, *54*(4), 553–578. https://doi.org/10.1007/bf03395492

Henslin, J. M. (1967). Craps and Magic. *American Journal of Sociology*, *73*(3), 316–330. https://doi.org/10.1086/224479

Herr, P. M. (1986). Consequences of priming: Judgment and behavior. *Journal of Personality and Social Psychology*, *51*(6), 1106-1115. https://doi.org/10.1037/0022-3514.51.6.1106

Holmes, E. A., & Bourne, C. (2008). Inducing and modulating intrusive emotional memories: A review of the trauma film paradigm. *Acta Psychologica*, *127*(3), 553–566. https://doi.org/10.1016/j.actpsy.2007.11.002

Holmes, E. A., Brewin, C. R., & Hennessy, R. G. (2004). Trauma Films, Information Processing, and Intrusive Memory Development. *Journal of Experimental Psychology: General*, *133*(1), 3–22. https://doi.org/10.1037/0096-3445.133.1.3

Hsee, C. K., & Ruan, B. (2016). The Pandora Effect. *Psychological Science*, *27*(5), 659–666. https://doi.org/10.1177/0956797616631733

Hutchinson, A. (2017, March 24). Instragram Rolls Out New 'Sensitive Content' Filter Tool, New Safety Measures. *Social Media Today*. Retrieved from https://www.socialmediatoday.com/social-networks/instragram-rolls-out-new-sensitive-content-filter-tool-new-safety-measures

Hyland, M., & Birrell, J. (1979). Government Health Warnings and the “Boomerang” Effect. *Psychological Reports*, *44*(2), 643–647. https://doi.org/10.2466/pr0.1979.44.2.643

Innocent Lives Foundation. (2020, December 16). Importance of Trigger Warnings: The Innocent Lives Foundation. Retrieved from https://www.innocentlivesfoundation.org/importance-of-trigger-warnings/#:~:text=The%20American%20Psychological%20Association%20shares,intentionally%20thinks%20about%20their%20trauma

Instagram (2017, March 24). *Last September we made a commitment to the community to keep Instagram a safe place*. [Facebook] Available at https://www.facebook.com/instagram/posts/last-september-we-made-a-commitment-to-the-community-to-keep-instagram-a-safe-pl/1279646722121169/

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. https://doi.org/10.1016/j.cpr.2016.04.010

Janiszewski, C., & Wyer, R. (2014). Content and process priming: A review. *Journal of Consumer Psychology*, *24*(1), 96–118. https://doi.org/10.1016/j.jcps.2013.05.006

Jean-Philippe, M. (2020, December 29). Bridgerton's Controversial Sex Scene Needs a Trigger Warning. *Oprah Daily*. Retrieved from https://www.oprahdaily.com/entertainment/tv-movies/a35090027/bridgertons-controversial-sex-scene-episode-6/

Jeffreys, H. (1961). *Theory of probability (3rd Ed.).* Oxford, UK: Oxford University Press.

Jing, H., Madore, K., & Schacter, D. (2016). Worrying about the future: An episodic specificity induction impacts problem solving, reappraisal, and well-being. *Journal Of Experimental Psychology: General*, *145*(4), 402–418. https://doi.org/10.1037/xge0000142

Johnson, K., Lynch, T., Monroe, E., & Wang, T. (2015, April 30). Our identities matter in Core classrooms. *Columbia Spectator.* Retrieved from https://www.columbiaspectator.com/opinion/2015/04/30/our-identities-matter-core-classrooms/

Jones, P. J., Bellet, B. W., & McNally, R. J. (2019). Helping or harming? The effect of trigger warnings on individuals with trauma histories. *Clinical Psychological Science*, *8*(5), 905–917. https://doi.org/10.31219/osf.io/axn6z

Kamenetz, A. (2016, September 7). NPR Cookie Consent and Choices. *NPR*. Retrieved from https://www.npr.org/sections/ed/2016/09/07/492979242/half-of-professors-in-npr-ed-survey-have-used-trigger-warnings

Kashdan, T. B., Disabato, D. J., Goodman, F. R., & McKnight, P. E. (2020). The Five-Dimensional Curiosity Scale Revised (5DCR): Briefer subscales while separating overt and covert social curiosity. *Personality and Individual Differences*, *157*, 109836. https://doi.org/10.1016/j.paid.2020.109836

Kato, T. (2012). Development of the Coping Flexibility Scale: Evidence for the coping flexibility hypothesis. *Journal of Counseling Psychology*, *59*(2), 262–273. https://doi.org/10.1037/a0027770

Kessler, R.C., Berglund, P., Delmer, O., Jin, R., Merikangas, K.R., & Walters, E.E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, *62*(6), 593–602.

Kilpatrick, D. G., Resnick, H. S., Milanak, M. E., Miller, M. W., Keyes, K. M., & Friedman, M. J. (2013). National Estimates of Exposure to Traumatic Events and PTSD Prevalence Using DSM-IV and DSM-5 Criteria. *Journal of Traumatic Stress*, *26*(5), 537–547. https://doi.org/10.1002/jts.21848

Kilpatrick, D. G., Resnick, H. S., Milanak, M. E., Miller, M. W., Keyes, K. M., & Friedman, M. J. (2013). National Estimates of Exposure to Traumatic Events and PTSD Prevalence Using DSM-IV and DSM-5 Criteria. *Journal of Traumatic Stress*, *26*(5), 537–547. https://doi.org/10.1002/jts.21848

Kimble, M., Flack, W., Koide, J., Bennion, K., Brenneman, M., & Meyersburg, C. (2021). Student reactions to traumatic material in literature: Implications for trigger warnings. *PLOS ONE*, 16(3), e0247579. https://doi.org/10.1371/journal.pone.0247579

King, S. (1981). Danse macabre. New York, NY: Everest House

Kirsch, I. (1985). Response expectancy as a determinant of experience and behavior. *American Psychologist*, *40*(11), 1189–1202. http://doi.org/ 10.1037/0003-066X.40.11.1189

Klugman, C. (2017, 14 June). Trigger Warnings on College Campuses Are Censorship. *Pacific Standard*. Retrieved from https://psmag.com/education/trigger-warnings-on-college-campuses-are-nothing-but-censorship

Krieger, T., Zimmermann, J., Huffziger, S., Ubl, B., Diener, C., Kuehner, C., & Grosse Holtforth, M. (2014). Measuring depression with a well-being index: Further evidence for the validity of the WHO Well-Being Index (WHO-5) as a measure of the severity of depression. *Journal of Affective Disorders*, *156*, 240–244. https://doi.org/10.1016/j.jad.2013.12.015

Krypotos, A. M., Effting, M., Kindt, M., & Beckers, T. (2015). Avoidance learning: a review of theoretical models and recent developments. *Frontiers in Behavioral Neuroscience*, *9*, 189. http://doi.org/10.3389/fnbeh.2015.00189

Lakens, D. (2021). Sample Size Justification. https://doi.org/10.31234/osf.io/9d3yf

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8, The Center for Research in Psychophysiology, University of Florida, Gainesville, FL.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, *32*(2), 311–328. https://doi.org/10.1037/0022-3514.32.2.311

Lee, S. W. S., Oyserman, D., & Bond, M. H. (2010). Am I doing better than you? That depends on whether you ask me in English or Chinese: Self-enhancement effects of language as a cultural mindset prime. *Journal of Experimental Social Psychology*, *46*(5), 785–791. https://doi.org/10.1016/j.jesp.2010.04.005

LeMoult, J., Colich, N., Joormann, J., Singh, M. K., Eggleston, C., & Gotlib, I. H. (2017). Interpretation Bias Training in Depressed Adolescents: Near- and Far-Transfer Effects. *Journal of Abnormal Child Psychology, 46*(1), 159–167. https://doi.org/10.1007/s10802-017-0285-6

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

Levine L. J., Prohaska V., Burgess S. L., Rice J. A., & Laulhere T. M. (2001). Remembering past emotions: The role of current appraisals, *Cognition & Emotion*, *15*(4), 393–417, https://doi.org/10.1080/02699930125955

Levy, H. C., & Radomsky, A. S. (2013). Safety Behaviour Enhances the Acceptability of Exposure. *Cognitive Behaviour Therapy*, *43*(1), 83–92. http://doi.org/10.1080/16506073.2013.819376

Lin, H., Werner, K. M., & Inzlicht, M. (2020). Promises and perils of experimentation: The mutual internal validity problem. *Perspectives on Psychological Science*, *16*(4), 854–863. http://doi.org/10.31234/osf.io/hwubj

Littleton, H., Horsley, S., John, S., & Nelson, D. V. (2007). Trauma coping strategies and psychological distress: A meta-analysis. *Journal of Traumatic Stress*, *20*(6), 977–988. https://doi.org/10.1002/jts.20276

Lockhart, E. (2016). Why trigger warnings are beneficial, perhaps even necessary. *First Amendment Studies*, *50*(2), 59–69. https://doi.org/10.1080/21689725.2016.1232623

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*(1), 75–98. https://doi.org/10.1037/0033-2909.116.1.75

Loftus, E. (1979). Informed consent may be hazardous to health. *Science*, *204*(4388), 11. https://doi.org/10.1126/science.373117

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, *12*(4), 361–366. https://doi.org/10.1101/lm.94705

Loftus, E. F., & Teitcher, J. (2018). Invasion of the Mind Snatchers: A Nation Full of Traumatic Memories. *Clinical Psychological Science*, *7*(1), 25–26. https://doi.org/10.1177/2167702618797107

Logan, E. (2020, December 26). This Controversial 'Bridgerton' Scene Is Causing a Lot of Frustration on Twitter. *Glamour*. Retrieved from https://www.glamour.com/story/controversial-bridgerton-scene-sexual-assault-twitter-response

Longo, Y., Coyne, I., & Joseph, S. (2018). Development of the short version of the Scales of General Well-Being: The 14-item SGWB. *Personality and Individual Differences*, *124*, 31–34. https://doi.org/10.1016/j.paid.2017.11.042

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behavior Research and Therapy*, *33*(3), 335–343. https://doi.org/10.1016/0005-7967(94)00075-u

LSA Inclusive Teaching Initiative. (2020). *An Introduction to Content Warnings and Trigger Warnings – Inclusive Teaching*. Retrieved from https://sites.lsa.umich.edu/inclusive-teaching/inclusive-classrooms/an-introduction-to-content-warnings-and-trigger-warnings/

Lukianoff, G., & Haidt, J. (2015, September 1). The coddling of the American mind. *Atlantic*. Retrieved from http://www.theatlantic.com/magazine/archive/2015/09/the-coddling-of-the-american-mind/399356/

Maddocks, F. (2016, March 17). Warning: art may broaden the mind. It should never apologise. *The Guardian*. Retrieved from https://www.theguardian.com/music/2016/mar/16/royal-opera-house-lucia-di-lammermoor-content-warning

Mähönen, T. A., Leinonen, E., & Jasinskaja-Lahti, I. (2013). Met expectations and the wellbeing of diaspora immigrants: A longitudinal study. *International Journal of Psychology*, *48*(3), 324–333. https://doi.org/10.1080/00207594.2012.662278

Malervy, R. (2018, December 3). Castigating Trigger Warnings isn’t Only Hypocritical – It’s Absurd. *University Times.* Retrieved from http://www.universitytimes.ie/2018/12/castigating-trigger-warnings-isnt-only-hypocritical-its-absurd/

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

Marchewka, A., Żurawski, Ł., Jednoróg, K., & Grabowska, A. (2014). The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behavior Research Methods*, *46*(2), 596-610. https://doi.org/10.3758/s13428-013-0379-1

Marsh, B. U., Pezdek, K., & Lam, S. T. (2014). Imagination perspective affects ratings of the likelihood of occurrence of autobiographical memories. *Acta Psychologica*, *150*, 114-119. https://doi.org/10.1016/j.actpsy.2014.05.006

Marshall, M., & Brown, J. (2006). Emotional reactions to achievement outcomes: Is it really best to expect the worst? *Cognition & Emotion*, *20*(1), 43–63. https://doi.org/10.1080/02699930500215116

Marteau, T. M. & Bekker, H. (1992) The development of a six-item short-form of the State Scale of the Spielberger State-Trait Anxiety Inventory (STAI). *British Journal of Clinical Psychology, 31,* 301–306. https://doi.org/10.1111/j.2044-8260.1992.tb00997.x

McGraw, A. P., Mellers, B. A., & Ritov, I. (2004). The affective costs of overconfidence. *Journal of Behavioral Decision Making*, *17*(4), 281–295. https://doi.org/10.1002/bdm.472

McNally, R. (2014, June 14). Hazards Ahead: The Problem With Trigger Warnings. *Pacific Standard.* Retrieved from https://psmag.com/education/hazards-ahead-problem-trigger-warnings-according-research-81946

McNeil, K. (2015, October 26). The Importance of Trigger Warnings. *Tufts Observer.* Retrieved from http://tuftsobserver.org/the-importance-of-trigger-warnings/

Medhora, S. (2021, January 11). Does Netflix have a responsibility to provide trigger warnings on its content?. *Hack.* Retrieved from https://www.abc.net.au/triplej/programs/hack/netflix-trigger-warning-content-sexual-assault/13048904#:~:text=Source%3A%20supplied-,Caitlin%20Norman%20said%20she%20was%20surprised%20to%20learn%20that%20Netflix,trigger%20warnings%20on%20its%20content.&text=To%20be%20clear%2C%20Caitlin%20wasn,change%20their%20own%20viewing%20behaviour.

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

Mellers, B. A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision Affect Theory: Emotional Reactions to the Outcomes of Risky Options. *Psychological Science*, *8*(6), 423–429. https://doi.org/10.1111/j.1467-9280.1997.tb00455.x

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Penn State Worry Questionnaire. *PsycTESTS Dataset*. https://doi.org/10.1037/t01760-000

Miller, L., & Grandjean, A. (2019, August 19). The News and Mental Health: Would You Filter Out Bad News if You Could?. *BBC.* Retrieved from https://www.bbc.co.uk/rd/blog/2019-08-news-mood-filter-mental-health

Millgram, Y., Joormann, J., Huppert, J. D., & Tamir, M. (2015). Sad as a Matter of Choice? Emotion-Regulation Goals in Depression. *Psychological Science*, *26*(8), 1216–1228. https://doi.org/10.1177/0956797615583295

Molla, R. (2020, January 6). Americans spent about 3.5 hours per day on their phones last year — a number that keeps going up despite the "time well spent" movement. *Vox.* Retrieved from: https://www.vox.com/recode/2020/1/6/21048116/tech-companies-time-well-spent-mobile-phone-usage-data

Moos, R.H. (1993). The Coping Responses Inventory: An update on research and applications and validity – manual supplement. Lutz, FL: Psychological Assessment Resources.

Morgan, C. A., Southwick, S., Steffian, G., Hazlett, G. A., & Loftus, E. F. (2013). Misinformation can influence memory for recently experienced, highly stressful events.

International Journal of Law and Psychiatry, *36*(1), 11–17. https://doi.org/10.1016/j.ijlp.2012.11.002

Mosseri, A. (2019a, February 7). Instagram Policy Changes on Self-Harm Related Content - Protecting Vulnerable Users. *Instagram Blog*. Retrieved from https://about.instagram.com/blog/announcements/supporting-and-protecting-vulnerable-people-on-instagram

Mosseri, A. (2019b, October 27). Taking More Steps To Keep The People Who Use Instagram Safe. *Instagram Blog.* Retrieved from https://about.instagram.com/blog/announcements/more-steps-to-keep-instagram-users-safe

Myers, M. G., Cairns, J. A., & Singer, J. (1987). The consent form as a possible cause of side effects. *Clinical Pharmacology & Therapeutics*, *42*(3), 250–253. https://doi.org/10.1038/clpt.1987.142

National Coalition Against Censorship. (2015). Trigger warnings: A National College Educator Survey, *National Coalition Against Censorship,* Retrieved from https://ncac.org/resource/ncac-report-whats-all-this-about-trigger-warnings

Newton, J. W., & Hobbs, S. D. (2015). Simulating Memory Impairment for Child Sexual Abuse. *Behavioral Sciences & the Law*, *33*(4), 407–428. https://doi.org/10.1002/bsl.2197

Newman, E., Willard, T., Sinclair, R., & Kaloupek, D. (2001). Empirically supported ethical research practice: The costs and benefits of research from the participants' view. *Accountability In Research*, *8*(4), 309–329. https://doi.org/10.1080/08989620108573983

Oosterwijk, S. (2017). Choosing the negative: A behavioral demonstration of morbid curiosity. *PLOS ONE*, *12*(7), e0178399. https://doi.org/10.1371/journal.pone.0178399

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. https://doi.org/0.1016/j.jesp.2009.03.009

Oulton, J. M., & Takarangi, M. K. T. (2017). (Mis)remembering negative emotional experiences. In R. Nash & J. Ost (Eds.), *False and Distorted Memories* (pp. 9–22). Abingdon, UK: Psychology Press.

Oxford Dictionaries. (2018). *Trigger warning*. Definition of trigger warning in English. Retrieved from https://en.oxforddictionaries.com/definition/trigger_warning

Oxford Languages. (2021). *Prepare*. Definition of prepare in English. Retrieved from https://languages.oup.com/google-dictionary-en/

Palmer, T. (2017, March 28). Monash University trigger warning policy fires up free speech debate. *ABC News*. Retrieved from https://www.abc.net.au/news/2017-03-28/monash-university-adopts-trigger-warning-policy/8390264

Pennebaker, J.W., Booth, R.J., Boyd, R.L., & Francis, M.E. (2015). Linguistic Inquiry and Word Count: LIWC2015. Austin, TX: Pennebaker Conglomerates (www.LIWC.net).

Pickrell, J. E., McDonald, D. L., Bernstein, D. M., & Loftus, E. F. (2017). Misinformation effect. In R. F. Pohl (Ed.), Cognitive illusions: Intriguing phenomena in thinking, judgment and memory (pp. 406–423). Routledge/Taylor & Francis Group.

Porter, S., ten Brinke, L., Riley, S., & Baker, A. (2014). Prime time news: The influence of primed positive and negative emotion on susceptibility to false memories. *Cognition and Emotion, 28*(8), 1422–1434. http://dx.doi.org/10.1080/02699931.2014.887000

Psychology Software Tools, Inc. [E-Prime Go]. (2020). Retrieved from https://support.pstnet.com/.

Qualtrics software. (Copyright © 2018) Qualtrics and all other Qualtrics product or service names are registered trademarks or trademarks of Qualtrics, Provo, UT, USA. https://www.qualtrics.com

Qualtrics software. (Copyright © 2020). Qualtrics and all other Qualtrics product or service names are registered trademarks or trademarks of Qualtrics, Provo, UT, USA. https://www.qualtrics.com

Quintana, D. (2018). How to calculate statistical power for your meta-analysis. *Towards Data Science*. Retrieved from https://towardsdatascience.com/how-to-calculate-statistical-power-for-your-meta-analysis-e108ee586ae8

Rachman, S., Radomsky, A. S., & Shafran, R. (2008). Safety behaviour: A reconsideration. *Behaviour Research and Therapy*, *46*(2), 163–173. http://doi.org/10.1016/j.brat.2007.11.008

Rains, S. A., & Tukachinsky, R. (2014). An Examination of the Relationships Among Uncertainty, Appraisal, and Information-Seeking Behavior Proposed in Uncertainty Management Theory. *Health Communication*, *30*(4), 339–349. https://doi.org/10.1080/10410236.2013.858285

Rauch, S. A. M., Eftekhari, A., & Ruzek, J. I. (2012). Review of exposure therapy: A gold standard for PTSD treatment. *The Journal of Rehabilitation Research and Development, 49*(5), 679. https://doi.org/10.1682/jrrd.2011.08.0152

Redmond, S., Jones, N. M., Holman, E. A., & Silver, R. C. (2019). Who watches an ISIS beheading—And why. *American Psychologist*, *74*(5), 555–568. https://doi.org/10.1037/amp0000438

Rickwood, D. J., & Braithwaite, V. A. (1994). Social-psychological factors affecting help-seeking for emotional problems. *Social Science & Medicine*, *39*(4), 563–572. https://doi.org/10.1016/0277-9536(94)90099-x

Ringold, D. J. (2002). Boomerang effects in response to public health interventions: Some unintended consequences in the alcoholic beverage market. *Journal of Consumer Policy*, *25*(1), 27–63. https://doi.org/10.1023/a:1014588126336

Robbins, S. P. (2016). From the Editor—Sticks and Stones: Trigger Warnings, Microaggressions, and Political Correctness. *Journal of Social Work Education*, *52*(1), 1–5. https://doi.org/10.1080/10437797.2016.1116850

Rubin, D. C., & Feeling, N. (2013). Measuring the Severity of Negative and Traumatic Events. *Clinical Psychological Science*, *1*(4), 375–389. https://doi.org/10.1177/2167702613483112

Rubin, D. C., Boals, A., & Klein, K. (2008). Autobiographical Memories for Very Negative Events: The Effects of Thinking About and Rating Memories. *Cognitive Therapy and Research*, *34*(1), 35–48. https://doi.org/10.1007/s10608-008-9226-6

Rubin, D. C., Deffler, S. A., & Umanath, S. (2019). Scenes enable a sense of reliving: Implications for autobiographical memory. *Cognition*, *183*, 44–56. https://doi.org/1010.1016/j.cognition.2018.10.024

Rubin, M. (2021). When to adjust alpha during multiple testing: A consideration of disjunction, conjunction, and individual testing. *Synthese*, 1-32. https://doi.org/10.22541/au.162569018.85105015/v1

Saint Louis, C. (2017, May 2). Netflix to Add Warning to Start of '13 Reasons Why'. *New York Times.* Retrieved from https://www.nytimes.com/2017/05/02/arts/television/netflix-to-add-warning-to-start-of-13-reasons-why.html#:~:text=Netflix%20will%20add%20a%20warning,the%20streaming%20network%20announced%20Monday.

Salters-Pedneault, K., Tull, M. T., & Roemer, L. (2004). The role of avoidance of emotional material in the anxiety disorders. *Applied and Preventive Psychology*, *11*(2), 95–114. https://doi.org/10.1016/j.appsy.2004.09.001

Sanson, M., Strange, D., & Garry, M. (2019). Trigger warnings are trivially helpful at reducing negative affect, intrusive thoughts, and avoidance. *Clinical Psychological Science*, *7*(4), 778–793. https://doi.org/10.1177/2167702619827018

Savitsky, K., Medvec, V. H., Charlton, A. E., & Gilovich, T. (1998). "What, Me Worry?": Arousal, Misattribution, and the Effect of Temporal Distance on Confidence. *Personality and Social Psychology Bulletin*, *24*(5), 529–536. https://doi.org/10.1177/0146167298245008

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Schönbrodt, F. D., & Perugini, M. (2018). Corrigendum to "At what sample size do correlations stabilize?" [J. Res. Pers. 47 (2013) 609–612]. *Journal of Research in Personality*, *74,* 194. https://doi.org/10.1016/j.jrp.2018.02.010

Senn, C. Y., & Desmarais, S. (2006). A new wrinkle on an old concern: Are the new ethics review requirements for explicit warnings in consent forms affecting the results of sexuality research?. *Canadian Journal of Human Sexuality*, *15*.

Shafir, R., & Sheppes, G. (2020). How anticipatory information shapes subsequent emotion regulation. *Emotion*, *20*(1), 68–74. https://doi.org/10.1037/emo0000673

Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. Philosophical transactions of the Royal Society of London. Series B, *Biological sciences*, *362*(1481), 773–786. http://doi.org/10.1098/rstb.2007.2087

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, *1*(2), 213–220. https://doi.org/10.1177/2167702612469015

Sheeran, P. (2002) Intention-Behavior Relations: A Conceptual and Empirical Review, *European Review of Social Psychology*, *12*(1), 1–36. https://doi.org/10.1080/14792772143000003

Sheeran, P. & Webb, T. L. (2016). The Intention-Behaviour Gap, *Social and Personality Psychology Compass*, *10*(9), 503–518. https://doi.org/10.1111/spc3.12265

Shepperd, J. A., & McNulty, J. K. (2002). The Affective Consequences of Expected and Unexpected Outcomes. *Psychological Science*, *13*(1), 85–88. https://doi.org/10.1111/1467-9280.00416

Shepperd, J. A., Ouellette, J. A., & Fernandez, J. K. (1996). Abandoning unrealistic optimism: Performance estimates and the temporal proximity of self-relevant feedback. *Journal of*

Personality and Social Psychology, *70*(4), 844–855. https://doi.org/10.1037/0022-3514.70.4.844

Shepperd, J. A., Perez, S., Walker, D., & Findley-Klein, C. (2000). Bracing for Loss. *Journal of Personality and Social Psychology*. *78*(4), 620–34. https://doi.org/10.1037//0022-3514.78.4.620

Shiota, M. N., & Levenson, R. W. (2009). Effects of aging on experimentally instructed detached reappraisal, positive reappraisal, and emotional behavior suppression. *Psychology and Aging*, *24*(4), 890–900. https://doi.org/10.1037/a0017896

Shiota, M. N., & Levenson, R. W. (2012). Turn down the volume or change the channel? Emotional effects of detached versus positive reappraisal. *Journal of Personality and Social Psychology*, *103*(3), 416–429. https://doi.org/10.1037/a0029208

Siemer, M., Mauss, I., & Gross, J. J. (2007). Same situation--Different emotions: How appraisals shape our emotions. *Emotion*, *7*(3), 592–600. https://doi.org/10.1037/1528-3542.7.3.592

Silvia, P. J., & Kashdan, T. B. (2009). Interesting things and curious people: Exploration and engagement as transient states and enduring strengths. *Social and Personality Psychology Compass*, *3*(5), 785–797. https://doi.org/10.1111/j.1751-9004.2009.00210.x

Simister, E.., Bridgland, V. M. E., Takarangi M. K. T. (2021). An investigation of sensitive screens: Understanding when and why people approach sensitive content. See: https://osf.io/2fdr7

Simonpillai, R. (2021, May 20). Social media platforms are benefiting from activism and trauma. *Now Toronto.* Retrieved from https://nowtoronto.com/lifestyle/social-media-platforms-are-benefitting-from-activism-and-trauma

Southwick, S. M., Morgan, C. A., Nicolaou, A. L., & Charney, D. S. (1997). Consistency of memory for combat-related traumatic events in veterans of Operation Desert Storm. American Journal of Psychiatry, 154(2), 173-177. https://doi.org/10.1176/ajp.154.2.173

Spacapan, S., & Cohen, S. (1983). Effects and aftereffects of stressor expectations. *Journal of Personality and Social Psychology*, *45*(6), 1243–1254. https://doi.org/10.1037/0022-3514.45.6.1243

Speisman, J. C., Lazarus, R. S., Davison, L., & Mordkoff, A. M. (1964). Experimental analysis of a film used as a threatening stimulus. *Journal of Consulting Psychology*, *28*(1), 23–33. https://doi.org/10.1037/h0047028

Spielberger, C. D. (1983). State-Trait Anxiety Inventory for Adults. *PsycTESTS Dataset*. https://doi.org/10.1037/t06496-000

Stapel, D. A., & Noordewier, M. K. (2009). Stop Making Sense: The Ultimate Fear. *Psychological Inquiry*, *20*(4), 245–248. https://doi.org/10.1080/10478400903448516

Steger, M. F., Frazier, P., Oishi, S., & Kaler, M. (2006). The meaning in life questionnaire: Assessing the presence of and search for meaning in life. *Journal of Counseling Psychology*, *53*(1), 80–93. https://doi.org/10.1037/0022-0167.53.1.80

Stigler, S. (1997). Regression towards the mean, historically considered. *Statistical Methods In Medical Research*, *6*(2), 103–114. https://doi.org/10.1177/096228029700600202

Stringer, R. (2016). Reflection From the Field: Trigger warnings in university Teaching. *Women's Studies Journal*, *30*(2), 62–66.

Strothman, L. (2021). Trigger Warnings: How to protect yourself from online triggers. *Dr Lisa Strohman*. Retrieved from https://drlisastrohman.com/how-to-protect-yourself-from-online-triggers/

Sutin, A. R., & Robins, R. W. (2007). Phenomenology of autobiographical memories: The memory experiences questionnaire. *Memory*, 15(4), 390-411. https://doi.org/10.1080/09658210701256654

Suurmond R, van Rhee, H, & Hak T. (2017). Introduction, comparison and validation of Meta-Essentials: A free and simple tool for meta-analysis. *Research Synthesis Methods*. *8*(4), 537-553. https://doi.org/10.1002/jrsm.1260.

Sweeny, K. M., & Shepperd, J. A. (2007). Being the Best Bearer of Bad Tidings. *Review of General Psychology*, *11*(3). https://doi.org/10.1037/1089-2680.11.3.235

Sweeny, K., & Dillard, A. (2013). The Effects of Expectation Disconfirmation on Appraisal, Affect, and Behavioral Intentions. *Risk Analysis*, *34*(4), 711–720. https://doi.org/10.1111/risa.12129

Sweeny, K., & Shepperd, J. A. (2010). The costs of optimism and the benefits of pessimism. *Emotion*, *10*(5), 750–753. https://doi.org/10.1037/a0019016

Sweeny, K., Carroll, P. J., & Shepperd, J. A. (2006). Is Optimism Always Best? *Current Directions in Psychological Science*, *15*(6), 302–306. https://doi.org/10.1111/j.1467-8721.2006.00457.x

Sweeny, K., Reynolds, C., Falkenstein, A., Andrews, S., & Dooley, M. (2015). Two definitions of waiting well. *Emotion*, *16*(1), 129–143. https://doi.org/10.1037/emo0000117

Sweeny, K., Shepperd, J. A., & Carroll, P. J. (2009). Expectations for others' outcomes: Do people display compassionate bracing?. *Personality and Social Psychology Bulletin*, *35*(2), 160–171. https://doi.org/10.1177/0146167208327050

Takarangi, M. K. T, & Strange, D. (2010). Emotional impact feedback changes how we remember negative autobiographical experiences. *Experimental Psychology*, *57*(5), 354–359. https://doi.org/10.1027/1618-3169/a000042

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*(8), 1041–1051. https://doi.org/10.1080/09658211.2013.865238

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

Talarico, J. M., & Rubin, D. C. (2003). Confidence, not consistency, characterizes flashbulb memories. *Psychological Science*, *14*(5), 455–461. https://doi.org/10.1111/1467-9280.02453

Taylor, K. M., & Shepperd, J. A. (1998). Bracing for the Worst: Severity, Testing, and Feedback Timing as Moderators of the Optimistic Bias. *Personality and Social Psychology Bulletin*, *24*(9), 915–926. https://doi.org/10.1177/0146167298249001

The classification & Rating Administration (2021). *History of Ratings*. Retrieved from https://www.filmratings.com/History

Topp, C. W., Østergaard, S. D., Søndergaard, S., Bech, P. (2015). The WHO-5 Well-Being Index: A systematic review of the literature. *Psychotherapy and Psychosomatics*, *84*(3), 167–76. https://doi.org/10.1159/000376585

Tosi, J., & Warmke, B. (2016). Moral grandstanding. *Philosophy & Public Affairs*, *44*(3), 197–217. https://doi.org/10.1111/papa.12075.

Troy, A. S., Shallcross, A. J., Brunner, A., Friedman, R., & Jones, M. C. (2018). Cognitive reappraisal and acceptance: Effects on emotion, physiology, and perceived cognitive costs. *Emotion*, *18*(1), 58–74. https://doi.org/10.1037/emo0000371

u/someoneman. (2017). Are trigger warnings a form of virtue signalling? *Reddit*. Retrieved from https://www.reddit.com/r/TiADiscussion/comments/5527be/are_trigger_warnings_a_form_of_virtue_signalling/

University of St. Thomas (2015), Trigger Warnings in the classroom – Resources for UST faculty (Fall 2015). Retrieved from https://www.stthomas.edu/media/lyris/facultydevelopment/TriggerWarningHandout.pdf

User: Morbid-thoughts. (2003). I have a headache. *Live Journal.* Retrieved from https://morbid-thoughts.livejournal.com/5153.html

Valentine, J., Pigott, T., & Rothstein, H. (2010). How Many Studies Do You Need?. *Journal of Educational and Behavioral Statistics*, *35*(2), 215-247. https://doi.org/10.3102/1076998609346961

Van Dijk, W. W., & van der Pligt, J. (1997). The Impact of Probability and Magnitude of Outcome on Disappointment and Elation. *Organizational Behavior and Human Decision Processes*, *69*(3), 277–284. https://doi.org/10.1006/obhd.1997.2688

Van Rhee, H.J., Suurmond, R., & Hak, T. (2015). User manual for Meta-Essentials: Workbooks for meta-analysis (Version 1.2) Rotterdam, The Netherlands: Erasmus Research Institute of Management. Retrieved from www.erim.eur.nl/research-support/meta-essentials

Vigo, J. (2018, November 30). Trigger Warnings Perpetuate Victimhood. Forbes. Retrieved from https://www.forbes.com/sites/julianvigo/2018/11/30/trigger-warnings-perpetuate-victimhood/?sh=459855023526

Vingiano, A. (2014, May 5). How The "Trigger Warning" Took Over The Internet. *Buzzfeed News*. Retrieved from https://www.buzzfeednews.com/article/alisonvingiano/how-the-trigger-warning-took-over-the-internet

Waldman, K. (2016, September 5). The Trapdoor of Trigger Words. *Slate*. Retrieved from http://www.slate.com/articles/double_x/cover_story/2016/09/what_science_can_tell_us_about_trigger_warnings.html

Walker, D. (2021, April 4). Students at Aberdeen University vote for trigger warnings in lectures. *Evening Express*. Retrieved from https://www.eveningexpress.co.uk/fp/news/local/students-at-aberdeen-university-vote-for-trigger-warnings-in-lectures/

Walker, W. R., & Skowronski, J. J. (2009). The fading affect bias: But what the hell is it for?. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *23*(8), 1122-1136. https://doi.org/10.1002/acp.1614

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. https://doi.org/10.1037/0022-3514.54.6.1063

Weathers F. W., Blake D. D., Schnurr P. P., Kaloupek D. G., Marx B. P., Keane T. M. (2013). Clinician-Administered PTSD Scale for DSM–5 (CAPS-5). Boston, MA: *National Center for PTSD*. https://doi.org/10.1037/e572192010-003

Weiss, D. S. (2007). The impact of event scale: revised. In *Cross-cultural assessment of psychological trauma and PTSD* (pp. 219–238). Springer, Boston, MA. https://doi.org/10.1007/978-0-387-70990-1_10

Welsh, D., & Ordonez, L. (2014). The dark side of consecutive high performance goals: Linking goal setting, depletion, and unethical behavior. *Organizational Behavior and Human Decision Processes*, *123*(2), 79–89. https://doi.org/10.1016/j.obhdp.2013.07.006

Wilson, C. J., Deane, F. P., Ciarrochi, J., & Rickwood, D. (2005). General Help Seeking Questionnaire. *PsycTESTS Dataset*. https://doi.org/10.1037/t42876-000

Wilson, T. D., Lisle, D. J., Kraft, D., & Wetzel, C. G. (1989). Preferences as expectation-driven inferences: Effects of affective expectations on affective experience. *Journal of Personality and Social Psychology*, *56*(4), 519–530. https://doi.org/10.1037/0022-3514.56.4.519

Wilson, A. E., & Ross, M. (2001). From chump to champ: people's appraisals of their earlier and present selves. Journal of personality and social psychology, 80(4), 572. https://doi.org/10.1037/0022-3514.80.4.572

Wolgast, M., Lundh, L.-G., & Viborg, G. (2011). Cognitive reappraisal and acceptance: An experimental comparison of two emotion regulation strategies. *Behaviour Research and Therapy*, *49*(12), 858–866. https://doi.org/10.1016/j.brat.2011.09.011

Woodward, M. J., & Beck, J. G. (2017). Using the trauma film paradigm to explore interpersonal processes after trauma exposure. *Psychological Trauma: Theory, Research, Practice, and Policy*, *9*(4), 445–452. https://doi.org/10.1037/tra0000169

Yeater, E., Miller, G., Rinehart, J., & Nason, E. (2012). Trauma and Sex Surveys Meet Minimal Risk Standards: Implications for Institutional Review Boards. *Psychological Science*, *23*(7), 780–787. https://doi.org/10.1177/0956797611435131

Zillmann, D. (1988). Mood Management Through Communication Choices. American *Behavioral Scientist*, *31*(3), 327–340. https://doi.org/10.1177/000276488031003005

Appendix A—Self Assessment Manikins (SAM)

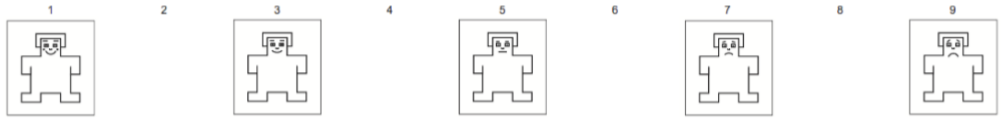


Figure 1. The 9-point valence dimension of the SAM depicting 5 figures (Bradley & Lang, 1994).

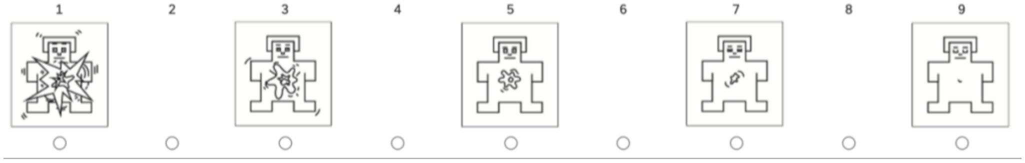
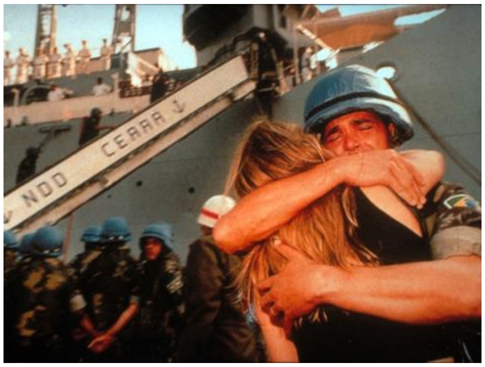


Figure 2. The 9-point arousal dimension of the SAM depicting 5 figures (Bradley & Lang, 1994).

Appendix B—Photo and Headline stimuli used in Studies 1a-1e

Example IAPS images used in Chapter 3 Studies 1a-1d



Example Shutterstock images used in Chapter 3 Study 1e

Example NAPS images used in Chapter 3 Study 1e

Table 1

Headline stimuli

Studies 1a-1d	
IAPS number	
1	8117
2	7620
3	4598
4	8400
5	8300
6	8190
Negative	
1	Hockey player in serious condition following spinal injury from Saturday's game
2	'I've lost everything' Mother takes photo of sons boarding plane shortly before fiery crash killing all
3	Seven soldiers dead, fresh explosions heard.
4	River rafting accident leaves man dead
5	Missouri pilot dies in plane crash after performing stunts
6	23-year-old Australian student, falls to death while skiing with friends at Whistler resort
Neutral	
1	New ice hockey stadium slated for city centre.
2	'Boeing starts shipping their new Dreamliners to airlines
3	Top 10 Best and Worst War films of all time.
4	Heavy rains put breaks on white water rafting on Kali River
5	Volunteers needed for aviation expo
6	Snow Watch: When it's coming and where it will be most heavy

	Study 1e
Shutterstock ID number/NAPS number	
1	Royalty-free stock photo ID: 416925112
2	NAPS: People_007
3	NAPS: Animals_036
4	Royalty-free stock photo ID: 110184284
5	Royalty-free stock photo ID: 549508300
6	Royalty-free stock photo ID: 359789747
7	Royalty-free stock photo ID: 75703819
8	Royalty-free stock photo ID: 110184296
Negative	
1	Mum chose to deliver terminally-ill baby just to say good-bye.
2	Serious two-car crash kills young father Warwick Hirvonen.
3	Malnourished horse and its baby forced to eat dirt and trash to survive
4	US soldiers critically injured after chemical attack
5	Floodwaters claim the lives of 20 people in Indonesia
6	11 year old’s mother tells story of his suicide from her perspective
7	3 firefighters dead after vicious factory blaze
8	Several chemical plant workers dead after dangerous leak
Neutral	
1	We take a look at how birthing procedures have changed over the past 100 years
2	Car accident simulation prepares students for real world.
3	The new wild horses: rogue horses spotted in inner city suburbs
4	Modern vs. historic warfare through pictures
5	Weather patterns across the world: we take a look
6	Tears of joy and tears of sadness look different under the microscope
7	Meet your local firefighters on open day
8	7 of the world’s most dangerous jobs

Appendix C—Positive affect negative affect schedule (PANAS)

Directions

This scale consists of a number of words that describe different feelings and emotions. Read each item and then circle the appropriate answer next to that word. Indicate to what extent you currently feel this way.

Use the following scale to record your answers.

(1) = Very slightly or not at all	(2) = A little	(3) = Moderately	(4) = Quite a bit	(5) = Extremely

	Very slightly or not at all	**A little**	**Moderately**	**Quite a bit**	**Extremely**
1. Interested	1	2	3	4	5
2. Distressed	1	2	3	4	5
3. Excited	1	2	3	4	5
4. Upset	1	2	3	4	5
5. Strong	1	2	3	4	5
6. Guilty	1	2	3	4	5
7. Scared	1	2	3	4	5
8. Hostile	1	2	3	4	5
9. Enthusiastic	1	2	3	4	5
10. Proud	1	2	3	4	5
11. Irritable	1	2	3	4	5
12. Alert	1	2	3	4	5
13. Ashamed	1	2	3	4	5
14. Inspired	1	2	3	4	5
15. Nervous	1	2	3	4	5
16. Determined	1	2	3	4	5
17. Attentive	1	2	3	4	5
18. Jittery	1	2	3	4	5
19. Active	1	2	3	4	5
20. Afraid	1	2	3	4	5

Appendix D—Six-item short form of the State Trait Anxiety Inventory (STAI-6)

A number of statements which people have used to describe themselves are given bellow. Read each statement and then select the most appropriate rating to the right of the statement to indicate how you feel right now, at this moment. *There are no right or wrong answers. Do not spend too much time on any one statement but give the answer which seems to describe your present feelings best.*

	Not at all	Somewhat	Moderately	Very much
1. I feel calm	1	2	3	4
2. I am tense	1	2	3	4
3. I feel upset	1	2	3	4
4. I am relaxed	1	2	3	4
5. I feel content	1	2	3	4
6. I am worried	1	2	3	4

Please make sure that you have answered *all* the questions.

Appendix E—Posttest-reactions questionnaire

<u>Questions About This Study</u>
Your answers will be kept absolutely anonymous and confidential.
You are free to skip any question that you feel uncomfortable answering for any reason.
If any question is very difficult or distressing for you, please write a big 'X' to the right of it.

Please indicate how much you <u>agree or disagree</u> with the following statements <u>about the study you just completed</u>, by circling the appropriate number on each scale.

	I strongly disagree			I feel neutral			I strongly agree
1. This study was boring	1	2	3	4	5	6	7
2. This study was mentally exhausting	1	2	3	4	5	6	7
4. This study was offensive to my values	1	2	3	4	5	6	7
5. This study made me feel stupid	1	2	3	4	5	6	7
6. This study gave me some insights into myself	1	2	3	4	5	6	7
9. This study kept my attention	1	2	3	4	5	6	7
11. This study was intellectually challenging	1	2	3	4	5	6	7
17. This study was interesting	1	2	3	4	5	6	7
31. This study gave me a headache	1	2	3	4	5	6	7
37. The subject matter of this study was similar to topics I have sometimes discussed with family or friends	1	2	3	4	5	6	7
38. The subject matter of this study was similar to topics sometimes covered in my class lectures	1	2	3	4	5	6	7
41. The consent form clearly described what this study would be like	1	2	3	4	5	6	7
42. I wish I had never signed up for this study	1	2	3	4	5	6	7
43. Some people could learn valuable things about themselves by participating in this study	1	2	3	4	5	6	7
46. I would like to participate in more studies like this	1	2	3	4	5	6	7

Scoring for post-tests reactions questionnaire

Perceived benefits:
1R
4R
6
9
17
37
41
42R
43
46

Mental costs:
2
5
11
31
38

R= reverse coded

Appendix F—Autobiographical memory instructions

Time 1

In this study we are interested in your autobiographical memory, that is, memory for events in your life that you can specify as occurring at one particular place and time.

Specifically, we would like you to think about a **NEGATIVE** event that happened to you in the last 2 weeks.

Try to re-experience that event as vividly as possible. To start off with, sit back in the chair, close your eyes and bring the event back into your mind in as much detail as possible.

When you have the event firmly in your mind, please write it down. Use as much of the box provided below as possible. The box will accommodate any length of text and you can drag the bottom right corner of the box to make it larger. Remember that what you write is completely confidential. Please be as complete and accurate as possible.

<u>Here are some questions designed to help you with the task.</u>
What sort of day was it? What was the weather like? What had you been doing beforehand? What happened? Where were you at the time? How did you find out about it? Who was involved? What did they say? What did you say? What did you think? How did you feel?

When you have written as much as you can remember about this event, please click the arrow button at the bottom of the page and answer the following questions about the event.

Time 2

Two weeks ago in Session 1, you were asked to recall and write about a negative autobiographical memory.

For today's session, we would like you to recall and write about this same event again (i.e., the event you wrote about in Session 1).

Try to re-experience that event as vividly as possible. To start off with, sit back in the chair, close your eyes and bring the event back into your mind in as much detail as possible.

When you have the event firmly in your mind, please write it down. Use as much of the box provided below as possible. The box will accommodate any length of text and you can drag the bottom right corner of the box to make it larger. Remember that what you write is completely confidential. Please be as complete and accurate as possible.

<u>Here are some questions designed to help you with the task.</u>
What sort of day was it? What was the weather like? What had you been doing beforehand? What happened? Where were you at the time? How did you find out about it? Who was involved? What did they say? What did you say? What did you think? How did you feel?

When you have written as much as you can remember about this event, please click the arrow button at the bottom of the page and answer the following questions about the event.

Appendix G—Memory phenomenology items

Category	**Variable name**	**Question**
Reliving (mean of items)	Reliving_1	While remembering the event, I feel as though I am reliving it (1 = not at all, 7 = as if it were happening now).
	Reliving_2	While remembering the event, I feel that I travel back to the time it happened (1 = not at all, 7 = completely).
	Emotioninten_2	While remembering the event, I can feel now the *emotion* I felt then (1 = not at all, 7 = as if it was happening right now).
	Emotioninten_3	Compared to how I felt at the time of the event, the *emotion* I feel about it now is (1 = completely different, 7 = identically the same).
Belief (mean of items)	Realimaginary	I believe the event in my memory really occurred in the way I remember it and that I have not imagined or fabricated anything that did not occur (1 = 100% imaginary, 7 = 100% real).
	R/K_Judgment	Sometimes people know something happened to them without being able to actually remember it. As I think about the event I can actually *remember* it rather than just knowing that it happened (1 = not at all, 7 = as much as any memory).
	FvO	While remembering the event, I feel that I see it out of my own eyes rather than that of an outside observer (1 = not at all, 7 = completely).

	LN_6	This memory is based on details *specific* to my life, not on general knowledge that I would expect most people to have (1 = not at all, 7 = completely).
Sensory details		Does your memory for this event contain sensory details?
	SensoryDetails_1	Visual (Yes/No)
	SensoryDetails_2	Auditory (Yes/No)
	SensoryDetails_3	Olfactory (smell; Yes/No)
	SensoryDetails_4	Tactile (touch; Yes/No)
	SensoryDetails_5	Gustatory (taste; Yes/No)
Vividness (mean of items)	Vivid_1	How vivid and clear is your memory for this event? (1 = not at all vivid and clear, 7 = completely vivid and clear).
	Sensory_1	While remembering the event, I can see it in my mind (1 = not at all, 7 = as if it were happening right now).
	Sensory_2	While remembering the event, I can hear it in my mind (1 = not at all, 7 = as if it were happening right now).
	Sensory_3	While remembering the event, I can smell it (1 = not at all, 7 = as if it were happening right now).
	Sensory_4	As I remember the event, I or other people are *talking* (1 = not at all, 7 = as if it were happening right now).
Content (mean of items)	Setting_1	While remembering the event, I know the setting where the event occurred (1 = not at all, 7 = as if it were happening right now).
	Setting_2	While remembering the event, I know the location of actions within the event

		(1 = not at all, 7 = as if it were happening right now).
Time (mean of items)	Time_1	My memory for the day when the event took place is clear (1 = not at all, 7 = extremely).
	Time_2	My memory for the hour when the event took place is clear (1 = not at all, 7 = extremely).
Emotional intensity (mean of items)	Emotioninten_1	While remembering the event, the emotions that I feel are extremely intense (1 = not at all, 7 = extremely).
	PR	While remembering the event, I had a physical reaction (I laughed, felt tense, sweaty, felt cramps or butterflies in my stomach, my heart pound or race, etc.;1 = not at all, 7 = extremely strong).
	ValR_Pos (reversed)	While remembering the event, the emotions are *extremely* positive (1 = not at all, 7 = entirely).
	ValE_pos (reversed)	My feelings at the time were positive (1 = not at all, 7 = entirely).
	ValR_neg	While remembering the event, the emotions are *extremely* negative (1 = not at all, 7 = entirely).
	ValE_Neg	My feelings at the time were negative (1 = not at all, 7 = entirely).
Rehearsal (mean of items)	Intrusive	This memory has previously come to me “out of the blue”, without my trying to think about it (1 = not at all, 7 = very often).
	Rehearsal 1:	Since it happened, I have thought about this event (1 = not at all, 7 = very often).

	Rehearsal 2	Since it happened, I have talked about this event (1 = not at all, 7 = very often).
Accessibility (mean of items)	Accessibility_1 (reversed)	This memory just sprang to my mind when I read the instructions, rather than having to search my “memory bank” for this event (1 = agree, 7 = disagree).
	Accessibility_2 (reversed)	Rather than being difficult to think of, this memory was easy for me to recall (1 = agree, 7 = disagree).
	Accessibility_3	It was difficult to think of this memory (1 = agree, 7 = disagree).
	Accessibility_4	I had to think for a while before I could recall this event (1 = agree, 7 = disagree).
	Accessibility_5	I really had to search my memory bank for this experience (1 = agree, 7 = disagree).
Coherence (mean of items)	LN_1	While remembering the event, it comes to me in words or in pictures as a coherent story or episode and not as an isolated fact, observation, or scene (1 = not at all, 7 = completely).
	LN_2 (reverse scored)	My memory comes in pieces with missing bits (1 = not at all, 7 = completely).
	LN_3	The order of the actions within the event in the memory is clear (1 = not at all, 7 = completely).
	LN_4	The order of the events before and after the in the memory is clear (1 = not at all, 7 = completely).
	LN_5	While remembering the event, it comes to me in

		words (1 = not at all, 7 = completely).

Appendix H—The Centrality of Events Scale (CES-20; 7-item versions marked with asterisk)

Please think back upon the negative event you recalled and answer the following questions in an honest and sincere way, by circling a number from 1 to 5.

1. This event has become a reference point for the way I understand new experiences.

totally disagree 1 2 3 4 5 totally agree

2. I automatically see connections and similarities between this event and experiences in my present life.

totally disagree 1 2 3 4 5 totally agree

* 3. I feel that this event has become part of my identity.

totally disagree 1 2 3 4 5 totally agree

4. This event can be seen as a symbol or mark of important themes in my life.

totally disagree 1 2 3 4 5 totally agree

5. This event is making my life different from the life of most other people.

totally disagree 1 2 3 4 5 totally agree

* 6. This event has become a reference point for the way I understand myself and the world.

totally disagree 1 2 3 4 5 totally agree

7. I believe that people who haven't experienced this type of event think differently than I do.

totally disagree 1 2 3 4 5 totally agree

8. This event tells a lot about who I am.

totally disagree 1 2 3 4 5 totally agree

9. I often see connections and similarities between this event and my current relationships with other people.

totally disagree 1 2 3 4 5 totally agree

*10. I feel that this event has become a central part of my life story.

totally disagree 1 2 3 4 5 totally agree

11. I believe that people who haven't experienced this type of event, have a different way of looking upon themselves than I have.

totally disagree 1 2 3 4 5 totally agree

*12. This event has colored the way I think and feel about other experiences.

totally disagree 1 2 3 4 5 totally agree

13. This event has become a reference point for the way I look upon my future.

totally disagree 1 2 3 4 5 totally agree

14. If I were to weave a carpet of my life, this event would be in the middle with threads going out to many other experiences.

totally disagree 1 2 3 4 5 totally agree

15. My life story can be divided into two main chapters: one is before and one is after this event happened.

totally disagree 1 2 3 4 5 totally agree

*16. This event permanently changed my life.

totally disagree 1 2 3 4 5 totally agree

*17. I often think about the effects this event will have on my future.

totally disagree 1 2 3 4 5 totally agree

*18. This event was a turning point in my life.

totally disagree 1 2 3 4 5 totally agree

19. If this event had not happened to me, I would be a different person today.

totally disagree 1 2 3 4 5 totally agree

20. When I reflect upon my future, I often think back to this event.

totally disagree 1 2 3 4 5 totally agree

Appendix I—Revised Impact of Events Scale (IES-R)

Below is a list of difficulties people sometimes have after stressful life events. Please read each item, and then indicate how distressing each difficulty has been for you **DURING THE PAST SEVEN DAYS** with respect to (your problem), how much were you distressed or bothered by these difficulties?

0 = Not at all	1 = A little bit	2 = Moderately	3 = Quite a bit	4 =Extremely

1. Any reminder brought back feelings about it
2. I had trouble staying asleep
3. Other things kept making me think about it
4. I felt irritable and angry
5. I avoided letting myself get upset when I thought about it or was reminded of it
6. I thought about it when I didn't mean to
7. I felt as if it hadn't happened or wasn't real
8. I stayed away from reminders about it
9. Pictures about it popped into my mind
10. I was jumpy and easily startled
11. I tried not to think about it
12. I was a ware I still had feelings about it, but I didn’t deal with them
13. My feelings about it were kind of numb
14. I found myself acting or feeling like I was back at that time
15. I had trouble falling asleep
16. I had waves of strong feelings about it
17. I tried to remove it from my memory
18. I had trouble concentrating
19. Reminders of it caused me to have physical reactions, such as sweating, trouble breathing, nausea, or a pounding heart
20. I had dreams about it
21. I felt watchful and on guard
22. I tried not to talk about it

Appendix J—Ways of Coping (Revised; WCS-R)

Please read each item below and indicate, by using the following rating scale, to what extent you used it in the situation you have just described.

0 = Not used, 1 = Used Somewhat, 2 = Used Quite A Bit, 3 = Used A Great Deal

_____ 1. Just concentrated on what I had to do next – the next step.

_____ 2. I tried to analyze the problem in order to understand it better.

_____ 3. Turned to work or substitute activity to take my mind off things.

_____ 4. I felt that time would make a difference – the only thing to do was to wait.

_____ 5. Bargained or compromised to get something positive from the situation.

_____ 6. I did something which I didn't think would work, but at least I was doing something.

_____ 7. Tried to get the person responsible to change his or her mind.

_____ 8. Talked to someone to find out more about the situation.

_____ 9. Criticized or lectured myself.

_____ 10. Tried not to burn my bridges, but leave things open somewhat.

_____ 11. Hoped a miracle would happen.

_____ 12. Went along with fate; sometimes I just have bad luck.

_____ 13. Went on as if nothing had happened.

_____ 14. I tried to keep my feelings to myself.

_____ 15. Looked for the silver lining, so to speak; tried to look on the bright side of things.

_____ 16. Slept more than usual.

_____ 17. I expressed anger to the person(s) who caused the problem.

_____ 18. Accepted sympathy and understanding from someone.

_____ 19. I told myself things that helped me to feel better.

_____ 20. I was inspired to do something creative.

_____ 21. Tried to forget the whole thing.

_____ 22. I got professional help.

_____ 23. Changed or grew as a person in a good way.

_____ 24. I waited to see what would happen before doing anything.

_____ 25. I apologized or did something to make up.

_____ 26. I made a plan of action and followed it.

_____ 27. I accepted the next best thing to what I wanted.

_____ 28. I let my feelings out somehow.

_____ 29. Realized I brought the problem on myself.

_____ 30. I came out of the experience better than when I went in.

_____ 31. Talked to someone who could do something concrete about the problem.

_____ 32. Got away from it for a while; tried to rest or take a vacation.

_____ 33. Tried to make myself feel better by eating, drinking, smoking, using drugs or medication, etc.

_____ 34. Took a big chance or did something very risky.

_____ 35. I tried not to act too hastily or follow my first hunch.

_____ 36. Found new faith.

_____ 37. Maintained my pride and kept a stiff upper lip.

_____ 38. Rediscovered what is important in life

_____ 39. Changed something so things would turn out all right.

_____ 40. Avoided being with people in general.

_____ 41. Didn’t let it get to me; refused to think too much about it.

_____ 42. I asked a relative or friend I respected for advice.

_____ 43. Kept others from knowing how bad things were.

_____ 44. Made light of the situation; refused to get too serious about it.

_____ 45. Talked to someone about how I was feeling.

_____ 46. Stood my ground and fought for what I wanted.

_____ 47. Took it out on other people.

_____ 48. Drew on my past experiences; I was in a similar situation before.

_____ 49. I knew what had to be done, so I doubled my efforts to make things work.

_____ 50. Refused to believe that it had happened.

_____ 51. I made a promise to myself that things would be different next time.

_____ 52. Came up with a couple of different solutions to the problem.

_____ 53. Accepted it, since nothing could be done.

_____ 54. I tried to keep my feelings from interfering with other things too much.

_____ 55. Wished that I could change what had happened or how I felt.

_____ 56. I changed something about myself.

_____ 57. I daydreamed or imagined a better time or place than the one I was in.

_____ 58. Wished that the situation would go away or somehow be over with.

_____ 59. Had fantasies or wishes about how things might turn out

______60. I prayed.

_____ 61. I prepared myself for the worst.

_____ 62. I went over in my mind what I would say or do.

_____ 63. I thought about how a person I admire would handle this situation and used that as a model.

_____ 64. I tried to see things from the other person’s point of view.

_____ 65. I reminded myself how much worse things could be.

_____ 66. I jogged or exercised.

Appendix K—Trauma History Screen (THS)

The events below may or may not have happened to you. Circle "YES" if that kind of thing has happened to you or circle "NO" if that kind of thing has not happened to you. If you circle "YES" for any events: put a number in the blank next to it to show how many times something like that happened. Event Circle "YES" if that kind of thing has happened to you Circle "NO" if that kind of thing has not happened to you:

Number of times something like this has happened
A. A really bad car, boat, train, or airplane accident YES NO _____ times
B. A really bad accident at work or home YES NO _____ times
C. A hurricane, flood, earthquake, tornado, or fire YES NO _____ times
D. Hit or kicked hard enough to injure - as a child YES NO _____ times
E. Hit or kicked hard enough to injure - as an adult YES NO _____ times
F. Forced or made to have sexual contact - as a child YES NO _____ times
G. Forced or made to have sexual contact - as an adult YES NO _____ times
H. Attack with a gun, knife, or weapon YES NO _____ times
I. During military service - seeing something horrible or being badly scared YES NO _____ times
J. Sudden death of close family or friend YES NO _____ times
K. Seeing someone die suddenly or get badly hurt or killed YES NO _____ times
L. Some other sudden event that made you feel very scared, helpless, or horrified YES NO _____ times
M. Sudden move or loss of home and possessions YES NO _____ times
N. Suddenly abandoned by spouse, partner, parent, or family YES NO _____ times

Briefly describe (in one or two sentences) the most stressful experience of your life in the box below. We are going to ask you a number of questions about this event.

Your age when this happened: ______
When this happened, did anyone get hurt or killed? NO YES
When this happened, were you afraid that you or someone else might get hurt or killed? NO YES
When this happened, did you feel very afraid, helpless, or horrified? NO YES
When this happened, did you feel unreal, spaced out, disoriented, or strange? NO YES
After this happened, how long were you bothered by it? not at all / 1 week / 2-3 weeks / a month or more
How much did it bother you emotionally? not at all / a little / somewhat / much / very much

Appendix L—Posttraumatic Stress Disorder Checklist 5 (PCL-5)

Below is a list of problems that people sometimes have in response to a very stressful experience. Keeping your worst event in mind, please read each problem carefully and then circle one of the numbers to the right to indicate how much you have been bothered by that problem in the past month.

No.	Response:	Not at all	A little bit	Moderately	Quite a bit	Extremely
1.	Repeated, disturbing, and unwanted memories of the stressful experience?	0	1	2	3	4
2.	Repeated, disturbing dreams of the stressful experience?					
3.	Suddenly feeling or acting as if the stressful experience were actually happening again (*as if you were actually back there reliving it*)?					
4.	Feeling very upset when something reminded you of the stressful experience?					
5.	Having strong physical reactions when something reminded you of the stressful experience (*for example, heart pounding, trouble breathing, sweating*)?					
6.	Avoiding memories, thoughts, or feelings related to the stressful experience?					
7.	Avoiding external reminders of the stressful experience (*for example, people, places, conversations, activities, objects, or situations*)?					
8.	Trouble remembering important parts of the stressful experience?					
9.	Having strong negative beliefs about yourself, other people, or the world (*for example, having thoughts such as: I am bad, there is something seriously wrong*					

with me, no one can be trusted, the world is completely dangerous)?

10. Blaming yourself or someone else for the stressful experience or what happened after it?
11. Having strong negative feelings such as fear, horror, anger, guilt, or shame?
12. Loss of interest in activities that you used to enjoy?
13. Feeling distant or cut off from other people?
14. Trouble experiencing positive feelings (*for example, being unable to feel happiness or have loving feelings for people close to you*)?
15. Irritable behaviour, angry outbursts, or acting aggressively?
16. Taking too many risks or doing things that could cause you harm?
17. Being “super alert” or watchful or on guard?
18. Feeling jumpy or easily startled?
19. Having difficulty concentrating?
20. Trouble falling or staying asleep?

Appendix M—Modified Autobiographical Questionnaire

Pre-experiencing	While imagining the event, I feel as though I am experiencing it: 1 = not at all, 7 = completely
Mental time travel	While imagining the event, I feel that I travel forward to the time when it would happen: 1 = not at all, 7 = completely
Visual details	My representation for this event involves visual details: 1 = none, 7 = a lot
Other sensory details	Average of sounds and smells/tastes
Sounds	My representation for this event involves sounds: 1 = none, 7 = a lot
Smells/tastes	My representation for this event involves smells/tastes: 1 = none, 7 = a lot
Spatial context	Average of location, spatial arrangement of objects, and spatial arrangement of people
Location	My representation for the location where the event takes place is: 1 = not at all clear, 7 = very clear.
Spatial arrangement of objects	Relative spatial arrangement of objects in my representation for the events is: 1 = not at all clear, 7 = very clear.
Spatial arrangement of people	Relative spatial arrangement of people in my representation for the event is: 1 = not at all clear, 7 = very clear.
Temporal information	My representation for the time of day when the event takes place is: 1 = not at all clear.
Feeling emotions	While imagining the event, I feel the emotions I would feel if the event occurred: 1 = not at all, 7 = completely
Intensity	If this event happened, my emotions would be: 1 = not intense, 7 = very intense
Valence	If this event happened, my emotions would be: -3 = very intense, 0 = neutral, +3= positive.
Personal importance	This event is very important to me (it involves an important theme or episode in my life): 1 = not at all important, 7 = very important.
In-words	While imagining the event, it comes to me in words: 1 = not at all, 7 = a lot.
Coherent story	While imagining the event, it comes to me as a coherent story and not as an isolated scene: 1 = not at all, 7 = completely.
Visual perspective	As I imagine the event, I see it out of my own eyes rather than those of an outside observer: -3 = entirely looking though my

	eyes, +3 = entirely observing myself as an outside observer.
Vividness	How vivid is this imagined event: 1 = vague, 7 = extremely vivid

Appendix N—Coping Response Inventory (CRI)

Modified version—Questions asked in relation to imagined event:

We are interested in how people respond when they confront difficult or stressful events in their lives. There are lots of ways to try to deal with stress. This questionnaire asks you to indicate what you think you would do and how you would feel, **in the scenario you just imagined and wrote about**.

Then respond to each of the following items by selecting one number for each, using the response choices listed. Please try to respond to each item separately in your mind from each other item. Choose your answers thoughtfully, and make your answers as true FOR YOU as you can. Please answer every item. There are no "right" or "wrong" answers, so choose the most accurate answer for YOU--not what you think "most people" would say or do. Indicate what YOU think you would do if YOU experienced the event you just imagined and wrote about.

1 = I wouldn’t do this at all
2 = I would do this a little bit
3 = I would do this a medium amount
4 = I would do this a lot

1. I try to think of this experience as something to grow as a person from
2. I turn to work or other substitute activities to take my mind off things.
3. I get upset and let my emotions out.
4. I try to get advice from someone about what to do.
5. I concentrate my efforts on doing something about it.
6. I say to myself "this isn't real."
7. I put my trust in God.
8. I laugh about the situation.
9. I admit to myself that I can't deal with it, and quit trying.
10. I restrain myself from doing anything too quickly.

11. I discuss my feelings with someone.
12. I use alcohol or drugs to make myself feel better.
13. I get used to the idea that it is happening.
14. I talk to someone to find out more about the situation.
15. I keep myself from getting distracted by other thoughts or activities.
16. I daydream about things other than this.
17. I get upset, and am really aware of it.
18. I seek God's help.
19. I make a plan of action.
20. I make jokes about it.

21. I accept that this is happening and that it can't be changed.
22. I hold off doing anything about it until the situation permits.
23. I try to get emotional support from friends or relatives.
24. I just give up trying to reach my goal.
25. I take additional action to try to get rid of the problem.
26. I try to lose myself for a while by drinking alcohol or taking drugs.

27. I refuse to believe that it is happening.
28. I let my feelings out.
29. I try to see it in a different light, to make it seem more positive.
30. I talk to someone who could do something concrete about the problem.

31. I sleep more than usual.
32. I try to come up with a strategy about what to do.
33. I focus on dealing with this problem, and if necessary let other things slide a little.
34. I get sympathy and understanding from someone.
35. I drink alcohol or take drugs, in order to think about it less.
36. I kid around about it.
37. I give up the attempt to get what I want.
38. I look for something good in what is happening.
39. I think about how I might best handle the problem.
40. I pretend that it isn’t really happening.

41. I make sure not to make matters worse by acting too soon.
42. I try hard to prevent other things from interfering with my efforts at dealing with this.
43. I go to movies or watch TV, to think about it less.
44. I accept the reality of the fact that it is happening.
45. I ask people who have had similar experiences what they did.
46. I feel a lot of emotional distress and I find myself expressing those feelings a lot.
47. I take direct action to get around the problem.
48. I try to find comfort in my religion.
49. I force myself to wait for the right time to do something.
50. I make fun of the situation.

51. I reduce the amount of effort I'm putting into solving the problem.
52. I talk to someone about how I feel.
53. I use alcohol or drugs to help me get through it.
54. I learn to accept with it.
55. I put aside other activities in order to concentrate on this.
56. I think hard about what steps to take.
57. I act as though it isn’t even happening.
58. I do what has to be done, one step at a time.
59. I learn something from the experience.
60. I pray more than usual.

Appendix O—Emotion Regulation Questionnaire (ERQ)

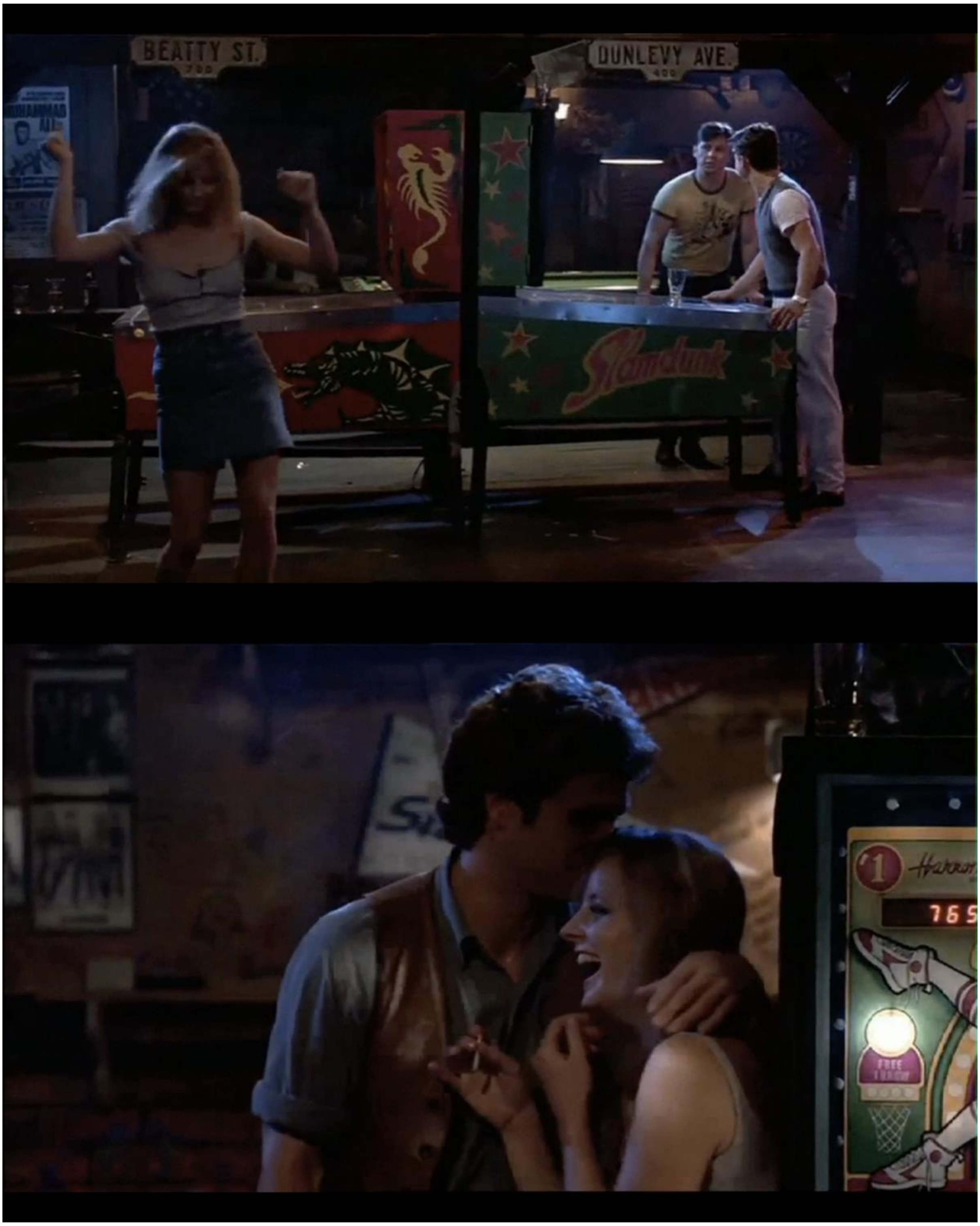
Modified version—Questions asked in relation to imagined event:

We would like to ask you some questions related to how you might control (that is, regulate and manage) your emotions **in relation to the scenario you just imagined and wrote about.** The questions below involve two distinct aspects of your emotional life. One is your emotional experience, or what you feel like inside. The other is your emotional expression, or how you show your emotions in the way you talk, gesture, or behave. Although some of the following questions may seem similar to one another, they differ in important ways. For each item, please answer using the following scale:

1	2	3	4	5	6	7
Strongly disagree			Neutral			Strongly

1. ____ If I wanted to feel a more positive emotion (such as joy or amusement), I would change what I was thinking about.
2. ____ I would keep my emotions to myself.
3. ____ If I wanted to feel a less negative emotion (such as sadness or anger), I would change what I was thinking about.
4. ____If I was feeling positive emotions, I would be careful not to express them.
5. ____ I would make myself think about it in a way that helps me stay calm.
6. ____ I would control my emotions by not expressing them.
7. ____If I wanted to feel more positive emotion, I would change the way I was thinking about the situation.
8. ____ I would control my emotions by changing the way I was thinking about the situation I was in.
9. ____If I was feeling negative emotions, I would make sure not to express them.
10. ____If I wanted to feel a less negative emotion, I would change the way I was thinking about the situation.

Appendix P—Example image still stimuli used in Study 5



Appendix Q—Anticipated traumatic stress symptoms

Below is a list of problems and complaints that people sometimes have following stressful experiences (such as the film and image viewing task you recently completed). Please read each one carefully, then select one of the numbers to indicate how much you think you might be bothered by that problem in the next 24 hours.

Not at all	**A little bit**	**Moderately**	**Quite a bit**	**Extremely**
0	1	2	3	4

1. Repeated, disturbing and unwanted images related to the film or images?
2. Repeated, disturbing dreams related to the film or images?
3. Suddenly acting or feeling as if you were exposed to the film or images again (as if you were re-living this experience)?
4. Feeling very upset when something reminds you of the film or images?
5. Having strong physical reactions (e.g., heart pounding, trouble breathing, sweating) when something reminds you of the film or images?
6. Avoiding imaginings, thoughts or feelings related to the film or images?
7. Avoiding external remindings of the film or images (for example people, places, conversations, activities, objects or situations)?
8. Trouble imagining important parts of the film or images?
9. Having strong negative beliefs about yourself, other people, or the world (for example, having thoughts such as: I am bad, there is something seriously wrong with me, no one can be trusted, the world is completely dangerous)?
10. Having strong negative feelings such as fear, horror, anger, guilt, or shame?
11. Loss of interest in activities that you used to enjoy?
12. Feeling distant or cut off from other people?
13. Trouble experiencing positive feelings (for example, being unable to feel happiness or have loving feelings for people close to you)?
14. Irritable behavior, angry outbursts, or acting aggressively?
15. Taking too many risks or doing things that could cause you harm?
16. Being "superalert" or watchful or on guard?
17. Feeling jumpy or easily startled?
18. Having difficulty concentrating?
19. Trouble falling or staying asleep?

Appendix R— Spielberger Trait Anxiety Inventory (STAI-T)

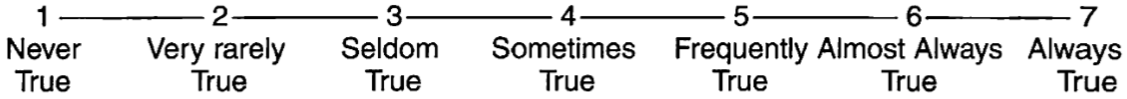
A number of statements which people have used to describe themselves are given below. Read each statement and then select the appropriate number to the right of the statement to indicate how you generally feel.

1 = almost never, 2 = sometimes, 3 = often, 4 = almost always.

1. I feel pleasant
2. I feel nervous and restless
3. I feel satisfied with myself
4. I wish I could be as happy as others seem to be
5. I feel like a failure
6. I feel rested
7. I am 'calm, cool and collected'
8. I feel that difficulties are piling up so that I cannot overcome them
9. I worry too much over something that really doesn't matter
10. I am happy
11. I have disturbing thoughts
12. I lack self-confidence
13. I feel secure
14. I make decisions easily
15. I feel inadequate
16. I am content
17. Some unimportant thought runs through my mind and bothers me
18. I take disappointments so keenly that I can't put them out of my mind
19. I am a steady person
20. I get in a state of tension or turmoil as I think over my recent concerns and interests

Appendix S—The Acceptance and Actions Questionnaire (AAQ)

Below you will find a list of statements. Please rate the truth of each statement as it applies to you. Use the following scale to make your choice.



1	2	3	4	5	6	7
Never True	Very rarely True	Seldom True	Sometimes True	Frequently True	Almost Always True	Always True

1. I am able to take action on a problem even if I am uncertain what is the right thing to do.
2. I often catch myself daydreaming about things I've done and what I would do differently next time.
3. When I feel depressed or anxious, I am unable to take care of my responsibilities.
4. I rarely worry about getting my anxieties, worries, and feelings under control.
5. I'm not afraid of my feelings.
6. When I evaluate something negatively, I usually recognize that this is just a reaction, not an objective fact.
7. When I compare myself to other people, it seems that most of them are handling their lives better than I do.
8. Anxiety is bad.
9. If I could magically remove all the painful experiences I've had in my life, I would do so.

Appendix T— Coping strategies following a traumatic event

1) We are interested in how people respond when they confront difficult or stressful events in their lives. There are lots of ways to try to deal with stress. This questionnaire asks you to indicate what you generally do and feel, when you experience stressful events. Obviously, different events bring out somewhat different responses, but think about what you usually do when you are under a lot of stress. Then respond to each of the following items. Please try to respond to each item separately in your mind from each other item. Choose your answers thoughtfully, and make your answers as true FOR YOU as you can. There are no "right" or "wrong" answers, so choose the most accurate answer for YOU--not what you think "most people" would say or do. Indicate what YOU usually do when YOU experience a stressful event/
2) How beneficial do you think each of the following strategies would be for your mental health following a traumatic/stressful event:

Avoidance coping (behaviour and emotion):

1. Avoiding reminders of the event in order to reduce distress (e.g., avoid people, places, media etc. related to the event).
2. Avoiding thoughts and feelings associated with the event in order to reduce distress (e.g., try not to think about the event, try to forget the event, attempt to “push away”, disengage, or avoid expressing thoughts and feelings).

Approach coping (behaviour and emotion):

1. Engaging with reminders of the event (e.g., people, places media etc.) in order to reduce distress (e.g., to learn more information about yourself and the event).
2. Engaging with thoughts and feelings associated with the event in order to reduce distress (e.g., trying to find personal meaning in the event, thinking of different ways to deal with the outcomes of the event, try to see the good side of the situation etc.).

Appendix U—Depression Anxiety Stress Scales (DASS-21)

Please read each statement and circle a number 0, 1, 2 or 3 which indicates how much the statement applied to you over the past week. There are no right or wrong answers. Do not spend too much time on any statement.

The rating scale is as follows:
0 Did not apply to me at all
1 Applied to me to some degree, or some of the time
2 Applied to me to a considerable degree or a good part of time
3 Applied to me very much or most of the time

1. I found it hard to wind down 0 1 2 3 2
2. I was aware of dryness of my mouth 0 1 2 3 3
3. I couldn't seem to experience any positive feeling at all 0 1 2 3 4
4. I experienced breathing difficulty (e.g. excessively rapid breathing, breathlessness in the absence of physical exertion) 0 1 2 3 5
5. I found it difficult to work up the initiative to do things 0 1 2 3 6
6. I tended to over-react to situations 0 1 2 3 7
7. I experienced trembling (e.g. in the hands) 0 1 2 3 8
8. I felt that I was using a lot of nervous energy 0 1 2 3 9
9. I was worried about situations in which I might panic and make a fool of myself 0 1 2 3 10
10. I felt that I had nothing to look forward to 0 1 2 3 11
11. I found myself getting agitated 0 1 2 3 12
12. I found it difficult to relax 0 1 2 3 13
13. I felt down-hearted and blue 0 1 2 3 14
14. I was intolerant of anything that kept me from getting on with what I was doing 0 1 2 3 15
15. I felt I was close to panic 0 1 2 3 16
16. I was unable to become enthusiastic about anything 0 1 2 3 17
17. I felt I wasn't worth much as a person 0 1 2 3 18
18. I felt that I was rather touchy 0 1 2 3 19
19. I was aware of the action of my heart in the absence of physical exertion (e.g. sense of heart rate increase, heart missing a beat) 0 1 2 3 20
20. I felt scared without any good reason 0 1 2 3 21
21. I felt that life was meaningless 0 1 2 3

Appendix V—14-item Scales of General Well-Being (14-SGWB)

Instructions Below you'll find fourteen statements about your experiences. Please indicate how true each statement is regarding the EXPERIENCES IN YOUR LIFE OVERALL. There are no right or wrong answers. Please, choose the answer that best reflects your experience rather than what you think your experience should be. (Not at all true, A bit true, Somewhat true, Mostly true, Very true).

1. I feel happy
2. I feel energetic
3. I feel calm
4. I'm optimistic
5. In my activities, I feel absorbed by what I'm doing
6. I'm in touch with how I really feel inside
7. I accept most aspects of myself
8. I feel great about myself
9. I am highly effective at what I do
10. I feel I am improving
11. I have a purpose
12. What I do in my life is worthwhile
13. What I do is consistent with what I believe I should do
14. I feel close and connected to the people around me

Appendix W—The Self-Triggering Questionnaire (STQ)

1. Some people who have experienced difficult events seek experiences (video, literature, places, etc.) that remind them of that event. This behavior is known by some as “self-triggering.” Have you ever self-triggered with reminders of the "worst event" you chose? This does not include “exposures” assigned by a therapist.

Methods of Self-Triggering

In your own words, please describe how you typically self-trigger in reference to your “worst event,” and what it is like: _________________________________

Below is a list of ways that some people self-trigger. Keeping your “worst event” in mind, please indicate how often you have used each method since your worst event occurred. (0 = *I have not done this*, 1 = *I have done this once*, 2 = *I have done this occasionally,* 3 = *I have done this often*, 4 = *I have done this quite often*).

1. Watching movies or videos that remind me of my worst event.
2. Looking at pictures that remind me of my worst event.
3. Reading things that remind me of my worst event.
4. Going to web pages or online forums that remind me of my worst event.
5. Going to places that remind me of my worst event.
6. Being around people who remind me of my worst event.
7. Collecting objects that remind me of my worst event.
8. Other (please describe):________________

****Participant will be presented with the methods that he/she selected****
Out of all the ways of self-triggering that you endorsed, which one have you used the most?

Frequency of Self-Triggering

Did you self-trigger before your worst event?

a. Yes
b. No

****If “Yes” is selected**** How long ago did you start self-triggering?
__years, __months, __days ago

How long since your worst event occurred did you start self-triggering in reference to your worst event?
__years, __months, __days after my worst event

Since you started self-triggering in reference to your worst event, how often have you self-triggered? (1 = *about once every two or more years*, 2 = *about once a year*, 3 = *once every few months*, 4 = *2-3 times a month*, 5 = *once a week*, 6 = *2-6 times a week*, 7 = *every day*).

In the past month, how often have you self-triggered? (0 = *not at all*, 1 = *once overall*, 2 = *2-3 times overall*, 3 = *about once a week*, 4 = *2-6 times a week*, 5 = *almost every day*).

Overall, how many different times have you self-triggered since your worst event? Please give you best estimate: _________

Motives for Self-Triggering

In your own words, why do you self-trigger? ___________________________

How often do you self-trigger for any of the reasons listed below? (0 = *Never*, 1= *Rarely*, 2 = *Sometimes*, 3 = *Often*, 4 = *Always*)

Sensation-Seeking

1. I self-trigger to generate excitement or exhilaration.
2. I self-trigger to entertain myself by doing something extreme.
3. I self-trigger to feel as if I'm doing something risky or dangerous.

Anti-Dissociation/Numbing

4. I self-trigger to stop feeling numb
5. I self-trigger to feel something (as opposed to nothing), even if it's distress.
6. I self-trigger to make sure I am still alive when I don't feel real

Affect Regulation

7. I self-trigger to calm myself down.
8. I self-trigger to release emotional pressure that has built up inside of me.
9. I self-trigger to reduce anxiety, frustration, anger, or other overwhelming emotions.
10. I self-trigger because if I'm feeling good, I don't want to "crash" all of a sudden.

Shame/guilt/self-punishment

11. I self-trigger because I want to punish myself
12. I self-trigger in order to express anger towards myself for being worthless or stupid
13. I self-trigger because I am feeling unhappy with myself or disgusted with myself

Mastery of Symptoms

14. I self-trigger because I'd rather know when symptoms will come rather than being surprised by them.
15. I self-trigger because I want to gain control over my symptoms.
16. I self-trigger because I want to be better at dealing with reminders of my worst event.

Affect Matching

17. I self-trigger because when I'm feeling "keyed up," or "on edge," I want to have an experience that matches my mood.
18. I self-trigger because when I'm feeling "down," or "blue," I want to have an experience that matches my mood.
19. I self-trigger because when I'm feeling emotional distress, I want to have an experience that matches my mood.

Meaning Making

20. I self-trigger in order to make sense of my worst event.
21. I self-trigger to try to remember parts of my worst event that I forgot
22. I self-trigger to figure out why my worst event happened

23. I self-trigger because without my symptoms, I don’t know who I am.
25. I self-trigger because I don’t want the memory of my worst event to “fade” or become forgotten.
26. I self-trigger because I want to change the memory of my worst event in some way (give it a different ending, or change things that happen in my memory of it).

Other Reasons

27. I self-trigger for other reasons not listed here (please describe)

****Participant is presented with a list of the reasons he/she endorsed**** Out of all of the reasons you endorsed, which one is *most often* the reason you self-trigger?

Appendix X—Mental Health Help Seeking Questionnaire

1. In the last six months, have you taken any medication to treat a personal or emotional problem(s) (e.g., anti-depressants)? (Yes/No)

2. In the last six months, have you seen a mental health professional (e.g., university counsellor, psychologist, psychiatrist) to get help for a personal or emotional problem(s)? (Yes/No)

If no, please go to question 5.

If yes, please complete questions 3 and 4 below.

3. How many visits did you have with the mental health professional?

___________ visits

4. What type of mental health professional(s) have you seen? Please list their titles (e.g., counsellor, psychologist, psychiatrist) below.

____

5. In the last six months, have you seen sought help from anyone (e.g., support, advice, talking it over) or anything (e.g., an app, the internet) other than a mental health professional for your personal or emotional problem? (Yes/No).

If yes, please select the people from whom or ways in which you have sought help.

1. Partner
2. Friend (not related to you)
3. Parent
4. Other relative/family member
5. Phone help line (e.g., Lifeline)
6. Doctor/GP
7. Tutor/Lecturer/Topic Coordinator
8. Someone else not listed above (please describe who this was)
9. Mobile phone application (e.g., headspace, calm)
10. The internet
11. Read a self-help book

Appendix Y—The 5-item World Health Organization Well-Being Index (WHO-5)

Please indicate for each of the five statements which is closest to how you have been feeling over the last two weeks. Notice that higher numbers mean better well-being

0 = At no time, 1 = Some of the time, 2 = Less than half of the time, 3 = More than half of the time, 4 = Most of the time, 5 = All of the time

1. I have felt cheerful and in good spirits
2. I have felt calm and relaxed
3. I have felt active and vigorous
4. I woke up feeling fresh and rested
5. My daily life has been filled with things that interest me