MUNI FSS

JOURNAL OF PSYCHOSOCIAL RESEARCH ON CYBERSPACE
CYBERPSYCHOLOGY

McDonough, C., Lewitzka, S., Bridgland, V. M. E., Moeck, E. K., Nixon, R. D. V., Semmler, C., & Takarangi, M. K. T. (2026). No evidence of protection: Imaginary bulletproof glass fails to reduce psychological harm in content moderation. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 20*(4), Article 8. https://doi.org/10.5817/CP2026-4-8

No Evidence of Protection: Imaginary Bulletproof Glass Fails to Reduce Psychological Harm in Content Moderation

Chloe McDonough[1], Sarah Lewitzka[1], Victoria M. E. Bridgland[1], Ella K. Moeck[2], Reginald D. V. Nixon[1], Carolyn Semmler[2], & Melanie K. T. Takarangi[1]

[1] College of Education, Psychology, and Social Work, Flinders University, Australia, Flinders Institute for Mental Health and Wellbeing, Adelaide, South Australia, Australia
[2] School of Psychology, Adelaide University, Adelaide, South Australia, Australia

Abstract

Content moderators exposed to graphic content online are at risk of negative mental health outcomes including depression, anxiety and distressing intrusive thoughts—a hallmark symptom of post-traumatic stress disorder (PTSD). Yet, there is little research on strategies to protect people exposed to graphic content. Here, we empirically tested a strategy suggested on a range of online sources for handling traumatic imagery–visualizing bulletproof glass exists between oneself and the screen. Using a content moderator simulation paradigm with Amazon Mechanical Turk participants (N = 202), we experimentally tested whether this intervention reduces outcomes associated with exposure to graphic imagery. We found no empirical support for this intervention influencing intrusive thought frequency, negative intrusion characteristics like intensity, anxiety, positive affect, or negative affect. We conclude that this bulletproof glass strategy should not be recommended to groups exposed to graphic content, like content moderators.

Keywords: content moderation; traumatic imagery; anxiety; intrusions; intervention

Editorial Record

First submission received:
February 9, 2026

Revisions received:
May 28, 2026
June 22, 2026

Accepted for publication:
June 22, 2026

Editor in charge:
Douglas A. Parry

Introduction

"*It is important to change your mental tack and put the bulletproof glass up before you deal with it*" a senior newsroom editor advises as a strategy for coping with graphic imagery (Dubberley et al., 2015, p. 23). As social media use grows, with over five billion global users spending over two hours online daily (Duarte, 2024), the volume of online content also grows. Strikingly, around 402 billion gigabytes of content is created daily (Duarte, 2024; Kemp, 2024). This user-generated content includes gruesome and graphic content—including rape, torture, animal abuse and violent deaths—that content moderators filter through to keep it from the public. Content moderators are at risk of developing negative outcomes, like anxiety and intrusive memories of the content (Spence, Bifulco et al., 2023). Despite this risk, there is limited harm mitigation research for people exposed to graphic content online, including content moderators, journalists, and newsroom editors. Visualizing something protective (i.e., bulletproof glass) between oneself and the screen has been suggested online as a strategy for handling traumatic imagery, but this strategy has not been empirically tested (Rees, 2017). Thus, here, we tested whether imagining bulletproof glass

when viewing graphic content in a content moderator simulation reduces the negative impact of viewing such content.

Exposure to a traumatic event—like actual or threatened death or serious injury—can lead to PTSD, marked by symptoms including persistent intrusive trauma memories, hypervigilance, and negative emotions like fear and anger (Ehlers & Steil, 1995). Such symptoms can also occur for events that do not involve actual or threatened death or serious injury (Robinson & Larson, 2010), and for indirect trauma exposure such as media exposure to collective trauma like bombings and exposure to others' suffering (e.g., through working with trauma victims; Collins & Long, 2003; Holman et al., 2020). According to the DSM-5, such indirect exposure to adverse details of the traumatic events through electronic media (e.g., images of actual or threatened death or serious injury) meets Criterion A for a PTSD diagnosis, when the exposure is work related (American Psychiatric Association, 2022).

Unsurprisingly then, professions involving prolonged exposure to graphic content experience psychological harm (Browne et al., 2012; Perez et al., 2010; Steiger et al., 2021). For example, journalists exposed to trauma—like incidents involving injured/dead children—report PTSD symptoms, with more frequent exposure associated with more symptoms (Browne et al., 2012). Similarly, online child exploitation investigators report anxiety, sadness, anger, and intrusive thoughts about material they see (Burns et al., 2008; Krause, 2009). More recently, the negative impact of content moderation has received attention (e.g., Roberts, 2016, 2019). For example, Newton (2019, 2020) reports cases where moderators have developed PTSD, depression, and chronic anxiety after working in moderation roles for as little as six months, and consequently, Meta paid $52 million in settlement with moderators who developed PTSD on the job (Newton, 2020). In another study, moderators exposed to child sexual abuse material reported experiencing symptoms such as anger, anxiety, and intrusive thoughts about the content (Spence, Bifulco et al., 2023; see also Dosono & Semaan, 2019; Schöpke-Gonzalez et al., 2022).

Despite the impact of viewing graphic content, there are few interventions for moderators, and those that exist, have limited efficacy. For example, as one of their key interventions, Meta allows their content moderators to modify (i.e., greyscale/blur) content (Meta, 2024). But recent independent research suggests this intervention is unsuccessful at mitigating harm in a content moderation context (Lewitzka et al., 2026). Positive emotional stimulation—that is, adding positive stimuli (e.g., baby animal images) to break times during a simulated moderation task—has also been tested as a method to reduce harm for experienced moderators. However, this intervention increased stress and emotional distress, which the authors suggest may be because the positive stimulation images triggered compassion fatigue (i.e., the cumulative negative effect on well-being due to an increased susceptibility to taking on other people's hardships; Cook et al., 2022). Another study tested a workplace resilience program for moderators over three months, finding stable burnout and secondary traumatic stress, and a slight decrease in resilience and compassion satisfaction from before to after the program (although scores were still in the standard range, suggesting this decrease was not clinically significant; Steiger et al., 2022).

Seeking other strategies that could benefit content moderators, we turned to professions—like journalism—where viewing graphic content is similarly common. The DART Centre for Journalism and Trauma provides an in-depth guide for handling traumatic imagery (Rees, 2017), including strategies such as visualizing that bulletproof glass exists between oneself and the screen. Similar references to putting up 'your own bulletproof glass' appear in other online guides for managing traumatic content (e.g., Centre for Information Resilience, 2023; Storm et al., 2022); such guides also suggest building distance from content more generally to prevent harm (e.g., Dunkley, 2023; Silverman, 2014; Smith, 2023; Smith & McLellan, 2023), for example, by focusing on details like clothing rather than faces (Porter, 2024). Like the image modification intervention Meta provides, the recommended bulletproof glass strategy does not appear to have an empirical basis and has not yet been tested. However, there is some theoretical basis for predicting that imagining something protective (e.g., bulletproof glass) could be effective.

Imagining bulletproof glass might be effective because it increases psychological distance from the content. Psychological distancing is a form of reappraisal; an emotion regulation strategy that involves thinking about an emotional stimulus differently to reduce its emotional impact (e.g., thinking about a mistake as a learning experience). Distancing specifically involves taking a new perspective that changes the perceived distance of a stimulus (Powers & LaBar, 2019). According to Construal Level Theory, psychological distance can change on four interrelated dimensions (Bar-Anan et al., 2007); temporal distance (how far the stimulus is from the self in the past or future), spatial distance (physical distance of the self to the stimulus), social distance (closeness of the relationship between the self and stimulus) and hypothetical distance (likelihood of stimulus occurring; Trope & Liberman, 2010). There is robust evidence that distancing reduces emotional outcomes. For example, a meta-

analysis including 230 effect sizes from 100 studies revealed an overall medium effect (hedges g = 0.52; 95% CI [0.406, 0.631]) of psychological distance reducing a range of emotional outcomes (e.g., anger and sadness) in response to a range of stimuli (e,g., recalling personal memories or imagining scenarios; Moran & Eyal, 2022).

Previous studies have manipulated psychological distance for personal memories by comparing a self-distanced (e.g., *…take a few steps back from your experience…*) to a self-immersed (e.g., *…relive the situation as if it were happening to you all over again…*) perspective (e.g., Kross & Ayduk, 2017). Taking a self-distanced perspective while recalling emotional experiences decreased negative affect (Kross & Ayduk, 2008; Kross et al., 2005; Kross et al., 2012). In another study, using third person vs. first person pronouns in self-talk (to increase psychological distance) decreased negative affect and shame following a public speech and decreased anxiety for a future speech (Kross et al., 2014). In a systematic review, adopting a third (versus first) person perspective to increase distance was generally associated with reduced positive and negative affect (Wallace-Hadrill & Kamboj, 2016). Distance has also been manipulated for images; for example researchers have asked observers to perceive (in an illusion; Mühlberger et al., 2008) and imagine (Davis et al., 2011) negative images to be moving away from them, resulting in decreased emotional arousal. Applied to the bulletproof glass intervention, perhaps imagining bulletproof glass can make viewers feel more distant from the content, thereby reducing its emotional impact. Indeed, anecdotal evidence from online child exploitation investigators suggests that taking a distanced or detached objective perspective helps to reduce the emotional impact of viewing graphic content (e.g., Burns et al., 2008; Krause, 2009). Reducing the emotional impact of viewing graphic content may also reduce the likelihood of intrusions about the content occurring and the problematic nature of any intrusions (e.g., distressing, vivid, intense), given that finding a task more emotional can contribute to intrusion formation (Marks et al., 2018).

A second explanation for why the bulletproof glass intervention might be effective is based on response expectancy theory (Kirsch, 1985), which suggests that a person's experiences and behaviors are influenced by their expectation of a particular outcome. This theory is one of the prevailing explanations for the placebo effect, which occurs when an inert substance or activity, or a suggestion, produces genuine physiological or psychological change (e.g., Kirsch 1985; Stewart-Williams, 2004). According to a response expectancy account, these improvements in response to a placebo arise from a person's expectation of improvement, rather than the intervention itself (Corsi & Colloca, 2017). For example, a sugar pill administered as an anti-depressant can relieve depression symptoms when the person expects the pill will be effective (Rief et al., 2009); expecting positive consequences to drinking alcohol (e.g., feeling more relaxed and sociable) is related to experiencing those consequences when drinking (LaBrie et al., 2007; Lee et al., 2020); and taking part in low-intensity physical activities (e.g., progressive muscle relaxation) known not to lead to psychological improvement nonetheless has mental health benefits (e.g., on depression; Lindheimer et al., 2015). Applied to the bulletproof glass intervention, perhaps simply having an intervention to use—and believing it will be effective—will reduce the negative impact of content moderation.

To test whether imagining bulletproof glass when viewing graphic content reduces the negative impact of content moderation, we randomly allocated participants to complete a content moderator simulation task either with or without the bulletproof glass instruction. We operationalized negative impact as the frequency and negative characteristics of intrusions, given that more frequent intrusions are associated with greater emotional distress and functional impairment, and reducing the problematic nature of intrusions (e.g., distress, vividness, and intensity) is considered in influencing the persistence and frequency of intrusions (Marks et al., 2018). We also operationalized negative impact as change in anxiety and negative affect given these are common responses to viewing distressing content (e.g., Spence, Bifulco et al., 2023) and finding a task more emotional (e.g., anxiety and negative affect) leads to more frequent intrusions (Marks et al., 2018). Given that the content moderator simulation task involves a titrated dose of content moderation exposure, the task functions as a trauma analogue. Consistent with the trauma analogue literature, immediate intrusion monitoring and measures of emotional impact are appropriate proxies for impact/distress (Holmes & Bourne, 2008; Lau-Zhu et al., 2018).

If imagining bulletproof glass is effective at mitigating the harmful effects of viewing graphic content, we expected that:

1. People exposed to this intervention would report less frequent and problematic intrusions about the content (e.g., distressing, vivid, and intense), than those not exposed to the intervention (control group) and;

2. Anxiety and negative affect would increase, and positive affect would decrease, from before to after completing the simulated content moderation task, but to a lesser extent for the intervention condition compared to the

control condition, demonstrated by an interaction between condition (i.e., intervention, control) and time (pre, post) for negative affect, positive affect, and anxiety[1].

Methods

This study was preregistered on the Open Science Framework (https://osf.io/qgdwu/overview) and the data and materials are available as files at: https://osf.io/kz9tr/overview. The Flinders University Human Research Ethics Committee approved this research.

Participants

Our desired sample size was 200 participants, determined by an a priori power analysis run in G*Power for a between-groups *t*-test at 80% power with α = .05 to detect a small/medium effect (*d* = 0.4; Faul et al., 2007). The same sample size is recommended for a mixed 2 x 2 ANOVA: small/medium effect (*f* = .2) at 80% power with α = .05 (Faul et al., 2007). We recruited participants online through Amazon MTurk, using Cloud Research to only source approved participants who had previously completed 500+ studies with a 95%+ HIT rate (Peer et al., 2014). We also prevented multiple submissions, so participants could not repeatedly complete the survey for financial gain. Participants received $3.00 USD compensation. Participation was entirely voluntary, and participants could withdraw from the study prior to the end of their participation without penalty. Data were collected in August 2024.

To avoid bots/server farmers in our online data (Aguinis et al., 2021), participants who failed a Qualtrics V2 Captcha, or a simple addition question (i.e., *13 + 4 =*) were ineligible to complete our survey. To ensure participants could understand and respond to questions, participants who scored less than 8/10 questions correctly on an English Proficiency Test or failed a cultural familiarity check (i.e., correctly naming an eggplant image) were ineligible to complete our survey (Moeck et al., 2022).

In total, 220 participants completed our survey. We excluded data from 16 participants following our preregistered plan (more detail appears in the Appendix). We also excluded data from one duration outlier (who spent > 190 minutes completing the survey), and one participant who reported technical issues. We retained data from 100 intervention, and 102 control participants.[2]

Design

We used a 2 (condition: bulletproof glass intervention, control) x 2 (time: pre, post) mixed experimental design. The key dependent variables were intrusion frequency and problematic intrusion characteristics, state anxiety, and affect (positive and negative). We also measured previous trauma and related symptoms, and items related to 'self-location' and the bulletproof glass intervention as mechanism checks to understand why the intervention might or might not work.

Procedure

After passing the Captcha, addition, eggplant question and English Proficiency Test, participants provided consent and demographic information (i.e., gender, age, ethnicity, income and education). Next, we measured their baseline state anxiety and affect. Participants then read background information about content moderators and the guidelines images must comply with, developed in previous pilot testing (Lewitzka et al., 2026), based on Facebook's community standards (i.e., a set of guidelines detailing what is and is not allowed on Facebook). We used Facebook's community standards given that Facebook is one of the most popular social media platforms, and shares the same standards with Instagram under Meta. Further, other major platforms such as YouTube, TikTok and X similarly prohibit the core violation we focus on (e.g., violent death, animal abuse, self-harm). Guideline comprehension was tested by six multiple-choice questions. Participants who responded incorrectly to more than two questions were ineligible to continue the study.

Next, participants in the intervention condition received the bulletproof glass instruction: *as you are doing the task, we would like you to imagine that there is a very thick pane of bulletproof glass between you and each image you are moderating*. To continue, participants were required to correctly answer an instruction comprehension question

to confirm they understood and read the instruction. Participants then described what visualizing bulletproof glass between themselves and the images would look like for them based on their current computer screen view (e.g., *it would be on a stand in between my monitor and my desk*).

Then, over a series of trials, participants were presented with an image and indicated whether (*yes/no*) the image violated any of the previously described guidelines. If they selected *yes*, participants chose the content violation type(s) from a list of possible violations (e.g., *violent death*). These decisions simulate those a content moderator would make. Participants completed this process for 32 images, including 16 "objective" violations (95%+ agreement in previous studies that the images violate guidelines; e.g., an image of a dead body), eight subjective content violations (between 38%–90% agreement; e.g., an image of a person at gunpoint), four ambiguous (do not violate guidelines, could be interpreted as negative or neutral; e.g., an image of a crying woman's face) and four neutral images (do not violate guidelines; e.g., an image of a cat). We used the subjective images to replicate the true nature of content moderator decisions (i.e., where there may not always be a high consensus as to whether the image violates guidelines); and the neutral and ambiguous images allowed us to check engagement and appropriate responding. Participants in the control condition moderated all 32 images in a randomised order with no further instruction. Participants in the intervention condition moderated four blocks of eight images (each block including one ambiguous, one neutral, two subjective and four objective images), and were given three brief intervention reminders between each block. The images within each block were presented in a randomized order. These blocks were used to distribute the reminders throughout the full set of images at fixed intervals.

After completing the moderation task, all participants re-completed the state anxiety and affect measures, then completed the intrusion frequency monitoring task and responded to the intrusion characteristic questions (if they reported at least one intrusion). Then, to check whether intrusions were only of the images they viewed, we asked participants to briefly describe their intrusion(s) content. Next, participants responded to the self-location items and participants in the intervention condition responded to the bulletproof glass feelings questions. We asked participants what they thought the study's aims were, and some final questions about their behaviour while doing the study. Participants then completed a mood repair task (writing down a positive memory), and were debriefed. We provided contact details of agencies they could contact for support.

Materials

State-Trait Anxiety Inventory for State Short Form (STAI-S-6)

The STAI-S-6 is a widely used, valid and reliable state anxiety self-report measure (Cronbach's α = .75; Fioravanti-Bastos et al., 2011). Participants rated three anxiety-present (e.g., *I feel tense*) and three anxiety-absent items (e.g., *I feel calm*) in relation to how they felt in the present moment, from (1) *not at all* to (4) *very much*. Summed scores range from 6 to 24, with higher scores indicating higher state anxiety. In our study, the internal consistency was high (α = .91)

Positive and Negative Affect Schedule (PANAS)

The PANAS is a reliable and valid tool measuring self-reported positive and negative affect (positive affect scale α = .89, negative affect scale α = .85; Watson et al., 1988). Participants indicated how they felt in the present moment on a 5-point Likert scale ranging from (1) *very slightly or not at all* (5) to *extremely* for 10 positive affect items (e.g., *excited*) and 10 negative affect items (e.g., *scared*). Scores on each affect scale range from 10 to 50, with higher scores indicating a higher positive or negative affect. Internal consistency was high (positive affect: α = .93, negative affect: α = .93).

Intrusion Frequency Monitoring Task

Participants were given a definition of intrusions and were asked to indicate if they experienced any intrusions of the images they viewed via a keypress while reading an unrelated article for 5-minutes (Oulton et al., 2018). Participants were told not to keypress for intrusive thoughts unrelated to the images viewed earlier.

Intrusion Characteristics Questions

Participants rated 12 statements measuring intrusion(s) characteristics (Oulton et al., 2018). Four statements checked whether intrusions were involuntary (e.g., *The images I viewed earlier came to mind spontaneously*) and eight statements tested problematic intrusion characteristics, namely: *distress, vividness, intrusiveness, unwantedness, emotional intensity, valence, here-and-nowness,* and *suppression* (e.g., *How distressing were the memories for the images you viewed earlier*). Participants responded to all items on a 7-point Likert scale ranging from (1) *not at all* to (7) *completely/extremely*. Valence was rated from (1) *extremely negative* to (7) *extremely positive*.[3]

Secondary Measures

We measured self-location, defined as the viewers' subjective feeling of "being there" (e.g., in the environment of the image; Hartmann et al., 2016) to check whether the bulletproof glass intervention led participants to feel more spatially distant from the images. Participants responded on a 5-point Likert scale ranging from (1) *I do not agree at all* to (5) *I fully agree* for four self-location items adapted from the *Spatial Presence Experience Scale* (e.g., *I felt like I was actually there in the environment of the images*; Hartmann et al., 2016). Scores were summed to provide a total score ranging from 4 to 20, with higher scores indicating closer spatial distance. Internal consistency in this study was high (α = .96).

It is possible that telling participants to imagine bulletproof glass could imply a threat, gunfire or prompt feelings of being unsafe. As such, we developed several items for the present study to assess how the intervention made participants feel. Participants in the intervention condition responded on a 5-point scale ranging from (1) *strongly disagree* to (5) *strongly agree* to four items related to their feelings about imagining the bulletproof glass (e.g., *imagining the bulletproof glass made me feel safer*). We combined the safe/protected items (α = .94) and the threatened/danger items (α = .87) to get two scores representing participants' feelings about the intervention. Participants also described, in an open text box, how they felt when they were imagining the bulletproof glass.

Results and Discussion

Statistical Overview

Our intrusion frequency, intrusion characteristics, and self-location variables did not meet assumptions of normal distribution (Kolmogorov-Smirnov *ps* < .05). For these variables, we ran non-parametric tests (Mann-Whitney U tests; see Appendix Tables A1 and A2) in addition to parametric tests (*t*-tests). Because interpretation (i.e., significance) did not change, we have reported *t*-test results throughout for ease of interpretation.

We ran all analyses using null-hypothesis significance testing. Because null-hypothesis significance testing does not allow for the evaluation of evidence in favor of the null hypothesis, for our main analyses, we also report Bayes factors indicating the odds ratio for the alternative over the null hypothesis (BF_{10}) and the null over the alternative hypothesis (BF_{01}).[4] We used default priors[5], given the absence of strong theoretical or empirical grounds to specify an informed prior and followed Wetzels et al. (2011) guidelines for interpretation.

Main Analyses

We first ran a series of independent samples *t*-tests to evaluate whether imagining bulletproof glass affected how frequent and problematic intrusions were. Means and standard deviations appear in Table 1. We retained one outlier (135 intrusions) in the dataset because removing this outlier did not change the results (Aguinis et al., 2013). Participants across the sample reported experiencing intrusions of the image content after completing the task. However, the intervention did not reduce intrusion frequency, $t(200) = 0.23$, $p = .822$, $d = .03$, 95% CI [−.24, .31]. We also ran a negative binomial regression, given that the intrusion count data were positively skewed, and the variance was greater than the mean. Condition did not account for significant variance in the number of intrusions participants reported, likelihood ratio: $\chi^2(1) = .04$, $p = .844$. That is, the bulletproof glass condition, $B = -0.04$, $SEB = .21$, Exp (B) = 0.96, 95% CI [0.64, 1.44], $p = .844$, did not significantly differ from the control condition, $B = 2.27$, $SEB = .10$, Exp (B) = 9.69, 95% CI [7.97, 11.93], $p < .001$. Indeed, Bayes factors show substantial evidence for the null over the alternative hypothesis ($BF_{01} = 6.38$).

Table 1. *Means and Standard Deviations for Self-Location, Intrusion Frequency and Intrusion Characteristics by Condition (Control, Intervention).*

	Condition		Total
	Control *M* (*SD*)	Intervention *M* (*SD*)	*M* (*SD*)
Self-location	7.94 (4.67)	7.44 (4.35)	7.69 (4.51)
Intrusion Frequency	9.89 (17.25)	9.42 (11.82)	9.66 (14.94)
Intrusion Characteristics			
Deliberate	1.69 (1.04)	1.90 (1.25)	1.80 (1.16)
Intentional	1.50 (0.92)	1.79 (1.25)	1.66 (1.12)
Spontaneous	6.10 (1.34)	5.79 (1.30)	5.92 (1.34)
Effortless	5.93 (1.46)	5.19 (1.58)	5.53 (1.57)
Intrusive	5.87 (1.35)	5.26 (1.69)	5.54 (1.57)
Distress	4.79 (1.59)	4.29 (1.74)	4.52 (1.68)
Vivid	5.60 (1.33)	4.80 (1.55)	5.17 (1.50)
Unwanted	6.40 (1.01)	5.75 (1.53)	6.05 (1.35)
Intensity	4.38 (1.67)	4.06 (1.67)	4.21 (1.67)
Valence[a]	1.90 (1.07)	2.08 (0.97)	1.99 (1.01)
Nowness	4.38 (1.51)	3.88 (1.63)	4.11 (1.59)
Suppress	5.24 (1.63)	4.79 (1.65)	4.99 (1.65)

Note. [a] valence is rated from (1) *very negative* to (7) *very positive*.

However, participants' intrusions were significantly less vivid ($t(146) = 3.36$, $p > .001$, $d = .55$, 95% CI [.22, .88]), unwanted ($t(146) = 2.98$, $p = .003$, $d = .49$, 95% CI [.16, .81]) and intrusive ($t(146) = 2.38$, $p = .019$, $d = .39$, 95% CI [.07, .72]) in the intervention condition than in the control condition, though the evidence according to Bayes factors was variable ($BF_{10} = 2.31 – 27.58$). Although mean scores showed a similar pattern (i.e., better outcomes for the intervention than control condition, with small-medium effect sizes), intrusion intensity ($t(146) = 1.16$, $p = .247$, $d = .19$, 95% CI [−.13, .52]), valence ($t(146) = 1.07$, $p = .289$, $d = −.18$, 95% CI [−.50, .15]), distress ($t(146) = 1.84$, $p = .068$, $d = .30$, 95% CI [−.02, .63]), nowness ($t(146) = 1.97$, $p = .052$, $d = .32$, 95% CI [−.00, .65]), and suppression ($t(146) = 1.65$, $p = .100$, $d = .27$, 95% CI [−.05, .60]) did not significantly differ by condition. Here, there was anecdotal to substantial evidence for the null over the alternative hypothesis ($BF_{01} = 0.99 – 3.35$; see Appendix Table A3). Thus, evidence for the efficacy of the bulletproof glass intervention at reducing problematic intrusion characteristics was mixed.[6]

Next, we ran a series of 2 (condition: intervention, control) x 2 (time: pre, post) mixed ANOVAs to evaluate whether imagining bulletproof glass when viewing graphic content affected participants' change in anxiety, positive affect, and negative affect from before to after completing the moderation task. Means and standard deviations appear in Table 2. Consistent with expectations that this task would have a negative impact, our analyses revealed significant main effects of time for anxiety, $F(1, 200) = 289.55$, $p < .001$, $\eta_p^2 = .59$, $BF_{10} = 8.989 \times 10^{+37}$, positive affect $F(1,200) = 189.97$, $p < .001$, $\eta_p^2 = .49$, $BF_{10} = 2.428 \times 10^{+27}$, and negative affect $F(1, 200) = 160.01$, $p < .001$, $\eta_p^2 = .44$, $BF_{10} = 1.524 \times 10^{+24}$. Bayes factors show decisive evidence for the alternative hypothesis (i.e., that there is a significant effect) over the null hypothesis (i.e., that there is no effect). However, the intervention was unsuccessful in altering emotional responses to the moderation task: we found no significant interactions between time and condition for anxiety, $F(1,200) = .75$, $p = .388$, $\eta_p^2 = .00$, $BF_{01} = 4.23$, positive affect $F(1, 200) = 1.03$, $p = .313$, $\eta_p^2 = .00$, $BF_{01} = 3.49$, or negative affect $F(1, 200) = .11$, $p = .742$, $\eta_p^2 = .00$, $BF_{01} = 6.42$. Bayes factors show substantial evidence for the null over the alternative hypothesis.

Overall, our main analyses suggest that the content moderator task negatively impacted emotional outcomes, consistent with previous content moderator studies (e.g., Karunakaran & Ramakrishan, 2019). Imagining bulletproof glass task reduced intrusion characteristics, particularly vividness, unwantedness and intrusiveness. However, other characteristics including intensity, valence, distress, here-and-nowness and suppression were not significantly different across conditions, suggesting the effect of the bulletproof glass intervention was variable and at best, small. Perhaps the intervention changed the experience of intrusions coming to mind more so than the emotional impact of the intrusions. Indeed, vividness, unwantedness and intrusiveness relate to the subjective

accessibility of the intrusion itself (e.g., how clear and unwanted the intrusion feels), while other characteristics are more about people's affective response to the intrusion (e.g., distress, emotional intensity, valence). However, taking this evidence together, we suggest that imagining bulletproof glass did not substantively and consistently reduce the negative impact of content moderation. We next consider several explanations for this pattern.

Table 2. *Means and Standard Deviations for Anxiety, Positive Affect and Negative Affect by Condition (Control, Intervention) and Time.*

	Condition		Total *M* (*SD*)
	Control *M* (*SD*)	Intervention *M* (*SD*)	
Anxiety			
Baseline	9.12 (3.13)	9.44 (3.95)	9.28 (3.55)
Post	14.34 (4.65)	14.16 (4.67)	14.25 (4.65)
Positive affect			
Baseline	30.11 (8.16)	30.53 (7.97)	30.32 (8.05)
Post	24.55 (7.48)	24.09 (8.00)	24.32 (7.73)
Negative affect			
Baseline	12.13 (4.40)	12.57 (4.76)	12.35 (4.58)
Post	17.49 (7.22)	17.66 (7.39)	17.57 (7.29)

First, we considered whether our bulletproof glass strategy did not consistently reduce negative outcomes because it made people feel threatened rather than safe. Participants tended to report that the bulletproof glass made them feel safer (*M* = 2.59, *SD* = 1.22) and more protected (*M* = 2.74, *SD* = 1.42), relative to making them feel more threatened (*M* = 1.42, *SD* = 0.84) or like they could be in danger (*M* = 1.63, *SD* = 1.09). However, the mean ratings for safe/protected were around the mid-point of the scale, suggesting participants tended to feel ambivalent (did not agree or disagree) that imagining bulletproof glass made them feel safer when viewing the graphic content; participants tended to disagree that the bulletproof glass made them feel threatened or in danger (unsafe).[7]

Second, recall we suggested the strategy might work because it encourages psychological distancing. But perhaps imagining bulletproof glass did not make people feel more distant from the images. Indeed, although we found that participants reported feeling spatially distant (relative to feeling spatially proximal) to the images—i.e., the average score for the self-location items was closer to the (1) *do not agree* anchor than the (5) *fully agree* anchor (intervention *M* = 1.69 − 2.02; control *M* = 1.70 − 2.24)—self-location scores did not significantly differ between the control (*M* = 7.94, *SD* = 4.67) and intervention (*M* = 7.44, *SD* = 4.35) conditions, *t*(200) = 0.79, *p* = .431, *d* = .1, 95% CI [−.17, .39]. Bayes factors show substantial evidence for the null over the alternative hypothesis, BF_{01} = 4.88. Thus, unlike other psychological distance manipulations that have been used previously (e.g., imagining objects are moving away from the observer; Davis et al., 2011), the bulletproof glass intervention did not appear to increase the perceived spatial psychological distance of the images. Perhaps participants already felt spatially distant from the images, given they were behind a computer screen. Indeed, this idea was reported anecdotally: *I don't know if it [the bulletproof glass] made a difference. I felt safe because it's over a computer*.

Despite our manipulation not appearing to directly affect perceived distancing, we wondered whether distancing itself was related to emotional outcomes, as Construal Level Theory suggests (Trope & Liberman, 2010). Indeed, reporting lower self-location scores (indicating higher perceived spatial psychological distance) was associated with decreased anxiety, negative affect, and intrusion distress, vividness, intensity, unwantedness, presence, suppression, and negative emotions (valence), and increased positive affect (*r*s = −.26 to .37; see Appendix Table A4). Self-location did not correlate with how intrusive or frequent intrusions were.[8] However, because the bulletproof glass intervention did not appear to influence perceived psychological distance in the same way other distancing interventions do (e.g., Kross & Ayduk, 2017), we cannot draw any conclusions about the efficacy of distancing from this study. Future research that manipulates perceived spatial distance using a more typical distancing manipulation—like imagining images are moving away from the observer (Davis et al., 2011) or imagining how one would feel about the content if they were very far away from it (Powers & LaBar, 2019) might be effective at reducing emotional reactions for moderators (e.g., Davis et al., 2011).

Third, although we found no overall evidence consistent with a placebo effect, we considered the possibility that the bulletproof glass strategy was only effective for those who believed it was helping them (e.g., through a placebo effect based on response expectancy theory; Corsi & Colloca, 2017; Kirsch, 1985). We note these analyses were exploratory and not included in our pre-registration. We classified participants (in the intervention condition) as either endorsers (thought the intervention was helpful – e.g., *I felt very safe with the bulletproof glass*; *n* = 40) or *non-endorsers* (thought it was unhelpful – e.g., *I don't feel like it [the bulletproof glass] helped*; *n* = 60) based on their qualitative responses to the question: *please describe how you felt when you were imagining the bulletproof glass between yourself and each image in the task*. We then ran a series of 2 (bulletproof glass endorsement: endorsers, non-endorsers) x 2 (time: pre, post) mixed ANOVAs to evaluate whether endorsing the bulletproof glass affected participants' change in anxiety, positive affect, and negative affect. Means, standard deviations, inferential statistics, and Bayes factors appear in Table 3. We found a significant interaction between group and time for anxiety; those who thought the intervention was helpful experienced a greater increase in anxiety from before to after completing the task compared to participants who thought the intervention was unhelpful. However, there was no evidence that believing the bulletproof glass was helpful influenced anxiety, positive affect, or negative affect overall, nor interactions between group and time for positive or negative affect.

Table 3. *Means, Standard Deviations, ANOVA and Bayes Factors (BF10, BF01) for Anxiety, Positive Affect and Negative Affect by Condition (Endorser vs Non-Endorser) and Time.*

Effect	Condition		$F(1, 98)$	p	η_p^2	BF_{10}	BF_{01}
	Endorser *M*(*SD*)	Non-endorser *M*(*SD*)					
Anxiety							
Baseline	8.78 (3.31)	9.88 (4.30)					
Post	14.98 (4.33)	13.62 (4.84)					
Time			137.47	< .001	.56	$8.82 \times 10^{+15}$	1.13×10^{-16}
Endorsement			0.03	.871	.00	0.25	4.00
Time x endorsement			8.48	.004	.03	7.54	0.13
Positive affect							
Baseline	33.08 (6.36)	28.83 (8.51)					
Post	25.63 (7.77)	23.07 (8.06)					
Time			104.58	< .001	.51	$7.09 \times 10^{+13}$	1.41×10^{-14}
Endorsement			5.40	.022	.01	0.27	3.60
Time x endorsement			1.70	.196	.01	0.43	2.44
Negative affect							
Baseline	11.75 (4.18)	13.12 (5.07)					
Post	17.95 (6.77)	17.47 (7.83)					
Time			79.18	< .001	.44	$7.73 \times 10^{+10}$	1.29×10^{-11}
Endorsement			0.15	.695	.00	0.29	3.47
Time x endorsement			2.44	.122	.01	0.59	1.70

We also used *t*-tests to evaluate whether endorsement group affected intrusion frequency, problematic intrusion characteristics, and self-location. Means, standard deviations, inferential statistics, and Bayes factors appear in Table 4. Although there was a pattern of worse outcomes for endorsers, there were no significant differences by group for intrusion frequency, or how vivid, distressing, intense, valanced, present (nowness), supressed, intrusive or unwanted intrusions were. Here, Bayes factors indicated that evidence was inconclusive (BF_{01} = 0.99 − 4.03).

Our data suggest that opposite to a placebo effect (whereby people who expect the intervention will help experience better outcomes), people who thought the intervention helped them tended to report experiencing greater anxiety when completing the task. One possible explanation for this finding is that participants who endorsed the intervention are also those who are more sensitive to negative imagery and thus experienced greater distress during the task. Because of this sensitivity and heightened distress, these participants may have been more motivated to believe that the intervention helped reduce the emotional impact of the task (e.g.,

through motivated reasoning; Kunda, 1990), thus they endorsed the intervention. A similar effect is seen with trigger warnings, where participants who think a trigger warning will help them react less negatively to distressing material report worse outcomes (Bridgland et al., 2022). However, we note that these analyses included only a subset of our participants (i.e., the bulletproof glass intervention condition) and as such are underpowered (n = 40 for endorsers, and n = 60 for non-endorsers). Future research could investigate whether belief in the effectiveness of online content-exposure interventions moderates their success, using adequately powered samples.

Table 4. *Means, Standard Deviations, Inferential Statistics and Bayes Factors (BF_{10}, BF_{01}) for Self-Location, Intrusion Frequency and Problematic Characteristics for Endorser and Non-Endorser Conditions.*

	Condition							
	Endorser	Non-endorser	*df*	*t*	*p*	*d* [95% CI]	BF_{10}	BF_{01}
	M(SD)	*M(SD)*						
Self-location	7.80 (4.40)	7.20 (4.33)	98	0.67	.502	−.14 [−.54, .26]	0.26	3.81
Intrusion frequency	10.25 (12.45)	8.87 (11.45)	98	0.57	.569	−.12 [−.52, .28]	0.25	4.03
Intrusion characteristics								
Intrusive	5.52 (1.46)	5.09 (1.83)	78	1.12	.265	−.26 [−.70, .19]	0.41	2.46
Distress	4.39 (1.48)	4.21 (1.91)	78	0.46	.649	−.10. [−.55, .34]	0.26	3.88
Vivid	4.88 (1.34)	4.74 (1.69)	78	0.38	.705	−.09 [−.53, .36]	0.25	3.99
Unwanted	6.12 (1.19)	5.49 (1.69)	78	1.85	.069	−.42 [−.87, .03]	1.01	0.99
Intensity	4.42 (1.39)	3.81 (1.81)	78	1.64	.105	−.37, [−.82, .08]	0.75	1.34
Valence	2.12 (0.93)	2.04 (1.00)	78	0.36	.722	−.08 [−.53, .37]	0.25	4.02
Nowness	4.09 (1.47)	3.72 (1.73)	78	1.00	.323	−.23 [−.67, .22]	0.36	2.77
Suppress	5.15 (1.40)	4.53 (1.70)	78	1.67	.099	−.38 [−.83, .07]	0.48	1.28

Aside from psychological distance, endorsing the intervention, or feeling unsafe, there are also methodological explanations for our results. For example, perhaps participants did not fully understand the intervention or apply the intervention consistently throughout the study. To combat this issue, we included a multiple-choice question after the first bulletproof glass instruction that aimed to screen out participants who did not understand (two participants were exited from the survey for this reason). Participants also described what visualizing the bulletproof glass would look like for them before starting the task, so they had an opportunity to start imagining the bulletproof glass before seeing the graphic content. We reminded participants of the intervention at three points throughout the study to minimize instances of forgetting. Finally, it is possible that this intervention is simply ineffective. This finding is important given this strategy is already suggested online as an effective way to handle viewing traumatic imagery (e.g., Rees 2017) and thus may provide a false sense of protection, while leaving moderators vulnerable to harm. This research highlights the importance of empirically testing interventions before recommending them to users, and a lack of evidence-based strategies for those exposed to negative content. There is an urgent need for future research to develop and test interventions that reduce the severe impact of content moderation.

Finally, the present study has various limitations worth considering. First, the way we measured intrusions may have inadvertently contributed to experiences of intrusions. Because participants were given a definition of intrusions and were instructed to actively monitor for them, it is possible that this instruction made intrusions more likely to occur (Wenzlaff & Wegner, 2000). Further, our study did not measure persistent distressing intrusive memories, which are the most problematic (Marks et al., 2018). Although measuring intrusions over time was beyond the scope of this study, future research could consider ongoing, persistent intrusion monitoring, for example using a daily diary method (Singh et al., 2023). Next, our study design and content moderator simulation did not fully capture the chronic exposure to graphic content that content moderators experience. Although our study demonstrates limited efficacy for the bulletproof glass intervention in the short term, perhaps effectiveness may differ for repeated exposure to graphic content, given that small effects accumulate over time (Funder & Ozer, 2019). Future research could further explore the efficacy of this intervention, or similar interventions, in a longitudinal study design or in-field with real moderators. Because a small subset of participants reported prior

content moderation experience (n = 13), we reran the primary analyses excluding these participants. The overall pattern and interpretation of results were largely unchanged (some p-values previously < .10 reached < .05 [dance of the p values; Cumming, 2014]; see Appendix / Table A5). Finally, the present study did not measure individual difference factors that predict intrusive memories—like pre-existing anxiety, depression or negative appraisal tendencies (Marks et al., 2018)—which could also be considered in future studies.

This study aimed to address the gap in harm mitigation for populations exposed to graphic content, specifically testing whether imagining bulletproof glass when viewing graphic content can reduce the negative impact of content moderation. Our results suggest the intervention did not consistently reduce the frequency or problematic characteristics of intrusive thoughts in a way that would make a meaningful impact on content moderators' experience. Similarly, the intervention did not influence anxiety, or positive or negative affect experienced in response to the content moderation task. Our findings indicate that the bulletproof glass intervention had limited effectiveness as a harm mitigation strategy for exposure to graphic content. Therefore, we conclude that this strategy should not be recommended to handle traumatic imagery for groups exposed to graphic content, like content moderators.

Footnotes

[1] We preregistered competing hypotheses, corresponding to the pattern of results we would expect for a successful vs. unsuccessful intervention outcome. For simplicity, here we frame our hypotheses around the expected pattern for a successful intervention outcome.

[2] We excluded data from the intrusion rating measures only for one participant in the intervention condition who completed the measures but later reported not having any intrusions (these data were not included in the intrusion frequency analysis). We included their responses to the affect and anxiety measures per our pre-registered plan.

[3] See https://osf.io/kz9tr/overview for full items and response formats.

[4] These analyses were not pre-registered.

[5] We acknowledge the limit of their interpretive value in the absence of strong prior information.

[6] Descriptive statistics (collapsed across conditions; reported in Table 1) for intrusion characteristics measuring retrieval ease and whether retrieval was intentional suggest that intrusions were involuntary (rather than voluntary; Berntsen, 2010). See Appendix.

[7] We used correlations to see whether the extent participants agreed they felt safe or unsafe influenced emotional reactions, noting these analyses were exploratory and not pre-registered. Feeling unsafe correlated with outcomes, suggesting those who felt more unsafe experienced worse outcomes. However, feeling safe did not correlate with outcomes, suggesting that feeling safe did not reduce the impact of viewing the graphic content. More detail appears in the Appendix.

[8] These correlations were exploratory and not pre-registered.

Conflict of Interest

The authors have no conflicts of interest to declare.

Use of AI Services

The authors declare they have not used any AI services to generate any part of the manuscript or data.

Data Availability Statement

This study was preregistered on the Open Science Framework (https://osf.io/qgdwu) and the data and materials are publicly available as files at: https://osf.io/kz9tr/files/osfstorage.

Authors' Contribution

Chloe McDonough: conceptualization, methodology, investigation, data curation, formal analysis, writing—original draft. **Sarah Lewitzka:** conceptualization, methodology, data curation, writing—original draft, writing—review & editing. **Victoria M. E. Bridgland:** methodology, writing—original draft. **Ella K. Moeck:** funding acquisition, writing—review & editing. **Reginald D. V. Nixon:** funding acquisition, writing—review & editing. **Carolyn Semmler:** funding acquisition, writing—review & editing. **Melanie K. T. Takarangi:** conceptualization, funding acquisition, investigation, supervision, project administration, writing—review & editing.

Acknowledgement

The authors thank the participants for their time and contribution to this research.

Funding Declaration

This research was supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP230100906).

References

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods, 16*(2), 270–301. https://doi.org/10.1177/1094428112470848

Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk research: Review and recommendations. *Journal of Management, 47*(4), 823–837. https://doi.org/10.1177/0149206320969787

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Association Publishing. https://doi.org/10.1176/appi.books.9780890425787

Bar-Anan, Y., Liberman, N., Trope, Y., & Algom, D. (2007). Automatic processing of psychological distance: evidence from a Stroop task. *Journal of Experimental Psychology: General, 136*(4), 610–622. https://doi.org/10.1037/0096-3445.136.4.610

Berntsen, D. (2010). The unbidden past: Involuntary autobiographical memories as a basic mode of remembering. *Current Directions in Psychological Science, 19*(3), 138–142. https://doi.org/10.1177/0963721410370301

Bridgland, V. M. E., Barnard, J. F., & Takarangi, M. K. T. (2022). Unprepared: Thinking of a trigger warning does not prompt preparation for trauma-related content. *Journal of Behavior Therapy and Experimental Psychiatry, 75*, Article 101708. https://doi.org/10.1016/j.jbtep.2021.101708

Browne, T., Evangeli, M., & Greenberg, N. (2012). Trauma-related guilt and posttraumatic stress among journalists. *Journal of Traumatic Stress, 25*(2), 207–210. https://doi.org/10.1002/jts.21678

Burns, C. M., Morley, J., Bradshaw, R., & Domene, J. (2008). The emotional impact on and coping strategies employed by police teams investigating internet child exploitation. *Traumatology, 14*(2), 20–31. https://doi.org/10.1177/1534765608319082

Centre for Information Resilience. (2023, November 16). *Let's talk about mental health*. https://www.info-res.org/cir/articles/lets-talk-about-mental-health/

Collins, S., & Long, A. (2003). Working with the psychological effects of trauma: Consequences for mental health-care workers – A literature review. *Journal of Psychiatric and Mental health Nursing, 10*(4), 417–424. https://doi.org/10.1046/j.1365-2850.2003.00620.x

Cook, C. L., Cai, J., & Wohn, D. Y. (2022). Awe Versus aww: The effectiveness of two kinds of positive emotional stimulation on stress reduction for online content moderators. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW2), 1–19. https://doi.org/10.1145/3555168

Corsi, N., & Colloca, L. (2017). Placebo and nocebo effects: The advantage of measuring expectations and psychological factors. *Frontiers in Psychology, 8*, Article 308. https://doi.org/10.3389/fpsyg.2017.00308

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Davis, J. I., Gross, J. J., & Ochsner, K. N. (2011). Psychological distance and emotional experience: What you see is what you get. *Emotion, 11*(2), 438–444. https://doi.org/10.1037/a0021783

Dosono, B., & Semaan, B. (2019). Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit. In *Proceedings of the 2019 CHI Conference on Human factors in Computing Systems.* (Article 142, pp. 1–13). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300372

Duarte, F. (2024, June 13). *Amount of Data Created Daily (2024)*. Exploding Topics. https://explodingtopics.com/blog/data-generated-per-day

Dubberley, S., Griffin, E., & Mert Bal, H. (2015). *Making secondary trauma a primary issue: A study of eyewitness media and vicarious trauma on the digital frontline.* Eyewitness Media Hub. https://firstdraftnews.org/wp-content/uploads/2018/03/trauma_report.pdf

Dunkley, F. (2023, June 30). *Exposure to traumatic material – 3 stages of support*. FD Consultants. https://fdconsultants.net/3-stages-of-support/

Ehlers, A., & Steil, R. (1995). Maintenance of intrusive memories in posttraumatic stress disorder: A cognitive approach. *Behavioural and Cognitive Psychotherapy, 23*(3), 217–249. https://doi.org/10.1017/S135246580001585X

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*(2), 175–191. https://doi.org/10.3758/BF03193146

Fioravanti-Bastos, A. C. M., Cheniaux, E., & Landeira-Fernandez, J. (2011). Development and validation of a short-form version of the Brazilian state-trait anxiety inventory. *Psicologia, reflexão e crítica, 24*(3), 485–494. https://doi.org/10.1590/S0102-79722011000300009

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

Hartmann, T., Wirth, W., Schramm, H., Klimmt, C., Vorderer, P., Gysbers, A., Böcking, S., Ravaja, N., Laarni, J., Saari, T., Gouveia, F., & Sacau, A. M. (2016). The Spatial Presence Experience Scale (SPES): A short self-report measure for diverse media settings. *Journal of Media Psychology: Theories, Methods, and Applications, 28*(1), 1–15. https://doi.org/10.1027/1864-1105/a000137

Holman, E. A., Garfin, D. R., Lubens, P., & Silver, R. C. (2020). Media exposure to collective trauma, mental health, and functioning: does it matter what you see? *Clinical Psychological Science, 8*(1), 111–124. https://doi.org/10.1177/2167702619858300

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

Karunakaran, S., & Ramakrishan, R. (2019). Testing stylistic interventions to reduce emotional impact of content moderation workers. In E. Law & J. Wortman Vaughan (Eds.), *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (Vol. 7, pp. 50–58). PKP Publishing Services. https://doi.org/10.1609/hcomp.v7i1.5270

Kemp, S. (2024, January 31). *Digital 2024: Global overview report*. Datareportal. https://datareportal.com/reports/digital-2024-global-overview-report

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

Krause, M. (2009). Identifying and managing stress in child pornography and child exploitation investigators. *Journal of Police and Criminal Psychology, 24*(1), 22–29. https://doi.org/10.1007/s11896-008-9033-8

Kross, E., & Ayduk, O. (2008). Facilitating adaptive emotional analysis: Distinguishing distanced-analysis of depressive experiences from immersed-analysis and distraction. *Personality and Social Psychology Bulletin, 34*(7), 924–938. https://doi.org/10.1177/0146167208315938

Kross, E., & Ayduk, O. (2017). Self-distancing: Theory, research, and current directions. In J. M. Olson (Ed.), *Advances in experimental social psychology* (Vol. 55, pp. 81–136). Elsevier. https://doi.org/10.1016/bs.aesp.2016.10.002

Kross, E., Ayduk, O., & Mischel, W. (2005). When asking "why" does not hurt: Distinguishing rumination from reflective processing of negative emotions. *Psychological Science, 16*(9), 709–715. https://doi.org/10.1111/j.1467-9280.2005.01600.x

Kross, E., Bruehlman-Senecal, E., Park, J., Burson, A., Dougherty, A., Shablack, H., Bremner, R., Moser, J., & Ayduk, O. (2014). Self-talk as a regulatory mechanism: How you do it matters. *Journal of Personality and Social Psychology, 106*(2), 304–324. https://doi.org/10.1037/a0035173

Kross, E., Gard, D., Deldin, P., Clifton, J., & Ayduk, O. (2012). "Asking why" from a distance: Its cognitive and emotional consequences for people with major depressive disorder. *Journal of Abnormal Psychology, 121*(3), 559–569. https://doi.org/10.1037/a0028808

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

LaBrie, J. W., Hummer, J. F., & Pedersen, E. R. (2007). Reasons for drinking in the college student context: The differential role and risk of the social motivator. *Journal of Studies on Alcohol and Drugs, 68*(3), 393–398. https://doi.org/10.15288/jsad.2007.68.393

Lau-Zhu, A., Holmes, E. A., & Porcheret, K. (2018). Intrusive memories of trauma in the laboratory: Methodological developments and future directions. *Current Behavioral Neuroscience Reports, 5*(1), 61–71. https://doi.org/10.1007/s40473-018-0141-1

Lee, C. M., Fairlie, A. M., Ramirez, J. J., Patrick, M. E., Luk, J. W., & Lewis, M. A. (2020). Self-fulfilling prophecies: Documentation of real-world daily alcohol expectancy effects on the experience of specific positive and negative alcohol-related consequences. *Psychology of Addictive Behaviors, 34*(2), 327–334. https://doi.org/10.1037/adb0000537

Lewitzka, S., Liebeknecht, A., Bridgland, V. M. E., Moeck, E. K., Nixon, R. D. V., Semmler, C., & Takarangi, M. K. T. (2026). Evaluating image modification as a harm reduction approach in content moderation. *Psychological Research, 90*(2), Article 62. https://doi.org/10.1007/s00426-026-02241-5

Lindheimer, J. B., O'Connor, P. J., & Dishman, R. K. (2015). Quantifying the placebo effect in psychological outcomes of exercise training: A meta-analysis of randomized trials. *Sports Medicine, 45*, 693–711. https://doi.org/10.1007/s40279-015-0303-1

Marks, E. H., Franklin, A. R., & Zoellner, L. A. (2018). Can't get it out of my mind: A systematic review of predictors of intrusive memories of distressing events. *Psychological Bulletin, 144*(6), 584–640. https://doi.org/10.1037/bul0000132

Meta. (2024, November 12). *Helping reviewers make the right calls.* Meta Transparency Center. https://transparency.meta.com/en-gb/enforcement/detecting-violations/making-the-right-calls/

Moeck, E. K., Bridgland, V. M. E., & Takarangi, M. K. T. (2022). Food for thought: Commentary on Burnette et al. (2021) "Concerns and recommendations for using Amazon MTurk for eating disorder research". *International Journal of Eating Disorders, 55*(2), 282–284. https://doi.org/10.1002/eat.23671

Moran, T., & Eyal, T. (2022). Emotion regulation by psychological distance and level of abstraction: Two meta-analyses. *Personality and Social Psychology Review, 26*(2), 112–159. https://doi.org/10.1177/10888683211069025

Mühlberger, A., Neumann, R., Wieser, M. J., & Pauli, P. (2008). The impact of changes in spatial distance on emotional responses. *Emotion, 8*(2), 192–198. https://doi.org/10.1037/1528-3542.8.2.192

Newton, C. (2019, December 16). *The terror queue. These moderators help keep Google and YouTube free of violent extremism — and now some of them have PTSD*. The Verge. https://www.theverge.com/2019/12/16/21021005/google-youtube-moderators-ptsd-accenture-violent-disturbing-content-interviews-video

Newton, C. (2020, May 12). *Facebook will pay $52 million in settlement with moderators who developed PTSD on the job*. The Verge. https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health

Oulton, J. M., Strange, D., Nixon, R. D., & Takarangi, M. K. T. (2018). PTSD and the role of spontaneous elaborative "nonmemories". *Psychology of Consciousness: Theory, Research, and Practice, 5*(4), 398–413. https://doi.org/10.1037/cns0000158

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods, 46*(4), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y

Perez, L. M., Jones, J., Englert, D. R., & Sachau, D. (2010). Secondary traumatic stress and burnout among law enforcement investigators exposed to disturbing media images. *Journal of Police and Criminal Psychology, 25*(2), 113–124. https://doi.org/10.1007/s11896-010-9066-7

Porter, B. (2024). *Staying healthy when working with graphic imagery & content*. Thrive Worldwide. https://thrive-worldwide.org/wp-content/uploads/2024/10/Staying-Healthy-When-Working-With-Graphic-Imagery-Content.pdf

Powers, J. P., & LaBar, K. S. (2019). Regulating emotion through distancing: A taxonomy, neurocognitive model, and supporting meta-analysis. *Neuroscience and Biobehavioral Reviews, 96*, 155–173. https://doi.org/10.1016/j.neubiorev.2018.04.023

Rees, G. (2017, April 4). *Developing your own standard operating procedure for handling traumatic imagery*. Global Center for Journalism & Trauma. https://dartcenter.org/resources/handling-traumatic-imagery-developing-standard-operating-procedure

Rief, W., Nestoriuc, Y., Weiss, S., Welzel, E., Barsky, A. J., & Hofmann, S. G. (2009). Meta-analysis of the placebo response in antidepressant trials. *Journal of Affective Disorders, 118*(1–3), 1–8. https://doi.org/10.1016/j.jad.2009.01.029

Roberts, S. T. (2016). Commercial content moderation: Digital laborers' dirty work. In S. U. Noble & B. Tynes (Eds.), *The intersectional internet: Race, sex, class and culture online*. Peter Lang Publishing. https://uwo.scholaris.ca/items/ae23137f-960c-422c-a862-5bf54ee9cfa2

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press. https://doi.org/10.2307/j.ctvhrcz0v

Robinson, J. S., & Larson, C. (2010). Are traumatic events necessary to elicit symptoms of posttraumatic stress? *Psychological Trauma: Theory, Research, Practice, and Policy, 2*(2), 71–76. https://doi.org/10.1037/a0018954

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?" *Journal of Research in Personality, 74*, 194. https://doi.org/10.1016/j.jrp.2018.02.010

Schöpke-Gonzalez, A. M., Atreja, S., Shin, H. N., Ahmed, N., & Hemphill, L. (2022). Why do volunteer content moderators quit? Burnout, conflict, and harmful behaviors. *New Media & Society, 26*(10), 5677–5701. https://doi.org/10.1177/14614448221138529

Silverman, C. (Ed.). (2014). *Verification handbook: A definitive guide to verifying digital content for emergency coverage*. European Journalism Centre. https://verificationhandbook.com/downloads/verification.handbook.pdf

Singh, L., Pihlgren, S. A., Holmes, E. A., & Moulds, M. L. (2023). Using a daily diary for monitoring intrusive memories of trauma: A translational data synthesis study exploring convergent validity. *International Journal of Methods in Psychiatric Research, 32*(1), Article e1936. https://doi.org/10.1002/mpr.1936

Smith, E. (2023, November 13). *Managing trauma exposure*. The Walkley Foundation. https://www.walkleys.com/professional-development/managing-trauma-exposure/

Smith, E., & McLellan, T. (2023, October 26). *Working with traumatic imagery can affect your mental health*. Melbourne Press Club. https://www.melbournepressclub.com/article/working-with-traumatic-imagery-can-affect-your-mental-health-mpcmnl

Spence, R., Bifulco, A., Bradbury, P., Martellozzo, E., & DeMarco, J. (2023). The psychological impacts of content moderation on content moderators: A qualitative study. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *17*(4), Article 8. https://doi.org/10.5817/CP2023-4-8

Spence, R., Harrison, A., Bradbury, P., Bleakley, P., Martellozzo, E., & DeMarco, J. (2023). Content moderators' strategies for coping with the stress of moderating content online. *Journal of Online Trust & Safety, 1*(5). https://doi.org/10.54501/jots.v1i5.91

Steiger, M., Bharucha, T. J., Torralba, W., Savio, M., Manchanda, P., & Lutz-Guevara, R. (2022). Effects of a novel resiliency training program for social media content moderators. In S. Sherratt, A. Joshi, N. Dey, & X.-S. Yang (Eds.), *Proceedings of 7th International Congress on Information and Communication Technology* (Vol. 465, pp. 283–298). Springer. https://doi.org/10.1007/978-981-19-2397-5_27.

Steiger, M., Bharucha, T. J., Venkatagiri, S., Riedl, M. J., & Lease, M. (2021). The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 341, pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445092

Stewart-Williams, S. (2004). The placebo puzzle: Putting together the pieces. *Health Psychology, 23*(2), 198–206. https://doi.org/10.1037/0278-6133.23.2.198.

Storm, H., Crowley, J., & Berrie, A. (2022). *Vicarious trauma. A guide for journalists and media professionals*. Headlines. https://www.mind.org.uk/media/4tybnie0/headlines-guide-to-vicarious-trauma.pdf

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*(2), 440–463. https://doi.org/10.1037/a0018963

Wallace-Hadrill, S. M. A., & Kamboj, S. K. (2016). The impact of perspective change as a cognitive reappraisal strategy on affect: A systematic review. *Frontiers in Psychology, 7*, Article 1715. https://doi.org/10.3389/fpsyg.2016.01715

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*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Wenzlaff, R. M., & Wegner, D. M. (2000). Thought suppression. *Annual Review of Psychology, 51*(1), 59–91. https://doi.org/10.1146/annurev.psych.51.1.59

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6*(3), 291–298. https://doi.org/10.1177/1745691611406923

Appendix

Table A1. *Mann-Whitney U Results for Self-Location, Intrusion Frequency and Intrusion Characteristics by Condition (Control, Distance).*

Variable	U	*p*
Self-location	4,981	.768
Intrusion frequency	4,661	.286
Intrusion characteristics		
Deliberate	2,484.5	.315
Intentional	2,364	.110
Spontaneous	2,188	.030
Effortless	1,929.5	.002
Intrusive	2,166.5	.027
Distress	2,263.5	.074
Vivid	1,888.5	.001
Unwanted	2,042	.005
Intensity	2,434	.264
Valence	2,361	.145
Nowness	2,160.5	.028
Suppress	2,252.5	.067

Table A2. *Mann-Whitney U Results for Intrusion Frequency, Self-Location and Intrusion Characteristics by Condition (Endorser, Non-Endorser).*

Variable	U	*p*
Self-location	1,059	.308
Intrusion frequency	1,065	.340
Intrusion characteristics		
Intrusive	688	.380
Distress	738.5	.714
Vivid	770.5	.960
Unwanted	612	.093
Intensity	621	.124
Valence	727	.619
Nowness	675	.317
Suppress	636	.165

Table A3. *Bayes Factors (BF_{10}, BF_{01}) for Intrusion Characteristics.*

Intrusion Characteristics	BF_{10}	BF_{01}
Deliberate	0.31	3.26
Intentional	0.55	1.83
Spontaneous	0.53	1.88
Effortless	8.55	0.12
Intrusive	2.31	0.43
Distress	0.83	1.21
Vivid	27.58	0.04
Unwanted	9.64	0.10
Intensity	0.33	3.04
Valence	0.30	3.35
Nowness	1.02	0.99
Suppress	0.62	1.62

Table A4. *Correlations (and 95% CIs) Between Threat/Danger Feelings, Safe/Protected Feelings, Change in Anxiety, Change in Positive Affect, Change in Negative Affect, Self-Location, Intrusion Frequency and Problematic Intrusion Characteristics.*

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Threat/danger feelings [a]	—													
2. Safe/protected feelings	−.08 [−.27, .12]	—												
3. Change in anxiety	.35*** [.17, .51]	.19 [−.00, .38]	—											
4. Change in positive affect	−.26** [−.43, −.07]	−.17 [−.35, .03]	−.43*** [−.54, −.31]	—										
5. Change in negative affect	.38*** [.20, .54]	.12 [−.08, .31]	.67*** [.58, .74]	−.40*** [−.51, −.28]	—									
6. Self-location	.52*** [.36, .65]	.15 [−.04, .34]	.29*** [.16, .41]	−.26*** [−.39, −.13]	.34*** [.21, .46]	—								
7. Intrusion frequency	.28** [.09, .45]	−.06 [−.26, .14]	.15* [.01, .28]	−.06 [.19, .08]	.08 [−.06, .22]	.13 [−.01. .26]	—							
8. Intrusion intrusiveness	.09 [−.13, .31]	−.08 [−.24, .08]	.18** [.02, .33]	.05 [−.11, .21]	.92*** [.50, .71]	.16 [−.01, .31]	.18* [.01, .33]	—						
9. Intrusion distress	.41*** [.21, .58]	−.04 [−.26, .18]	.47*** [.33, .58]	−.27*** [−.42, −.12]	.52*** [.39, .63]	.36*** [.22, .50]	.19* [.03, .34]	.49*** [.36, .60]	—					
10. Intrusion vividness	.20 [−.02, .40]	−.01 [−.23, .21]	.30*** [.14, .44]	−.15 [−.30, .01]	.25** [.09, .39]	.32*** [.17, .46]	.17* [.02, .32]	.45*** [.32, .57]	.57*** [.45, .67]	—				
11. Intrusion unwantedness	.14 [−.08, .35]	.09 [−.14, .30]	.33*** [.18, .46]	−.08 [−.23, .09]	.29*** [.14, .43]	.16* [.00, .32]	.11 [−.05, .27]	.61*** [.50, .71]	.42*** [.27, .54]	.41*** [.27, .54]	—			
12. Intrusion intensity	.40*** [.19, .57]	.14 [−.08, .35]	.44*** [.30, .56]	−.37*** [−.51, −.23]	.50*** [.37, .62]	.37*** [.22, .50]	.21* [.05, .36]	.35*** [.20, .49]	.76*** [.68, .82]	.54*** [.42, .65]	.32*** [.17, .46]	—		
13. Intrusion valence	−.39*** [−.56, −.18]	.09 [−.13, .30]	−.36*** [−.49, −.21]	.25** [.09, .39]	−.36*** [−.49, −.21]	−.23** [−.38, −.07]	−.08 [−.24, .08]	−.40*** [−.52, −.25]	−.51*** [−.62, −.38]	−.40*** [−.53, −.26]	−.36*** [−.50, −.21]	−.53*** [−.64, −.40]	—	
14. Intrusion presence (nowness)	.31** [.10, .50]	.22 [−.00, .42]	.40*** [.25, .52]	−.22** [−.37, −.06]	.39*** [.24, .52]	.39*** [.24, .52]	.18 [.02, .33]	.40*** [.25, .52]	.51*** [.38, .62]	.51*** [.38, .62]	.40*** [.26, .53]	.54*** [.42. .65]	−.36*** [−.50, −.21]	—
15. Intrusion suppression	.13 [−.09, .34]	.19 [−.03, .39]	.26** [.10, .40]	−.13 [−.29, .03]	.21** [.05, .36]	.18* [.02, .33]	.05 [−.11, .21]	.45*** [.31, .60]	.33*** [.18, .47]	.16* [.00, .32]	.54*** [.41, .64]	.34*** [.19, .48]	−.35*** [−.48, −.20]	.35*** [.20, .49]

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$; [a] that is, feelings about the bulletproof glass intervention.

Table A5. *Sensitivity Analyses Excluding Participants With Prior Content Moderation Experience.*

Outcome	Full Sample	Excluding Content Moderators
Anxiety × Condition	$F(1, 200) = 0.75$, $p = .388$, $\eta_p^2 = .00$	$F(1, 187) = 1.08$, $p = .300$, $\eta_p^2 = .01$
Negative Affect × Condition	$F(1, 200) = 0.11$, $p = .742$, $\eta_p^2 = .00$	$F(1, 187) = 0.23$, $p = .636$, $\eta_p^2 = .00$
Positive Affect × Condition	$F(1, 200) = 1.02$, $p = .313$, $\eta_p^2 = .01$	$F(1, 187) = 0.32$, $p = .570$, $\eta_p^2 = .00$
Intrusion frequency	$t(200) = 0.23$, $p = .822$, $d = 0.03$	$t(187) = 0.00$, $p = .999$, $d = 0.00$
Intrusiveness	$t(146) = 2.38$, $p = .019$, $d = 0.39$	$t(135) = 2.13$, $p = .035$, $d = 0.37$
Distress	$t(146) = 1.84$, $p = .068$, $d = 0.30$	$t(135) = 2.06$, $p = .041$, $d = 0.36$
Vividness	$t(146) = 3.36$, $p = .001$, $d = 0.55$	$t(135) = 3.55$, $p < .001$, $d = 0.61$
Unwantedness	$t(146) = 2.98$, $p = .003$, $d = 0.49$	$t(135) = 2.73$, $p = .007$, $d = 0.47$
Intensity	$t(146) = 1.16$, $p = .247$, $d = 0.19$	$t(135) = 1.40$, $p = .164$, $d = 0.24$
Valence	$t(146) = -1.06$, $p = .289$, $d = -0.18$	$t(135) = -1.48$, $p = .142$, $d = -0.25$
Here-Nowness	$t(146) = 1.96$, $p = .052$, $d = 0.32$	$t(135) = 2.29$, $p = .023$, $d = 0.40$
Suppression	$t(146) = 1.65$, $p = .100$, $d = 0.27$	$t(135) = 1.34$, $p = .182$, $d = 0.23$

We excluded data from 16 participants including 11 participants who incorrectly responded to multiple objective violations and 5 participants who incorrectly responded to neutral images.

Descriptive statistics (collapsed across conditions; reported in Table 2) for intrusion characteristics measuring retrieval ease and whether retrieval was intentional suggest that intrusions were involuntary (rather than voluntary; Berntsen, 2010). This check was important because the intrusion monitoring task instructions (i.e., *indicate every time you experience an involuntary memory of the images*) may have led participants to intentionally retrieve memories of the images. We found low ratings (compared to the scale anchor (7) *completely/extremely*) of how deliberate and intentional intrusive thoughts were (measuring whether retrieval was intentional), and higher ratings (compared to the scale anchor (1) *not at all*) for how spontaneous and effortless intrusions were (measuring retrieval ease).

The threatened/danger items (combined) correlated with change in anxiety, positive affect and negative affect, self-location, intrusion frequency and intrusion distress, intensity valence and presence (nowness). This pattern suggests that as people felt more threatened/in danger (perhaps because of the intervention), they also reported more anxiety, negative affect, intrusions and problematic intrusion characteristics, and lower positive affect. Thus, those who did feel unsafe may have experienced even worse emotional reactions. Therefore overall, applying the intervention might have detrimental effects, given that it may make emotional reactions worse for a sub-set of people who feel unsafe. This finding is important given that the intervention has been suggested openly online as a tactic to handle traumatic imagery (Rees, 2017). Furthermore, the safe/protected items (combined) did not correlate with any outcomes. In other words, feeling safer did not influence emotional outcomes, perhaps because the content viewed is so graphic and the emotional detriment of content moderation is so severe (Spence, Harrison, et al., 2023). However, a limitation of these findings is that the relationship between safe/protected feelings and threatened/in danger feelings and emotional outcomes is correlational and likely underpowered. Correlations are known to stabilise with a sample size of 260 (Schönbrodt & Perugini, 2013, 2018) and our sample was $n = 100$ (and $n = 80$ for intrusion characteristics, because not all participants experienced intrusions). Therefore, our confidence in this finding is tentative.

About Authors

Chloe McDonough was an honours student in psychology at Flinders University in 2024.

https://orcid.org/0009-0005-5003-131X

Sarah Lewitzka is a doctoral researcher at Flinders University whose research focuses on the psychological impact of online content moderation.

https://orcid.org/0009-0003-6045-9745

Victoria M. E. Bridgland is a Lecturer at Flinders University with research interests in digital media and wellbeing, trauma, expectancy effects, and social media.

https://orcid.org/0000-0002-8865-8426

Ella K. Moeck is a Lecturer at the University of Adelaide whose research focuses on cognition, emotion, and wellbeing.

https://orcid.org/0000-0002-5300-7316

Reginald D. V. Nixon is a Professor at Flinders University with research interests in posttraumatic stress disorder and mental health more broadly.

https://orcid.org/0000-0003-1507-8428

Carolyn Semmler is a Professor at the University of Adelaide whose research focuses on cognition, judgement and decision-making, and applied psychological research.

https://orcid.org/0000-0001-7912-293X

Melanie K. T. Takarangi is a Professor at Flinders University whose research focuses on trauma, memory, and psychological wellbeing.

https://orcid.org/0000-0002-6006-8045

Correspondence to
Melanie Takarangi, School of Psychology, Flinders University, GPO Box 2100, Adelaide, 5001 SA, Australia,
melanie.takarangi@flinders.edu.au