Empirical Article

ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

Clinical Psychological Science
1–18
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/21677026221097618
www.psychologicalscience.org/CPS
SAGE

Curiosity Disturbed the Cat: Instagram's Sensitive-Content Screens Do Not Deter Vulnerable Users From Viewing Distressing Content

Victoria M. E. Bridgland[1], Benjamin W. Bellet[2], and Melanie K. T. Takarangi[1]

[1]College of Education, Psychology & Social Work, Flinders University, and [2]Department of Psychology, Harvard University

Abstract

In an attempt to mitigate the negative impact of graphic online imagery, Instagram has introduced *sensitive-content screens*—graphic images are obfuscated with a blur and accompanied by a warning. Sensitive-content screens purportedly allow "vulnerable people" with mental-health concerns to avoid potentially distressing content. However, no research has assessed whether sensitive-content screens operate as intended. Here we examined whether people, including vulnerable users (operationalized as people with more severe psychopathological symptoms, e.g., depression), use the sensitive-content screens as a tool for avoidance. In two studies, we found that the majority of participants (80%–85%) indicated a desire (Study 1) or made a choice (Study 2) 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.

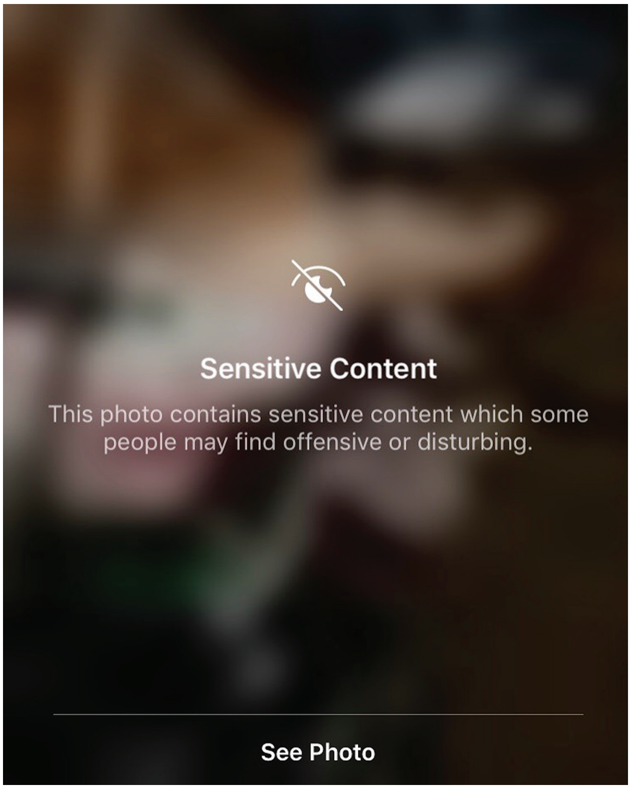
Keywords

avoidance, cognition and emotion, curiosity, social media, trigger warnings, open data, open materials, and preregistration

Received 3/9/21; Revision accepted 3/9/22

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). In addition to completely removing content related to self-harm, part of Instagram's mental-health initiative involves adding sensitive-content 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, 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 et al., 2007), Instagram claims that minimizing exposure to negative content via sensitive-content screens helps to preserve mental health (Mosseri, 2019b). However, there is currently no research that has assessed whether sensitive-content screens operate as intended. To address this gap in knowledge, we examined whether people, including

Corresponding Author:
Melanie Takarangi, College of Education, Psychology & Social Work, Flinders University
Email: melanie.takarangi@flinders.edu.au



Sensitive Content

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

Fig. 1. Example of a real Instagram sensitive-content screen used in Study 1. In Study 1, we used the sensitive-content 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 (for details, see https://osf.io/2fdr7), 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.

vulnerable users, use the sensitive-content screens to minimize their exposure to negative content. First, we investigated whether sensitive-content 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 sensitive-content screens. It is thus surprising that no research has investigated sensitive-content 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 toward material (Bellet et al., 2020; Boysen et al., 2021; Bridgland et al., 2019; Sanson et al., 2019). At worst, trigger warnings create anticipatory anxiety before people view 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 et al., 2018). In fact, early research has shown that trigger warnings may be the most deleterious for the very people they are intended to protect. For example, Jones et al. (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 (vs. unwarned). Event centrality—the belief that a traumatic event marks a turning point in one's life story—is associated with posttraumatic stress disorder (PTSD) symptoms (Berntsen & Rubin, 2006) and prospectively predicts more severe PTSD (Boals & Ruggero, 2016). Moreover, Bridgland and Takarangi (2021) found that warning messages prolonged the negative characteristics (e.g., PTSD-like symptoms) associated with recalling a negative memory over time.

Although trigger warnings have trivial effects on responses to potentially distressing material at best, the primary purpose of sensitive-content screens is to allow people who may have mental-health vulnerabilities to avoid or minimize exposure to potentially distressing content. Therefore, whether there is any evidence that such warning methods actually deter people from approaching potentially negative material must first be considered. 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). Likewise, in Bruce and Roberts (2020), members of the general population and trauma survivors showed equal preference for news articles labeled 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 (vs. no warning). Therefore, early evidence that focuses specifically on trigger warnings suggests that sensitive-content 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) warning about social-media censorship led to stronger feelings of aggression and support of social protests (Ng et al., 2021). Similar patterns have been observed for warnings on violent television shows (Bushman & Stack, 1996) and video games (Bijvank et al., 2009). Therefore, it is possible that sensitive-content screens make viewing images more attractive, which makes 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 sensitive-content 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). Likewise, Oosterwijk (2017) found that participants deliberately chose to view images that portrayed death, violence, and harm over nonnegative 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 (Silvia & Kashdan, 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 sensitive-content screens do not provide any information about the kind of content that is blurred, they foster uncertainty. Furthermore, the accompanying warning message informs the viewer that the content will be negative. Thus, it is possible that because of the Pandora effect, these screens do not deter users.

Given previous research, it seems likely that sensitive-content 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 sensitive-content 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 et al., 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). Likewise, people with or at risk of depression (vs. healthy control subjects) 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 et al., 2015).

There are several theoretical perspectives that may help researchers 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 (Brashers & Hogan, 2013). Indeed, the desire to make meaning of a traumatic experience was the best predictor of how often participants self-triggered (Bellet, Jones, & McNally, 2020). Second, in line with Zillmann's (1988) Mood Management Theory, we know that people often use media to regulate mood. Although it might be expected 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 "counterhedonistic" consumption behavior. For instance, clinically depressed people (vs. nondepressed) 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 people with depression. Therefore, it is possible that people with a tendency toward negative mood states—perhaps because of depression or low well-being—or with a desire to make meaning about their 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 because of 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 that relates to negative beliefs about uncertainty and its implications (Carleton, 2012). Thus, sensitive-content 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 sensitive-content 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 sensitive-content screens or trigger warnings in a social-media context. The present research investigated how participants interact with sensitive-content screens in two ways.

In Study 1, we asked participants how likely they would be to uncover a blurred image if they came across it on Instagram. Because the primary purpose of sensitive-content screens is to allow people who may be vulnerable to mental-health issues to avoid potentially distressing content, we also measured a series of factors covering psychopathology and psychological-vulnerability variables (i.e., depression, anxiety and stress, PTSD symptoms, general well-being, trauma history, centrality of traumatic event to identity, and treatment-seeking behaviors) that we thought might relate to the likelihood that people would uncover these images. We had no specific hypotheses for Study 1, but the previous literature and psychological theory reviewed above suggest that sensitive-content screens would not be effective in deterring the majority of people—including people with mental-health vulnerabilities—from desiring to view or deciding to view negative content. Therefore, our analyses of vulnerability characteristics in Study 1 were exploratory in nature. In Study 2, we presented participants with a mock Instagram photo-viewing task in which participants had the option to click to uncover ("see photo") a single blurred image or select "next photo" to skip uncovering the image. As well as attempting to replicate our main findings, we also included additional measures of well-being in Study 2 to further explore the association between uncovering behavior and these variables.

Study 1

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 1 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 report all measures, conditions, and data exclusions.

Participants. Participants were recruited online through Amazon's Mechanical Turk (MTurk) and received U.S. $2.50. The study was open to respondents above 18 years of age who were located in the United States. Because we wanted to recruit only Instagram users, participants who indicated that they did not use Instagram at the beginning of the survey were screened out.[1] 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, 2018). Therefore, we used a power-based stopping criterion and collected 260 participants after exclusions.

Participants ranged from 20 to 71 years old (M = 36.0 years, SD = 10.69) and were more likely to be female (54.2%, 45% male; 0.4% preferred not to specify sex). Our sample was predominantly White (63.8%); other participants were of African American (14.2%), Asian (7.3%), Latinx (4.2%), or other (5%; e.g., mixed race/biracial) descent; 5.4% of participants specified nationality (e.g., American/United States). The majority of participants (55.8%) reported an income between $45,000 and $140,000 and were predominantly (58.8%) college graduates.[2]

Measures.

Social-media/news-media use. We asked participants to indicate (from a list) which social-media sites they used on a regular basis. We also asked participants to indicate how many days of the past 7 days (*never*, *1 day*, *2 days . . . every day*) and for how many hours each day (*I don't use*, *less than half an hour*, *1 hr*, *2–3 hr*, *4–5 hr*, *> 6 hr*) they used social media.

Instagram sensitive-content screens. Participants were presented with one example of a real Instagram sensitive-content screen (from a pool of six examples) taken from the site (Fig. 1) and were told, "Imagine you are scrolling (i.e., browsing) through Instagram posts and come across the following image." Participants were then asked, "Would you click to uncover this image?" (1 = *definitely no*, 6 = *definitely yes*); "What factors would affect whether you would uncover the image?" (open-box response); and

"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, "Would you turn off the sensitive-content 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).

Depression Anxiety Stress Scales–21. The Depression Anxiety Stress Scales–21 (DASS-21; Lovibond & Lovibond, 1995) is a self-report instrument that measures the severity of depression (present study: α = .95), anxiety (α = .88), and stress (α = .91) in the past week. The scales demonstrate convergent validity with other well-validated measures of depression and anxiety (Antony et al., 1998).

The Scales of General Well-Being short form. The Scales of General Well-Being short form (SGWB-14; Longo et al., 2018) is a brief assessment that measures 14 dimensions of well-being (present study: α = .96). The scales demonstrate convergent validity with other validated measures that tap various aspects of well-being (Longo et al., 2018).

Trauma History Screen. The Trauma History Screen (THS; Carlson et al., 2011) 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 whether 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*).[3] The THS has been validated for use in both clinical and nonclinical populations and has excellent psychometric properties and high reliability (r = .93 for HMS events in clinical samples and rs = .74–.87 for nonclinical samples) and correlates strongly (rs = .73–.76) with more detailed trauma-exposure measures (i.e., the Traumatic Life Events Questionnaire; Carlson et al., 2011).

Posttraumatic Stress Disorder Checklist. The Posttraumatic Stress Disorder Checklist (PCL-5; Bovin et al., 2016) is a self-report measure that corresponds to the symptom criteria for PTSD from the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 2013). Participants were asked to indicate, on a scale from 0 (*not at all*) to 4 (*extremely*), in relation to their most stressful/traumatic event—identified on the THS—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, seven-item version. The Centrality of Events Scale, seven-item version (CES-7; Berntsen & Rubin, 2006) 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, on a scale from 1 (*totally disagree*) to 5 (*totally agree*), a series of questions (e.g., "I feel that this event has become part of my identity"; present study: α = .93). The scale correlates highly with the full 20-item version (r = .96) and displays a robust association with PTSD symptom severity (r = .37; Berntsen & Rubin, 2006).

The Self-Triggering Questionnaire. Using the Self-Triggering Questionnaire (STQ; Bellet et al., 2020), we piped back participants' most stressful/traumatic event text response from the THS and asked whether they had 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 whether 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: people who had self-triggered in reference to either their most stressful/traumatic event or other stressful/traumatic event and people who had not self-triggered at all.[4]

Treatment-seeking behaviors. The treatment-seeking-behaviors questionnaire comprises items from the past-help-seeking section of the General Help-Seeking Questionnaire (Items 2–4; Wilson et al., 2005) and the Actual Help-Seeking Questionnaire (Item 5; Rickwood

& Braithwaite, 1994) and Eisenberg et al. (2009). Participants were asked to indicate whether they have taken any medication, have seen a health professional, or sought help from a source other than a professional—in the past 6 months—to help with a personal problem.

Procedure. Participants were required to pass a Qualtrics V2 Captcha and correctly answer eight of 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 used, and completed the sensitive-content screen task. Next, participants indicated the frequency of their social-media use, completed the THS, indicated PTSD symptoms (PCL-5) and the centrality of the most stressful/traumatic event they had identified during the THS (CES-7), and answered questions on self-triggering. Participants then answered questions about depression, anxiety, and stress symptoms (DASS-21); well-being (SGWB-14); and individual difference characteristics[5] 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.

Results and discussion

Statistical overview. We ran analyses using null-hypothesis significance tests (α = .05) in IBM SPSS (Version 25) and JASP for MacOS (Version 0.13.1). Because our analyses on vulnerability characteristics in Study 1 were exploratory in nature, we opted not to correct for multiple comparisons.

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, fire (38.8%), or a really bad car, boat, train, or airplane accident (31.9%). Furthermore, 24.6% of the sample met criteria for probable PTSD according to the conservative cutoff (sum score > 33; Bovin et al., 2016). For depression, 51.5% of our participants were in the normal range, 25.7% were in the mild-to-moderate range, and 22.7% were in the severe-to-extremely-severe range. For anxiety, 53.8% of our participants were in the normal range, 23.9% were in the mild-to-moderate range, and 22.4% were in the severe-to-extremely-severe range. For stress, 61.2% of our participants were in the normal range, 22.3% were in the mild-to-moderate range, and 16.6% were in the severe-to-extremely-severe range (DASS-21 manual cutoffs). Most participants (85.8%) reported that they used social media every day in the past 7 days (the other responses: 5 days = 5%, 6 days = 4.2%, 3 days = 1.2%, ≤ 2 days = 0.8%) for an hour or more per day (the other responses: 2–3 hr per day = 35%, 1 hr per day = 27.3%, > 6 hr per day = 15.8%, 4–5 hr per day = 12.7%, less than half an hour per day = 9.2%).

Desire to uncover sensitive-content screens and prior experience with sensitive-content screens on Instagram. We asked participants whether they would click to uncover a sensitive-content screen (1 = *definitely no*, 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 1–3) or yes (Responses 4–6); the majority (80%) of participants fell into the yes, or uncover, category.

Aside from asking participants about hypothetically encountering a sensitive-content screen, we also asked them about encounters and interactions with sensitive-content screens in real life. More than 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; 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-content screen feature (so that all photos were not screened when browsing) if they had the option to do so.

Thus, sensitive-content screen 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 1) using the thematic-analysis technique described by Braun and Clarke (2006): Data are coded and labeled according to overarching themes identified across the data set. More than one third of participants (35.8%) indicated that they simply wanted to see the image/picture; of these participants, 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. More than one third (36.2%) of participants indicated they would decide whether to uncover

Table 1. Coded Text Responses to the Question "What Factors Would Affect Whether You Would Uncover the Image?" for Studies 1 and 2

Category	Value
Study 1	
Simply "Would want to see picture" or for more specific reason:	35.8% (93)
Curiosity/intrigue (specific mention)	26.9% (70)
Depend on interest level in the image/at the time	6.2% (16)
Want to see why an image is covered	3.1% (8)
Context provided (e.g., posting account/comments/caption)	36.2% (94)
Type of content expected would influence choice (e.g., nudity, gore, violence.)	14.6% (38)
Physical location/other people present	9.6% (25)
Mood	6.9% (18)
If they believe it would be something they did not want to see/something negative	5.4% (14)
Typically would uncover/would always uncover	2.7% (7)
Typically would not uncover/would never uncover	1.5% (4)
Internet security concerns	1.2% (3)
"No factor would prevent me"/"none"	1.9% (5)
If they could visually guess what the image was	1.2% (3)
Trust in the warning that it is for one's own good	0.8% (2)
Personality traits (e.g., cite general tendency to cope/not cope with sensitive content)	0.8% (2)
Miscellaneous (categories with < two people)/unclassifiable	4.2% (11)
Study 2	
Simply "Wanted to see the picture" or for more specific reason:	72.5% (190)
Curiosity/intrigue (specific mention)	46.2% (121)
Interested in seeing image	4.2% (11)
Want to see why the image is covered	3.8% (10)
Did not want to see something negative	12.6% (33)
Personality traits (e.g., cite general tendency to cope/not cope with sensitive content)	10.7% (28)
Type of content expected would influence choice (e.g., nudity, gore, violence)	8.4% (22)
Uncertainty	2.7% (7)
Did not expect negative content on Instagram/in the study	2.3% (6)
Typically would uncover/would always uncover	1.9% (5)
Context provided (e.g., posting account/comments/caption)	1.5% (4)
Mood	1.5% (4)
If they could visually guess what the image was	1.4% (4)
Not interested	0.8% (2)
Physical location/other people present	0.8% (2)
Miscellaneous (categories with < two people)/unclassifiable	5.0% (13)

Note: Values are percentages with *n*s in parentheses.

depending 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 sensitive-content 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-content screen (14.6%; e.g., nudity or gore); participants' physical surroundings (9.6%), such as 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%).

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 exploratory interest in whether particular psychological vulnerabilities are associated with the desire to uncover sensitive images. We correlated

Table 2. Correlations Between the Desire to Uncover and Continuous Variables

Variable	Study 1	Study 2
Age	−.16**	.004
Social-media use (general)	.05	.06
Instagram use	—	.04
DASS-21		
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-7	−.01	—

Note: Values are correlation coefficients (*r*). DASS-21 = Depression Anxiety Stress Scales–21 (Lovibond & Lovibond, 1995); SGWB-14 = Scales of General Well-Being short form (Longo et al., 2018); WHO-5 = World Health Organization Well-Being Index (Bech et al., 1996); PCL-5 = Posttraumatic Stress Disorder Checklist (Bovin et al., 2016); CES-7 = Centrality of Events Scale, seven-item version (Berntsen & Rubin, 2006). *$p < .05$. **$p < .01$.

participants' reported desire to uncover the Instagram sensitive-content screen as measured on the 6-point scale with our continuous measurements of these variables (Table 2). We also ran a series of χ^2 analyses on the desire to uncover as a dichotomous variable and our categorical dependent variables (Table 3). In terms of participant demographics, we found that age was negatively associated with the desire to uncover, whereas a higher percentage of males (biological sex), compared with 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 scores on the DASS-21 and the Criterion D (negative cognition/mood) and Criterion E (hyperarousal) subscales of the PCL-5 were positively associated with the desire to uncover the sensitive-content screen and that well-being was negatively associated. We also found that people who indicated they self-trigger (yes) compared with people who do not (no) were more likely to fall into the yes/uncover classification.

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 whether any variables were correlated at more than .70 with one another (as per our preregistration). No variables were correlated at more than .70. We also ran a standard linear regression, using the 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 variance inflation factor value of more than 10, or high variance proportions on the same eigenvalue (Field, 2005), which indicates no issue of multicollinearity among our predictors. In our main analysis, we entered all of our significant vulnerability predictors (DASS-21 total, well-being, PCL-5 total, and self-triggering, yes/no) in a single step (Table 3). We found that our model significantly predicted the desire to uncover (or not uncover) the sensitive-content screen, $\chi^2(4) = 15.24$, $p = .004$ (R^2: Hosmer and Lemeshow = .06, Cox and Snell = .06, Nagelkerke = .09). Well-being and the tendency to self-trigger with a reminder of participants' 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-content screen increased and that people with a tendency to self-trigger (vs. not) were more likely to indicate a desire to uncover the image (and thus fall into the uncover category).

Taken together, our Study 1 findings demonstrate that sensitive-content 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, in a set of exploratory analyses, we found that various psychological vulnerability factors are associated with the desire to approach sensitive content. Therefore, it is likely that sensitive-content screens are even less effective in encouraging avoidance behaviors for vulnerable users (e.g., people with mental-health vulnerabilities; especially people with lower well-being) than nonvulnerable users.

Study 2

In Study 2, we aimed to replicate and extend the findings of Study 1. In Study 1, we measured the intent to uncover sensitive-content 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 sensitive-content screen). However, although intentions *generally* map onto future behavior ($r = .53$; Sheeran, 2005), intentions may be inconsistent with actual behavior—the intention–behavior gap

Table 3. Desire to Uncover or Keep Covered by Key Categorical Dependent Variables

Variable	Keep covered	Uncover	$\chi^2(df)$	*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)				—	—			
Posttraumatic stress disorder 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 past 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 past 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)				—	—	—	—	—

Note: Values are percentages with *ns* in parentheses unless otherwise noted. For analyses of other categorical variables not mentioned here (i.e., gender, household income, highest level of education), see https://osf.io/acpeu/.

(Sheeran & Webb, 2016). Therefore, in Study 2, we presented participants with a mock Instagram-photo-viewing task in which 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 whether participants' hypothetical responses in Study 1 mapped onto a behavioral task that more closely matches Instagram. In addition to attempting to replicate our main findings, we included additional measures of well-being. There are many ways of defining and therefore measuring well-being (Dodge et al., 2012). Whereas the SGWB-14 focuses on 14 aspects of well-being (happiness, vitality, calmness, optimism, involvement, self-awareness, self-acceptance, self-worth, competence, development, purpose, significance, self-congruence, and connection), the World Health Organization Well-Being Index (WHO-5; Bech et al., 1996) focuses on well-being as a single construct: positive well-being as a signifier of mental health and absence of mental illness (e.g., depression; Krieger et al., 2014).

Given the findings of Study 1, we predicted that a majority of participants (≈80%) would click to uncover the sensitive-content screen. We further predicted that lower levels of well-being and higher levels of PTSD symptoms and 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-content screen. Finally, because self-triggering is associated with PTSD severity (Bellet et al., 2020), we predicted that the relationship between PTSD symptoms 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 participants who endorsed self-triggering versus those who did not.

Method

We preregistered this study (https://osf.io/8n7er). Study 2 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 report all measures, conditions, and data exclusions.

Participants. Participants were recruited online through MTurk. Participants received a payment of U.S. $2.00. As in Study 1, the study was open to respondents above 18 years of age who were located in the United States, and participants who indicated at the beginning of the survey that they did not use Instagram were screened out. We excluded one participant who failed all three embedded attention checks (Berinsky et al., 2021; 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 preregistered requirement because we were seeking to understand uncovering behaviors as those behaviors typically occur on Instagram). In total, we collected 262 participants after exclusions.

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

Measures.

Participants completed a mock Instagram task where they viewed a set offive neutral and five positive Nencki Affective Picture System (NAPS) photos (Marchewka et al., 2014)—randomly selected from one of 14 sets of 10 images—in a random order; there was 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 (Fig. 2). Participants then viewed a single sensitive-content 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,"[6] 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 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*); and (c) "Would you turn off the sensitive-content 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).

As in Study 1, participants completed measures of social-media use and Instagram specifically; depression, anxiety, and stress symptoms (DASS-21; Study 2: depression α = .93; anxiety α = .89; stress α = .89); well-being (SGWB-14; α = .94); the THS; and PTSD symptoms (PCL-5; α = .96). We also measured self-triggering by piping back participants' most stressful/traumatic event text response from the THS and asked whether they had ever self-triggered with reminders of this event (yes/no; i.e.,

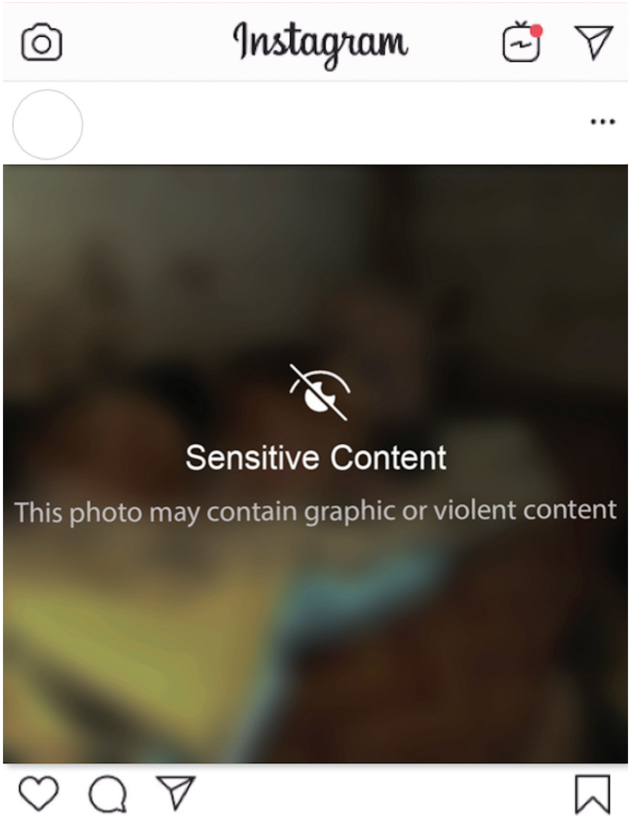


Fig. 2. Example of a Nencki Affective Picture System (NAPS) photo modified to look like an image with a sensitive-content overlay used in Study 2.

from the STQ). In addition, participants completed the WHO-5 (Bech et al., 1996). Participants rated how five statements (e.g., "I have felt calm and relaxed") applied to them over the past 2 weeks (0 = *at no time*, 5 = *all of the time*). Total scores (0–25) are multiplied by 4 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 eight of 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 1, 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). Next, participants completed the mock Instagram task and related questions about sensitive-content screens, followed by the THS, PTSD symptoms (PCL-5), the single self-triggering question, and coping questionnaires[7] in a randomized order. Participants then answered questions about depression, anxiety, and stress symptoms (DASS-21); well-being (the SGWB-14 and the WHO-5); and individual difference characteristics[8] in a randomized order. Participants were then asked to indicate whether they behaved as they normally would on Instagram (yes/no), and if yes, they were asked to explain how they behaved differently and why. They were also asked whether they left the task for any period of time (if no, when and for how long) and whether they had any technical issues. Participants were then fully debriefed.

Results

Statistical overview. We ran analyses using null-hypothesis significance tests (α = .05) in IBM SPSS (Version 25) and JASP for MacOS (Version 0.13.1). In cases in which data were missing, we used subscale-level mean substitution. One person missed three items on the SGWB-14.

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, fire (44.7%), or a really bad car, boat, train, or airplane accident (28.2%). Furthermore, 21.4% of the sample met criteria for a likely PTSD diagnosis according to the conservative cutoff on the PCL-5 (> 33; Bovin et al., 2016). For depression, 54.6% of our participants were in the normal range, 29.7% were in the mild-to-moderate range, and 18.7% were in the severe-to-extremely-severe range. For anxiety, 59.2% of our participants were in the normal range, 19.8% were in the mild-to-moderate range, and 21% were in the severe-to-extremely-severe range. For stress, 61.8% of our participants were in the normal range, 26% were in the mild-to-moderate range, and 12.3% were in the severe-to-extremely-severe range. The majority of participants (87.8%) reported that they used social media every day in the past 7 days (the other responses: 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 hr per day = 39.7%, 1 hr = 21.8%, > 6 hr = 17.6%, 4–5 hr = 11.8%, less than half an hour = 9.2%). Most participants (51.1%) had used Instagram every day in the previous 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 the previous 7 days = 0.8%).

Table 4. Logistic Regression Results for Predicting Uncovering Desire and Vulnerability Characteristics

Predictor	β	exp *b*	95% CI for exp *b*	*p*
Included				
Constant	4.00 (1.16)	54.39	—	< .001
DASS-21 total	–0.01 (0.02)	0.99	[0.96, 1.03]	.621
Well-being	–0.04 (0.02)	0.96	[0.93, 1.00]	.028
PCL-5 total	0.01 (0.01)	1.01	[0.98, 1.03]	.605
Self-triggering (yes/no)	–0.94 (0.40)	0.39	[0.18, 0.85]	.018
Included				
Constant	1.68 (1.53)	5.75		.274

Note: Values in parentheses are standard errors. exp *b* = the exponential value of *b* or odds ratio, which is the predicted change in odds for a unit increase in the predictor; DASS-21 = Depression Anxiety Stress Scales–21 (Lovibond & Lovibond, 1995); PCL-5 = Posttraumatic Stress Disorder Checklist (Bovin et al., 2016).

Decision to uncover sensitive-content screens and prior experience with sensitive-content 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 the findings of Study 1, the majority of participants (84.7%) fell into the uncover category. Again, as in Study 1, we also asked our participants about encounters and interactions with sensitive-content screens in real life. More than half of our participants (64.5%) indicated that they had previously seen a sensitive-content screen on Instagram. Participants who said they had seen the screens on Instagram reported that they almost always (*M* = 4.21, *SD* = 1.60; 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-content screen feature (so that all photos were not screened when browsing) if they had the option to do so.

Like Study 1, we coded participants' text responses to the question "Why did you or did you not uncover the screened image?" (Table 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 on the basis of a general tendency/personality trait to cope with or not cope with distressing content (10.7%). The type of content that might be behind the screen (e.g., nudity or gore) was mentioned by 8.4% of participants. Although 36.2% of participants in Study 1 mentioned contextual factors that may accompany a photo in real life (e.g., posting account, caption, comments), only 1.5% of participants mentioned it in Study 2.

Taken together, Study 2 confirms Study 1, that 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. Although curiosity remains a popular reason for approaching muted content, participants in Study 2 were more likely to cite not wishing to see negative content and personality traits and were 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 psychological vulnerabilities? Unlike in Study 1, we did not find any significant associations between psychological vulnerabilities and the decision to uncover the screened image in Study 2 (Tables 2–4). One potential statistical explanation for this pattern of results is the relatively high base rate of people who chose to uncover the image (*n* = 220) relative to people who avoided the image (*n* = 40), which may have led to variance heteroscedasticity. However, Levene's test for equality of variances—comparing people who uncovered and those who did not in Study 2—did not reveal any significant violations of homogeneity for any of our continuous dependent variables. Moreover, as noted earlier, because our analyses of vulnerability characteristics in Study 1 were exploratory in nature, we opted not to correct for multiple comparisons. Therefore, it is possible that the significant associations found in Study 1 can be explained by an inflated Type I error rate. We discuss further potential explanations below.

General Discussion

Instagram claims that sensitive-content 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 1) or made a choice (Study 2) 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 1, we found that the desire to uncover a muted image was associated with a number of vulnerability factors, including depression, well-being, and PTSD symptoms. Although we did not replicate this pattern in Study 2 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, sensitive-content 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 toward it increases (e.g., Ringold, 2002). Our results also fit with 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 2, only 15% of participants took it.

We also note that our results from Study 2 could also be explained by boredom-induced novelty seeking. Boredom creates an emotional state that causes people to seek novel counterhedonic 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 2 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 2 viewed only 10 photos (a task lasting from around 30 s to 1 min), it seems unlikely boredom would be a pertinent factor. We also note that in real-world conditions, sensitive-content 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.

Whereas our findings from Study 1 and 2 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 1, we found that certain vulnerability characteristics (e.g., poorer ratings of general well-being and higher ratings of depression) were associated with a greater desire to uncover screened content. These findings fit with data that show people 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 sensitive-content 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. In addition, 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 to ascertain meaning from doing so.

However, in Study 2, 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 1), 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 nonvulnerable users are not deterred from approaching distressing content.

Why else may we have found differences between Study 1 and Study 2? 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%), $\chi^2(1) = 2.01$, $p = .156$, 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 1 were different from the types of participants who actually uncovered the photo in Study 2. To get at this possibility, we compared participants in Study 1 and Study 2 on our key vulnerability factors of interest (i.e., variables that were significantly associated with uncovering behavior in Study 1). We found no significant differences (*p*s = .061–.922, *d*s = 0.01–0.16, ϕs = .02–.07).[9] Another possibility involving individual differences is that our dichotomous variable in Study 2 was less sensitive to our vulnerability factors than our ordinal variable in Study 1.

A second explanation lies within participants' qualitative responses about the decision to approach or avoid muted content. Specifically, participants in Study 1 seemed to place a higher importance on contextualizing the Instagram post and considering elements such as posting account, captions, and comments on the post as an important factor when deciding whether they would uncover the photo. In Study 1, these reasons were listed more than one third of the time, whereas they were minimally mentioned by participants in Study 2, who instead placed a high importance on feelings of curiosity. It is possible that this difference occurred because Study 1 asked a hypothetical question and therefore participants may have been more likely to contextualize the sensitive-content screen in their own imagination (e.g., the type of account that may have posted it) or think about past experiences with sensitive-content 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 1, we captured people's broader pattern of approach behavior (across different scenarios). In contrast, when we presented a sensitive-content 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 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 whether these features influence the desire to uncover the screened images. Furthermore, it may also be necessary to investigate whether alternative warning messages on the sensitive-content 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 vulnerable populations are less (as found in Study 1) or more deterred by sensitive-content 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; Schönbrodt & Perugini, 2013). Therefore, we believed this sample size was adequate. For 95% power for a small effect (e.g., r = .10), a sample of roughly double this size (470 participants) would have been required (Schönbrodt & Perugini, 2013). However, this sample size was not feasible because of resource constraints (Lakens, 2022).

A fourth explanation involves the time we collected data. Study 1 was collected in December 2019, before COVID-19 became a global pandemic, and Study 2 was collected in December 2020. Given that the COVID-19 pandemic has ravaged all areas of human life, including exacerbating mental-health issues (e.g., Bridgland, Moeck, et al., 2021), it is possible that the pandemic's impact had some unmeasurable impact on the way our mental-health variables interacted with uncovering behavior.

A fifth explanation is that there is no association between our vulnerability measures and uncovering intentions or behavior and the results of Study 1 were simply Type I errors. Indeed, because these analyses in Study 1 were exploratory, we opted not to correct for multiple comparisons, which means that the Type I error rate was likely inflated.

Our study has several limitations. First, although we found that sensitive-content 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 sensitive-content 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 toward negative material (Bellet et al., 2020; Boysen et al., 2021; Bridgland et al., 2019; Sanson et al., 2019). Moreover, because sensitive-content 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 (vs. unscreened). Regardless of the exact mechanism, an essential next step in this area of research is to assess how sensitive-content screens affect or do not affect emotional reactions to negative images.

Second, our open-text responses showed that contextual elements (e.g., the posting account name, captions, and comments) are likely an important factor in uncovering behavior. Because 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-content screen on Instagram. It is not uncommon for photos with sensitive-content screens to be posted with ambiguous or no clear captions/context.

Third, sensitive-content screens—and trigger warnings—have historically been primarily intended for people with mental-health vulnerabilities (e.g., PTSD, exposure to trauma). Therefore, it is possible that our results would have been different had we specifically recruited and powered our sample for particular 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 et al., 2013).

Fourth, although 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 choose 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. nondepressed) 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 these limitations in mind, we believe our findings from Study 1 and 2 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 articles that have examined the effect of trigger warnings on avoidance behaviors. However, neither of these articles focused on approach or avoidance behaviors as a main aim. In addition, 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, and Buzzfeed. Furthermore, no research has examined whether vulnerability factors (e.g., well-being, depressive symptoms) relate to the rates of approaching or avoiding content accompanied by trigger-warning messages. Therefore, we believe our key finding that sensitive-content 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 nonvulnerable 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.

Transparency

Action Editor: Kelsie T. Forbush
Editor: Jennifer L. Tackett
Author Contributions

V. M. E. Bridgland developed the study design with the guidance of M. K. T. Takarangi. V. M. E. Bridgland collected the data, performed the data analysis and interpretation, and drafted the manuscript. M. K. T. Takarangi and B. W. Bellet made critical revisions to the manuscript. All of the authors approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Open Practices

All data and materials have been made publicly available via OSF and can be accessed at https://osf.io/rj987. The design and analysis plans for the experiments were preregistered at OSF (Study 1: https://osf.io/m6d9g; Study 2: https://osf.io/8n7er). This article has received badges Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at https://osf.io/s6fdz.

ORCID iDs

Benjamin W. Bellet https://orcid.org/0000-0002-4338-3393
Melanie K. T. Takarangi https://orcid.org/0000-0002-6006-8045

Notes

1. One hundred and fifty-five participants completed both the Instagram-screen questions and the news-filter questions.
2. For full demographics, see https://osf.io/acpeu/.
3. These data are not reported here. See https://osf.io/rj987/.
4. Analyses that involve self-triggering frequency, methods, and motives are not reported here. See https://osf.io/acpeu/.
5. These data were secondary to our main research aims in this study (i.e., which focused on how vulnerable users interact with sensitive-content screens) and are not reported here. See https://osf.io/mjrq8/.

6. Participants were randomly assigned to see a “see photo” or “uncover photo” button—however, rates of selecting “Next Photo” did not significantly differ per button type, $\chi^2(1) = 0.47$, $p = .492$; thus, we collapsed our analyses across button type.

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

8. See https://osf.io/mjrq8/.

9. 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 et al., 2020) than participants in Study 2 (M = 17.52, SD = 5.31; d = 0.23). See https://osf.io/mjrq8/.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.).

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

Bech, P., Gudex, C., & Johansen, K. S. (1996). The WHO (Ten) well-being index: Validation in diabetes. *Psychotherapy and Psychosomatics*, *65*(4), 183–190. https://doi.org/10.1159/000289073

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

Bench, S. W., & Lench, H. C. (2019). Boredom as a seeking state: Boredom prompts the pursuit of novel (even negative) experiences. *Emotion*, *19*(2), 242–254. https://doi.org/10.1037/emo0000433

Berinsky, A. J., Margolis, M. F., Sances, M. W., & Warshaw, C. (2021). 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., & 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

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. (2016). Event centrality prospectively predicts PTSD symptoms. *Anxiety, Stress, & Coping*, *29*(5), 533–541. https://doi.org/10.1080/10615806.2015.1080822

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

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

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

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

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* [Manuscript submitted for publication.] College of Education, Psychology & Social Work, Flinders University. https://osf.io/7n85z/

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., Hutchison, N. P., & Takarangi, M. K. T. (2021). Why the COVID-19 pandemic is a traumatic stressor. *PLOS ONE*, *16*(1), Article e0240146. https://doi.org/10.1101/2020.09.22.307637

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

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.

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

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., Burling, T. A., & Spain, D. A. (2011). Development and validation of a brief self-report measure of trauma exposure: The trauma history screen. *Psychological Assessment*, *23*(2), 463–477. https://doi.org/10.1037/a0022294

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

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

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). Sage.

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

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

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

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

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*, Article 109836. https://doi.org/10.1016/j.paid.2020.109836

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

Lakens, D. (2022). *Sample size justification*. *Collabra: Psychology* (), *8*(1). Article 33267. https://doi.org/10.1525/collabra.33267

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

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

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. *Behaviour Research and Therapy*, *33*(3), 335–343. https://doi.org/10.1016/0005-7967(94)00075-u

Marchewka, A., Z˙urawski, 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*, 596–610. https://doi.org/10.3758/s13428-013-0379-1

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

Mosseri, A. (2019a, February 7). Instagram policy changes on self-harm related content - Protecting vulnerable users. *Instagram Blog*. 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*. https://about.instagram.com/blog/announcements/more-steps-to-keep-instagram-users-safe

Ng, A. H., Kermani, M. S., & Lalonde, R. N. (2021). Cultural differences in psychological reactance: Responding to social media censorship. *Current Psychology*, *40*(6), 2804–2813. https://doi.org/10.1007/s12144-019-00213-0

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

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

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

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

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

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. (2005). 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-behavior gap. *Social and Personality Psychology Compass*, *10*(9), 503–518. https://doi.org/10.1111/spc3.12265

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

Wakefield, M., Terry-McElrath, Y., Emery, S., Saffer, H., Chaloupka, F. J., Szczypka, G., Flay, B., O'Malley, P. M., & Johnston, L. D. (2006). Effect of televised, tobacco company–funded smoking prevention advertising on youth smoking-related beliefs, intentions, and behavior. *American Journal of Public Health*, *96*(12), 2154–2160. https://doi.org/10.2105/ajph.2005.083352

Wilson, C. J., Deane, F. P., Ciarrochi, J., & Rickwood, D. (2005). *General Help Seeking Questionnaire* [Database record]. APA PsycTests. https://doi.org/10.1037/t42876-000

Zillmann, D. (1988). Mood management through communication choices. *American Behavioral Scientist*, *31*(3), 327–340. https://doi.org/10.1177/000276488031003005