Journal of Experimental Psychology: Applied

Sending Signals: Trigger Warnings and Safe Space Notifications

Samuel Pratt, Payton J. Jones, Victoria M. E. Bridgland, Benjamin W. Bellet, and Richard J. McNally

Online First Publication, July 3, 2025. https://dx.doi.org/10.1037/xap0000541

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: Applied

ISSN: 1076-898X

https://doi.org/10.1037/xap0000541

Sending Signals: Trigger Warnings and Safe Space Notifications

Samuel Pratt[1], Payton J. Jones[1], Victoria M. E. Bridgland[2], Benjamin W. Bellet[3], and Richard J. McNally[4]

[1] Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill

[2] College of Education, Psychology, and Social Work, Flinders University

[3] Massachusetts Mental Health Center, Boston, Massachusetts, United States

[4] Department of Psychology, Harvard University

Trigger warnings and safe space notifications are common in higher education. Although researchers have evaluated these practices as mental health tools, little attention has been paid to the interpersonal signals they send. In this experiment conducted in fall 2024, we examined how trigger warnings and safe space notifications shape students' perceptions of instructors and the classroom environment. We randomly assigned 738 American undergraduate students to view videos of instructors delivering a brief lecture on trauma, preceded by the instructor providing a trigger warning, a safe space notification, both, or neither. Participants rated the instructor's epistemic trustworthiness, concern for student well-being, political orientation, and Left-Wing Authoritarianism scale, as well as their own feelings of psychological safety and willingness to discuss controversial topics in the classroom. Analyses using Bayes Factors provided substantial evidence that trigger warnings had no overall impact on students' perceptions. In contrast, safe space notifications increased students' feelings of psychological safety and willingness to discuss controversial topics. Safe spaces also increased perceptions of instructors as caring and trustworthy but signaled that instructors were liberal and left-wing authoritarian, including the subscale measuring support for top-down censorship. Implications for the use of trigger warnings and safe spaces in educational contexts are discussed.

Public Significance Statement

This study suggests that giving a trigger warning before potentially distressing course material does not improve students' perceptions of the instructor or classroom environment. In contrast, telling students that the classroom is a safe space increases their feelings of psychological safety and willingness to discuss controversial topics while signaling that the instructor is caring and trustworthy. However, safe space notifications also make instructors seem more liberal and supportive of censorship. These findings highlight the need to weigh the costs and benefits of these practices in context.

Keywords: trigger warning, safe spaces, student perceptions, signals

Supplemental materials: https://doi.org/10.1037/xap0000541.supp

Melody Wiseheart served as action editor.

Samuel Pratt https://orcid.org/0000-0002-5816-8397

Payton J. Jones https://orcid.org/0000-0001-6513-8498

Victoria M. E. Bridgland https://orcid.org/0000-0002-8865-8426

Benjamin W. Bellet https://orcid.org/0000-0002-4338-3393

Richard J. McNally https://orcid.org/0000-0002-5228-8777

Samuel Pratt and Payton J. Jones are contributed equally to this work. All Supplemental Materials associated with this project are available on the Open Science Framework (https://osf.io/9sbnt/?8fd741f038574d3783d8c94f1724c04a). The authors thank the Institute for Quantitative Social Science at Harvard University for supporting and funding through their Extraordinary Claims and Extraordinary Evidence partnership program (Grant 204.03, awarded to Richard J. McNally). Assistance on the analytic plan was provided by data science specialist Joshua Cetron at the Institute for Quantitative Social Science, Harvard University. The authors thank Donna M. McNally for her assistance on this project.

Samuel Pratt played a lead role in writing the original draft and an equal role in funding acquisition, methodology, and reviewing and editing. Payton J. Jones played a lead role in conceptualization, formal analysis, and writing the original draft and an equal role in funding acquisition, methodology, and reviewing and editing. Victoria M. E. Bridgland played an equal role in funding acquisition, methodology, and reviewing and editing. Benjamin W. Bellet played an equal role in funding acquisition, methodology, and reviewing and editing. Richard J. McNally played a lead role in funding acquisition and an equal role in methodology and reviewing and editing.

Correspondence concerning this article should be addressed to Samuel Pratt, Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, 315 Davie Hall, CB 3150, Chapel Hill, NC 27599, United States. Email: sampratt@g.ucla.edu

Trigger warnings and safe space notifications are common practices in higher education. Trigger warnings—originally intended to warn individuals with posttraumatic stress disorder (PTSD) about potential reminders of their trauma—are now commonly used to alert students to a range of potentially distressing course topics (e.g., sexual assault, racism; Boysen, 2017). Similarly, some professors designate the classroom as a "safe space," which the Merriam-Webster dictionary defines as "a place (as on a college campus) intended to be free of bias, conflict, criticism, or potentially threatening actions, ideas, or conversations" (Merriam-Webster, n.d.). The use of safe spaces ranges from brief verbal notifications to designated physical spaces where students are encouraged to feel safe from identity-based prejudice and discrimination (Witherup & Verrecchia, 2020). Although there is limited direct data on how frequently instructors use safe space notifications, survey research suggests that the broader concept of safe spaces is widely recognized and discussed in educational settings (Witherup & Verrecchia, 2020).

Whether such practices are beneficial to students is the subject of widespread debate within higher education. Proponents argue that warnings and safe spaces are necessary to support the emotional well-being of students and foster a positive learning environment (George & Hovey, 2020). Critics argue that these practices restrict academic freedom (Engle, 2023; George & Hovey, 2020), contribute to a culture of fragility (Filipovic, 2014; Lukianoff & Haidt, 2018), and are ineffective as mental health resources (McNally, 2016; Suk Gersen, 2021). In 2016, the Dean of Students at the University of Chicago directly addressed the controversy surrounding trigger warnings and safe spaces, stating that "Our commitment to academic freedom means that we do not support so-called trigger warnings, we do not cancel invited speakers because their topics might prove controversial, and we do not condone the creation of intellectual 'safe spaces' where individuals can retreat from ideas and perspectives at odds with their own" (Ellison, 2016).

Public opinion on these issues is also divided. In a 2020 study, 46.5% of students agreed that "safe spaces encourage a positive learning environment," while 40.8% believed that safe spaces actively detract from learning (attitudes were comparable among faculty; Witherup & Verrecchia, 2020). Among these students, liberals were 40% more likely than conservatives to hold positive attitudes toward safe spaces, and women were six times more likely than men. Similarly, studies suggest that between 39% and 51% of college professors use warnings in the classroom (Boysen & Prieto, 2018; Kamenetz, 2016), and estimates of the prevalence of students who support the use of trigger warnings are as low as 20% in some studies (Burch et al., 2018) and as high as 80% or more in other studies (Bellet et al., 2018; Celniker et al., 2022; Sevincer et al., 2024). Ultimately, whether these practices help or harm is an empirical question, and this question has become a recent focus of psychological research (Bridgland et al., 2024; Pratt, Bellet, et al., 2025).

Advocates of trigger warnings and safe spaces initially promoted these policies because they assumed these practices would aid vulnerable students (Stokes, 2014). However, at least in the case of trigger warnings, research suggests they are ineffective or counterproductive (Bridgland et al., 2024). A meta-analysis of all empirical studies of trigger warnings up to March 2022 found that warnings increase anxiety prior to distressing content and do not reduce anxiety in response to "triggering" material (Bridgland et al., 2024). Other research finds that warnings increase the belief that trauma is central to one's identity (Jones et al., 2020)—a countertherapeutic belief linked to stronger symptoms of PTSD (e.g., Berntsen & Rubin, 2006; Robinaugh & McNally, 2011). There is comparatively very little research on safe space notifications. One study found that students infer from the presence of safe spaces that there are high levels of prejudice on their campus (Gainsburg & Earl, 2022). Despite mounting evidence for the inefficacy of trigger warnings and a lack of research on safe space notifications, both practices continue.

Why do so many offer warnings if they fail to work as intended? One possibility is that the public is simply misinformed about the purported clinical benefits of these practices. For example, one study found that 42% of students believed that trigger warnings reduce negative emotional reactions to distressing content (Sevincer et al., 2024), a belief refuted by empirical evidence (Bridgland et al., 2024). However, another possibility is that trigger warnings and safe spaces have benefits or costs that transcend any clinical effects. Some educators and scientists have argued that even if trigger warnings or safe spaces fail in their original intent—to help students effectively regulate their emotions—these practices might be useful as an interpersonal signal to students, demonstrating solicitude, trust, and respect for students' personal struggles (Bloom, 2023; Sastre, 2017). Interestingly, critics of warnings and safe spaces often agree about the intent of such signals but have decried them as mere "virtue signaling"—that is, highly public behaviors lacking the sincerity or personal sacrifice characteristic of truly moral actions (e.g., Tosi & Warmke, 2016). Critics have also accused these practices of signaling affiliation with a specific political orientation or ideological group, which is traditionally discouraged in educational settings (Suk Gersen, 2021). No research has investigated either set of claims.

The Present Study

We investigated whether trigger warnings and safe space notifications alter perceptions of the person offering them. We randomly assigned American undergraduate students to view brief videos of an instructor lecturing on the topic of psychological trauma, preceded by a trigger warning, a safe space notification, both, or neither. Participants then rated the instructor and classroom environment along various dimensions. We also tested whether any effects of trigger warnings and/or safe space notifications on students' perceptions were moderated by theoretically relevant participant characteristics, including political ideology and trauma history.

Method

Transparency and Openness

This study was approved by the Harvard University Committee on the Use of Human Subjects (IRB24-0559). All data were collected in fall 2024. The preregistered study design and analysis plan along with all data, materials, and code are publicly available on the Open Science Framework (https://osf.io/9sbnt/?8fd741f038574d3783d8c94f1724c04a; Pratt, Jones, et al., 2025). All analyses were conducted in the R software environment (Version 4.4.2;

R Core Team, 2024). We report how we determined our sample size and have reported all measures, conditions, and data exclusions.

Participants

Participants were American undergraduate college students recruited online via Prolific.[1] Participants were 18 years or older and fluent in English. Participants were prescreened within Prolific to provide a roughly equal number of liberals, moderates, and conservatives. To determine the appropriate sample size for our study, we conducted a series of simulations using R (the code is available in the Supplemental Materials). The simulations modeled two different scenarios: a simple case examining the main effects of trigger warnings and safe space notifications, and a more complex case testing moderation effects, where individual characteristics (e.g., political orientation) may interact with the presence of warnings. In the simple case, we simulated data with a small effect size (Cohen's $d = 0.2$) or no effect ($d = 0.0$) and tested Bayesian regression models across sample sizes ranging from 400 to 1,500 participants. The results indicated that to correctly detect a true positive effect in 80% of cases while maintaining a false positive rate below 5%, we would need approximately 1,350 total observations—or, in our case, 675 participants providing two observations each. For the moderation analysis, the simulations revealed that even for small interaction effects ($f^2 = 0.02$), 675 participants would be sufficient to detect true positive effects in greater than 80% of cases with a false positive rate below 5%. To account for potential attrition, we collected data from 800 participants. Based on our preregistered analysis plan, we excluded participants if they failed an English fluency verifier, were identified by Qualtrics as a bot, failed an attention check (i.e., a single question asking them to recall the topic of the lecture video, PTSD, among three incorrect topics, linear algebra; the French language; postmodern economic theory), or provided a valid reason for their data to be excluded in an open-ended response at the end of the survey. This left a final sample of 738 participants (n = 1,439 total observations).[2]

Procedure

Participants were invited to participate in a study on "Evaluating Instructors" based on brief videotapes of the instructors delivering a lecture. After consenting to participate, passing all screening checks, and completing a brief battery of standard demographic questions, participants were randomly assigned to watch two short lecture videos.

Each lecture video included an instructor delivering a segment of an introductory psychology lecture about PTSD (the videos ranged from 2:14 min to 3:53 min and are available in the online materials). We experimentally manipulated whether the lecture was preceded by the instructor providing a trigger warning, a safe space notification, a combination of both, or neither (control). We also varied the lecturer (one male, one female) and the wording of the lecture (two similar scripts), resulting in a total set of 16 videos. The first video seen by each participant was chosen randomly. The second video was always a distinct lecturer with a distinct script from the first video, and a nonidentical condition (any of the other three conditions, chosen randomly). The full wording of each experimental condition is listed below. The full wording of the lecture scripts is available in the online materials.

Trigger Warning

"Before we begin today's lecture, I want to issue a trigger warning. The content we're about to cover includes discussions about interpersonal trauma, such as sexual violence. This content may evoke a distressing emotional reaction for some people, particularly those with a history of trauma."

Safe Space Notification

"Before we begin today's lecture, I want to emphasize to everyone that this classroom is a safe space. If at any point the material becomes too distressing, please feel free to disengage as necessary. It's essential to prioritize your emotional safety."

Trigger Warning + Safe Space Notification

"Before we begin today's lecture, I want to issue a trigger warning. The content we're about to cover includes discussions about interpersonal trauma, such as sexual violence. This content may evoke a distressing emotional reaction for some people, particularly those with a history of trauma. Also, I want to emphasize to everyone that this classroom is a safe space. If at any point the material becomes too distressing, please feel free to disengage as necessary. It's essential to prioritize your emotional safety."

Control

"Before we begin today's lecture, I'd like to mention that the material we will be covering today concerns life experiences and interpersonal variables. Okay, let's get started."

After viewing each lecture segment, participants rated the instructor and the classroom environment by responding to the following measures. After completing the survey, participants were debriefed about the purpose of the study and the reasons for incomplete disclosure of study aims.

Measures

Epistemic Trustworthiness

Epistemic trustworthiness refers to the degree that a person trusts the truthfulness of a given source. We measured the epistemic trustworthiness of the instructor with the Muenster Epistemic Trustworthiness Inventory (METI), which contains 14 items on 7-point Likert scale across three domains: expertise, integrity, and benevolence (Hendriks et al., 2015). Participants rated the extent to which the instructor fit pairs of opposing constructs relating to

[1] To ensure that we collected a sample of undergraduate students, we used Prolific internal demographic targeting to select for current undergraduate students. Additionally, we embedded a question at the beginning of the survey asking participants whether they were currently an undergraduate at an American college or university. Participants who indicated "no" were screened out.

[2] There was a coding error in the survey software for saving the specific combination of male lecturer, first lecture version, and control condition. The result was that for participants who saw this specific combination in the first video, the second video was not properly randomized. In some cases, these participants saw a repeat instructor or lecture in the second video (n = 37). We omitted the data for the second video for all such cases.

expertise (e.g., *competent* vs. *incompetent*), integrity (e.g., *sincere* vs. *insincere*), and benevolence (e.g., *moral* vs. *immoral*).

Concern for Student Well-Being

We measured perceptions of the instructor's concern for student well-being with a single face-valid item "The instructor cares deeply about the well-being of his or her students" rated on a 7-point Likert scale from *strongly disagree* to *strongly agree*.

Instructor Political Orientation

Perceived political orientation of the instructor was measured on a 6-point Likert scale from *very liberal* to *very conservative*. The scale midpoint was intentionally removed to more effectively binarize this variable in subgroup analyses.

Instructor Left-Wing Authoritarianism Scale

Left-Wing Authoritarianism scale (LWA) refers to a set of political beliefs characterized by the desire to forcefully overturn existing hierarchies, prejudice and intolerance toward ideologically dissimilar others, and willingness to suppress dissenting viewpoints (Costello et al., 2022). Perceptions of instructor LWA were measured using a 13-item scale (LWA-13; Costello & Patrick, 2023) consisting of three subscales: antihierarchical aggression (e.g., "We need to replace the established order by any means necessary"), anticonventionalism (e.g., "The 'old-fashioned ways' and 'old-fashioned values' need to be abolished"), and top-down censorship (TDC) (e.g., "Getting rid of inequality is more important than protecting the so-called 'right' to free speech"). Participants estimated the extent to which the instructor endorsed each of the 13 statements on a 7-point Likert scale from *the instructor strongly disagrees* to *the instructor strongly agrees*.

After viewing both videos and rating the instructor and the classroom environment, participants responded to the following battery of measures assessing their own beliefs and background.

Psychological Safety

Psychological safety refers to the feeling that one is safe to take risks, make mistakes, and express oneself in a shared environment without negative repercussions. We measured participant perceptions of psychological safety by using Edmondson's (1999) Psychological Safety Scale but adapted the wording to refer to the classroom rather than the workplace. Participants imagined that they were a student in the instructor's classroom and rated their agreement with seven statements (e.g., "If you make a mistake in this class, it is often held against you") on a 5-point Likert scale from *strongly disagree* to *strongly agree*.

Reluctance to Discuss Controversial Issues

Reluctance to discuss controversial topics in the classroom was measured by an adapted version of the 4-item scale present across the multiple iterations of the Campus Expression Survey (Zhou & Barbaro, 2023). Participants imagined discussing six different topics (politics, race or ethnicity, religion, sexual orientation, gender roles, trans identity) in a relevant discussion in the instructor's class and rated on a 4-point Likert scale how comfortable or reluctant they would be to give their honest thoughts, ideas, and questions on each topic. Next, participants were asked, "If you were to give YOUR HONEST THOUGHTS, IDEAS, AND QUESTIONS on one of the issues during a class discussion, would you be concerned that any of the following would happen with regard to your instructor?" Participants selected all that apply from a dropdown list (e.g., "The instructor would give me a lower grade because of what I said"). For analysis, we created an aggregate variable tracking if participants selected any of these options (but omitted the open responses category "Other concerns of consequences," which many students checked but indicated in the text field they had no concerns).

Trigger Warnings Attitudes Assessment

The trigger warnings attitudes assessment (Bellet et al., 2018) assesses participants' attitudes toward trigger warnings. Participants responded *Yes* or *No* to a single question: "A 'trigger warning' is a statement given prior to presented material that allows the viewer to prepare for or avoid distress that it may cause, particularly if the viewer has clinical mental health issues. Do you think that trigger warnings should be given prior to potentially distressing material?"

Safe Spaces Attitudes Assessment

Participants indicated their attitudes toward safe spaces by responding *yes* or *no* to a single question: "A 'safe space' is a place (as on a college campus) intended to be free of bias, conflict, criticism, or potentially threatening actions, ideas, or conversations. Do you think that the classroom at colleges and institutions of higher learning should be a 'safe space'?" The definition of a safe space was taken verbatim from the most recent Merriam-Webster dictionary entry for "safe space" (Merriam-Webster, n.d.).

Words Can Harm Scale (WCHS)

WCHS (Bellet et al., 2018) measures the extent to which an individual believes that words can cause lasting psychological harm. Participants indicated their agreement with 10 statements (e.g., "Vulnerable people should not be exposed to certain kinds of speech, as this might harm them") on a 100-point slider from *strongly disagree* to *strongly agree*.

Participant LWA

In addition to rating their perceptions of the instructor's left-wing authoritarian beliefs, participants reported their own endorsement of left-wing authoritarian beliefs using the LWA-13 (Costello & Patrick, 2023). Participants rated their agreement with each of the 13 items on a 7-point Likert scale from *strongly agree* to *strongly disagree*.

Participant Political Orientation

Participants reported their political orientation on a 7-point Likert scale ranging from *very liberal* to *very conservative* with a neutral midpoint of *neither liberal nor conservative*.

Trauma Screen and Life Events Checklist (LEC)

Participants completed a brief trauma screener in which they read a description of Criterion A traumatic events and indicated whether

they had ever directly experienced a traumatic event (*yes/no*). Participants who reported experiencing a traumatic event completed the LEC-5 (Weathers, Blake, et al., 2013) to indicate the type of traumatic event and their degree of exposure to it. Additionally, participants indicated whether any of the content the instructors discussed reminded them of the traumatic event they experienced.

PTSD Checklist for Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition

Participants who indicated *yes* to the trauma screener completed the PTSD Checklist for the *Diagnostic and Statistical Manual of Mental Disorders, fifth edition* (PCL-5; Weathers, Litz, et al., 2013), reporting how often they had been bothered by 20 symptoms of PTSD over the past month on a 5-point Likert scale from *not at all* to *extremely*.

Demographic Variables

We assessed participants' age, gender, race, religion, level of education, and socioeconomic status using standard measures available in the online materials.

Analyses

Analyses were conducted with Bayesian mixed-effects linear regression models using the lmBF function from the BayesFactor R package (Morey & Rouder, 2024). The model included the participant identifier as a random effect to account for repeated measures, as each participant viewed two lecture videos. This allowed us to capture both within- and between-subjects variation. The condition was counterbalanced in the randomization design. The full analysis code is available in the Supplemental Materials. We used Bayes Factors to compare the relative evidence for the presence of a parameter against a comparison "null" model lacking that parameter (i.e., using a "top" style comparison in the Bayes Factor package with the default "medium" prior). We used the conventional decision threshold of 3 (or ⅓)—indicating that the alternative (or null) hypothesis is three times as likely as its counterpart—as interpretation of "substantial evidence" in favor of the leading hypothesis (Kass & Raftery, 1995). We interpreted Bayes Factors between ⅓ and 3 as providing insufficient evidence to distinguish between the null and alternative hypotheses.

To aid interpretation, we calculated effect size estimates to assess the magnitude of each effect. For main effects, we computed Cohen's *d* to provide a standardized measure of differences across experimental conditions. Notably, Cohen's *d* captures only the raw magnitude of difference across experimental conditions and does not incorporate within-participant clustering or covariates, as our primary models do. For both main effects and interactions, we also computed ΔR^2 values, which reflect the change in marginal R^2 when adding the relevant predictor to the model. These ΔR^2 values were calculated post hoc using frequentist mixed-effects models (via the lmer package in *R*) matched to the structure of our Bayesian models. Although ΔR^2 values are typically small in models with multiple covariates and random effects, they offer a consistent index of the additional variance explained by a given predictor and account for covariates and within-subject clustering.

For ease of interpretation, the analyses were grouped into the three sections listed below.

Perceptions of the Instructor

The following experimental questions were addressed in this section: Does the presence of a trigger warning and/or safe space notification affect students' perceptions of the instructor's (a) overall epistemic trustworthiness (METI)? (b) expertise (METI Expertise)? (c) integrity (METI Integrity)? (d) benevolence (METI Benevolence)? (e) concern for student well-being? (f) political orientation? (g) overall LWA (LWA-13)? (h) attitudes toward TDC(LWA-TDC)?[3]

Perceptions of the Classroom Environment

The following experimental questions were addressed in this section: Does the presence of a trigger warning and/or safe space notification affect students' perceptions of the classroom environment in terms of (a) psychological safety (Edmondson's Psychological Safety Scale)? (b) overall reluctance to honestly discuss controversial issues? (c) reluctance to honestly discuss (i) politics? (ii) race or ethnicity? (iii) religion? (iv) sexual orientation? (v) gender roles? (vi) trans identity?

Perceptions of the Instructor and Classroom Environment Dependent on Student Characteristics

In this section, we were interested in whether the effect of providing trigger warnings and/or safe space notifications on student perceptions depends on student characteristics. For example, we might imagine that trigger warnings have no overall effect on perceptions of instructor epistemic trustworthiness. However, it might be the case that liberal students perceive greater epistemic trustworthiness when instructors administer trigger warnings, whereas conservative students perceive less epistemic trustworthiness. These questions were analyzed by using linear regression on each main effect, with the variable of interest added as an interaction term. Evidence for the presence of an interaction would suggest that the effect of the condition depends on the students' characteristics.

The following experimental questions were addressed in this section: Does the effect of providing a trigger warning and/or a safe space notification on students' perceptions of the instructor and the classroom environment depend on students' (a) political orientation (1–7 from *very liberal* to *very conservative*)? (b) overall LWA(LWA-13)? (c) attitudes toward TDC(LWA-TDC)? (d) trauma exposure (LEC-5)? (e) PTSD symptoms related to trauma exposure (PCL-5)? (f) belief that WCHS? (g) gender? (h) race?[4,5]

Results

Sample Characteristics

The sample was roughly equally divided between women (n = 370, 50.1%) and men (n = 348, 47.2%) and the mean age was 27.9 years old (SD = 9.2) with a modal age of 21. Most participants were

[3] We earmarked the benevolence subscale of the METI (METI Benevolence) and the TDC subscale of the LWA-13 (LWA-TDC) in our preregistration as being especially theoretically important.

[4] In all analyses for this section, we used political orientation as a control variable and collapsed groups with fewer than 150 observations.

[5] We earmarked trauma exposure and the TDC subscale of the LWA-13 (LWA-TDC) in our preregistration as being especially theoretically important.

White or Caucasian ($n = 420$, 56.9%), with 18.3% Black or African American participants ($n = 135$), 11% Asian participants ($n = 81$), and 13.8% multiracial or other ($n = 102$). In terms of social attitudes, the sample was 39.2% liberal ($n = 289$), 30% moderate ($n = 221$), and 30.9% conservative ($n = 228$). A large majority of participants endorsed the use of trigger warnings ("Yes" = 654, 88.6%; "No" = 84, 11.4%) and safe spaces ("Yes" = 571, 77.4%; "No" = 167, 22.6%). The median time that participants took to complete the survey was 23:29 min.

Table 1 displays bivariate correlations between selected participant characteristics.

Effect of Condition on Perceptions of the Instructor

Overall, trigger warnings had no substantial impact on ratings of the instructor. For almost all instructor ratings, analyses suggested that a null model was much more likely than a model including the trigger warning manipulation ($BF_{10} < ⅓$). Specifically, providing a trigger warning had no substantial impact on perceptions of the instructor as epistemically trustworthy (METI, $BF_{10} = 0.18$, $d = -0.02$, $\Delta R^2 = 9.31$; including the subscales measuring expertise, $BF_{10} = 0.08$, $d = 0.02$, $\Delta R^2 = 1.05$; integrity, $BF_{10} = 0.20$, $d = -0.04$, $\Delta R^2 = 0.001$; and benevolence, $BF_{10} = 0.65$, $d = -0.07$, $\Delta R^2 = 0.002$) or perceptions of the instructor as liberal ($BF_{10} = 0.23$, $d = 0.05$, $\Delta R^2 = 0.002$). For the remaining two instructor ratings—instructor concern about student well-being ($BF_{10} = 2.33$, $d = -0.13$, $\Delta R^2 = 0.005$) and instructor LWA ($BF_{10} = 1.00$, $d = -0.09$, $\Delta R^2 = 0.003$)—analyses suggested insufficient data to adjudicate between the models ($⅓ < BF_{10} < 3$).

In contrast, safe space notifications consistently had a substantial impact on instructor ratings ($BF_{10} > 3$). Providing a safe space notification increased perceptions of instructor epistemic trustworthiness (METI, $BF_{10} = 812.78$, $d = -0.19$, $\Delta R^2 = 0.009$; including the subscales measuring expertise, $BF_{10} = 21.36$, $d = -0.16$, $\Delta R^2 = 0.006$; integrity, $BF_{10} = 124.22$, $d = -0.18$, $\Delta R^2 = 0.008$, and benevolence, $BF_{10} = 13{,}885.65$, $d = -0.22$, $\Delta R^2 = 0.01$) and concern for student well-being ($BF_{10} = 8.39 \times 10^{11}$, $d = -0.41$, $\Delta R^2 = 0.04$). Safe space notifications also increased perceptions that the instructor was liberal ($BF_{10} = 242.29$, $d = 0.20$, $\Delta R^2 = 0.01$) and that the instructor was high in LWA ($BF_{10} = 8.81$, $d = -0.12$, $\Delta R^2 = 0.006$), including the subscale measuring the tendency to engage in TDC (LWA-TDC, $BF_{10} = 108.67$, $d = -0.14$, $\Delta R^2 = 0.007$). The METI benevolence subscale (see Figure 1) and LWA-TDC (see Figure 2) are notable as they were the two scales earmarked in the preregistration as theoretically important dependent variables.

In several cases, the two conditions interacted in an interesting way. In all cases, where a substantial interaction occurred, the pattern was the same: Safe space notifications alone caused a shift in perceptions but when both a safe space notification and a trigger warning were given, the trigger warning dampened or nullified the effect of the safe space notification. This is especially interesting given that there were no substantial main effects of trigger warnings. This interaction was substantially supported ($BF_{10} > 3$) for perceptions of instructor epistemic trustworthiness (METI, $BF_{10} = 6.81$, $\Delta R^2 = 0.004$), including the subscales measuring benevolence ($BF_{10} = 40.90$, $\Delta R^2 = 0.006$; Figure 3) and integrity ($BF_{10} = 11.38$, $\Delta R^2 = 0.004$) but was ambiguous for the subscale measuring expertise ($BF_{10} = 0.61$, $\Delta R^2 = 0.002$). This interaction was also substantially supported for perceptions of instructor concern for student well-being ($BF_{10} = 115{,}117.39$, $\Delta R^2 = 0.02$) and instructor support for TDC ($BF_{10} = 3.82$, $\Delta R^2 = 0.004$), but was ambiguous for overall perceptions of instructor LWA ($BF_{10} = 1.40$, $\Delta R^2 = 0.003$).

Effect of Condition on Perceptions of the Classroom Environment

Trigger warnings had no substantial impact on perceptions of the hypothetical classroom environment. All analyses suggested a greater likelihood of a null model ($BF_{10} < ⅓$). Specifically, trigger warnings had no impact on participant ratings of psychological safety ($BF_{10} = 0.08$, $d = -0.001$, $\Delta R^2 < 0.001$) and reluctance to discuss controversial topics in the classroom ($BF_{10} = 0.07$, $d = 0.02$, $\Delta R^2 < 0.001$), including politics ($BF_{10} = 0.08$, $d = 0.07$, $\Delta R^2 < 0.001$), race ($BF_{10} = 0.08$, $d = 0.004$, $\Delta R^2 < 0.001$), religion ($BF_{10} = 0.08$, $d = 0.05$, $\Delta R^2 < 0.001$), sexual orientation ($BF_{10} = 0.13$, $d = 0.04$, $\Delta R^2 < 0.001$), gender ($BF_{10} = 0.07$, $d = 0.01$, $\Delta R^2 < 0.001$), and trans identity ($BF_{10} = 0.08$, $d = 0.004$, $\Delta R^2 < 0.001$).

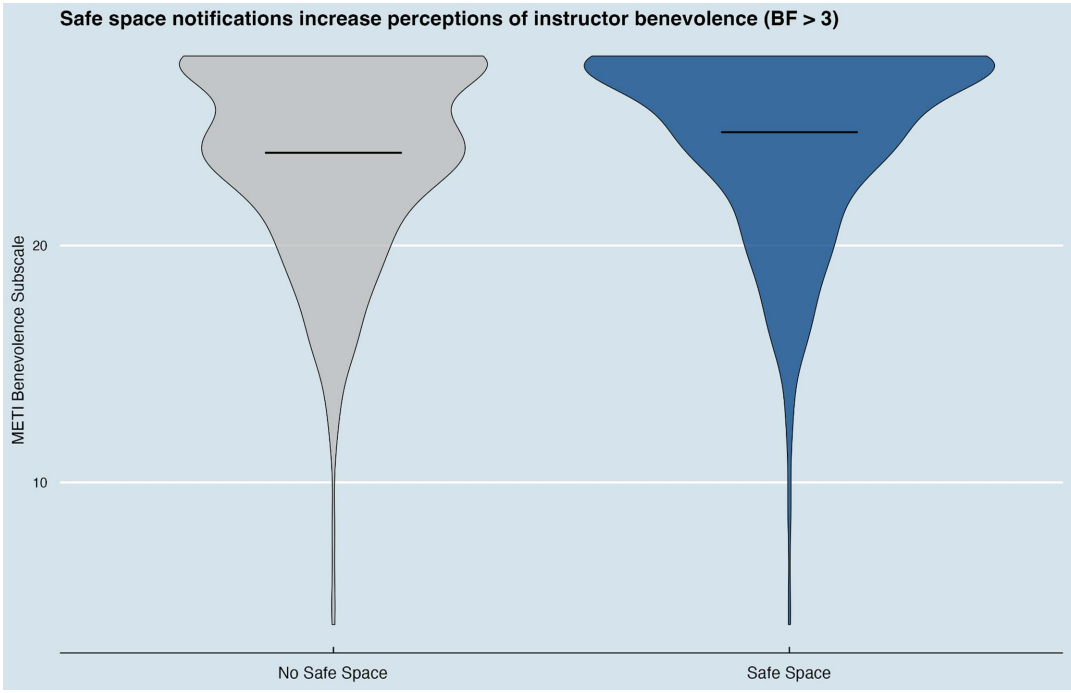
Table 1
Bivariate Correlations Between Selected Participant Characteristics

Variable	1	2	3	4	5	6	7	8	9
1. Conservatism	—								
2. SES	.07	—							
3. TW endorsement	−.17	−.06	—						
4. SS endorsement	−.09	.01	.35	—					
5. WCHS	−.18	−.03	.43	.33	—				
6. Trauma history[a]	−.06	−.07	.02	−.03	.11	—			
7. PCL-5[b]	−.03	−.13	.09	.07	.20	n/a[c]	—		
8. LWA-13	−.55	−.03	.31	.36	.35	.02	.16	—	
9. LWA-13 TDC	−.29	.01	.38	.52	.48	.03	.12	.74	—

Note. $N = 738$. SES = continuous measure of total household income in the past 12 months; TW Endorsement = support for trigger warnings; SS Endorsement = support for safe space notifications; WCHS = belief that words can harm; PCL-5 = PTSD symptoms; LWA-13 = Left-Wing Authoritarianism scale; LWA-13 TDC = top-down censorship subscale of the Left-Wing Authoritarianism scale.
[a] $n = 329$. [b] $n = 328$. [c] Cell missing because participants only completed the PCL-5 if they had a history of trauma.

Figure 1

Safe Space Notifications Increase Perceptions of Instructor Benevolence ($BF_{10} > 3$)



Note. METI = Muenster Epistemic Trustworthiness Inventory. See the online article for the color version of this figure.

In contrast, safe space notifications had substantial favorable impacts on perceptions of the classroom environment ($BF_{10} > 3$). Safe space notifications increased participant perceptions of psychological safety ($BF_{10} = 8{,}993.18$, $d = 0.23$, $\Delta R^2 = 0.01$). Safe space notifications also decreased participant reluctance to speak about controversial issues in the classroom ($BF_{10} = 182.41$, $d = -0.15$, $\Delta R^2 = 0.008$), including politics ($BF_{10} = 6.25$, $d = -0.12$, $\Delta R^2 = 0.004$), religion ($BF_{10} = 139.39$, $d = -0.15$, $\Delta R^2 = 0.007$), sexual orientation ($BF_{10} = 166.07$, $d = -0.14$, $\Delta R^2 = 0.008$), gender ($BF_{10} = 4.31$, $d = -0.11$, $\Delta R^2 = 0.005$), and trans identity ($BF_{10} = 13.53$, $d = -0.13$, $\Delta R^2 = 0.005$). The effect of discussing race was ambiguous ($BF_{10} = 0.60$, $d = -0.08$, $\Delta R^2 = 0.002$). Additionally, safe space notifications substantially decreased perceptions that the instructor would react negatively for voicing one's honest thoughts when aggregated across all types of concerns about instructor actions (e.g., giving a lower grade, $BF_{10} = 3.55$, $d = -0.09$, $\Delta R^2 = 0.003$). Breaking down by individual types of retaliation, the only individual effect that remained substantial was a small decrease in the perception that the instructor would dislike the student for voicing their honest thoughts ($BF_{10} = 15.29$, $d = -0.14$, $\Delta R^2 = 0.006$).

Once again, the two conditions interacted such that safe space notifications alone had large impacts, but these were dampened when a trigger warning was also included. This interaction was substantially supported ($BF_{10} = 3.34$, $\Delta R^2 = 0.003$) for reluctance to speak about controversial topics overall (the sum across all topics; the effect was ambiguous when breaking down into individual topics). This interaction was also present for participant perceptions of psychological safety ($BF_{10} = 1{,}179.83$, $\Delta R^2 = 0.01$).

Interactions With Participant Characteristics

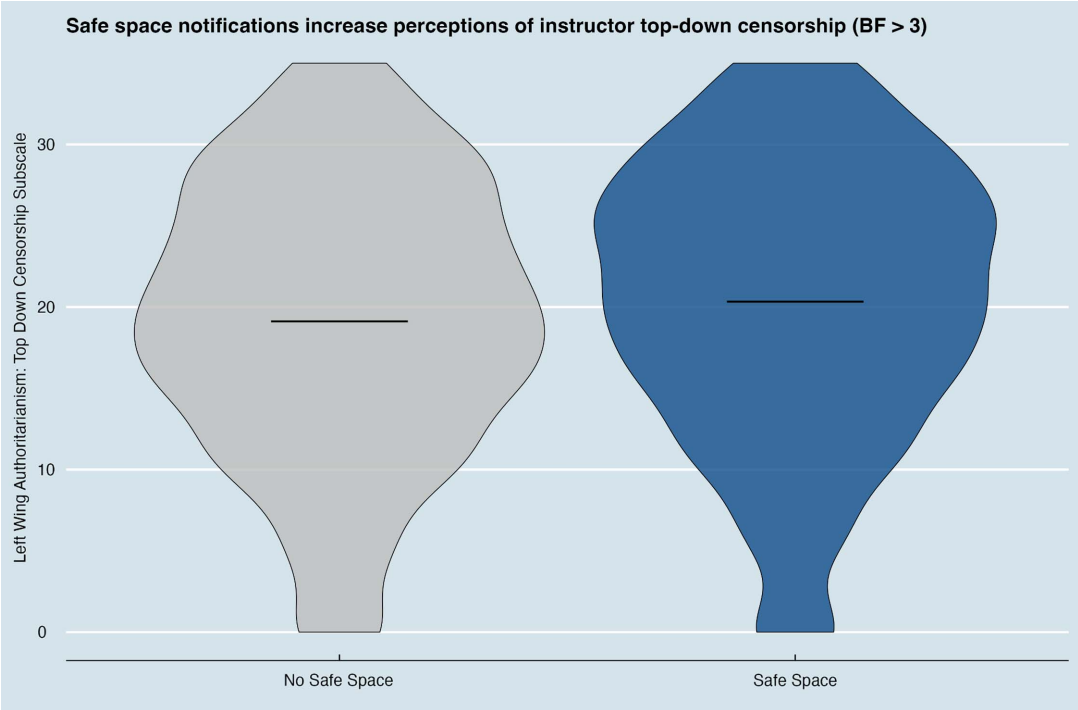
Although the overall effects of the conditions are interesting, these are, of course, dependent on the specific sample we collected. We were particularly interested in understanding shifts in perceptions among certain kinds of students. For instance, we preregistered that student trauma exposure and support for TDC were theoretically important characteristics that may moderate how trigger warnings or safe space notifications affect perceptions.

Indeed, student support for TDC (LWA-TDC) did substantially moderate perceptions of instructors. Specifically, trigger warnings increased perceptions of instructor LWA ($BF_{10} = 26.77$, $\Delta R^2 = 0.006$), TDC ($BF_{10} = 14.37$, $\Delta R^2 = 0.005$), and likelihood to criticize students' honest thoughts as being offensive ($BF_{10} = 5.08$, $\Delta R^2 = 0.004$), but only among students who scored low in TDC. Interestingly, we found consistently null moderation effects for student political orientation, suggesting that as predicted, LWA-TDC is the more theoretically relevant variable for moderating perceptions of trigger warnings (Figure 4).

In contrast, when considering students' past exposure to trauma as a moderating variable, we found consistent support for the null model ($BF_{10} < ⅓$). That is, trigger warnings and safe space notifications were not any more impactful for students with a history of trauma than for students who had never experienced trauma (Figure 5). Among trauma-exposed students, the level of PTSD symptoms reported also had no substantial effect on any perception. Importantly, among the 329 participants who reported a history of trauma, the two most common events were (a) having directly experienced sexual assault ($n = 71$) and (b) having directly experienced physical assault ($n = 31$), suggesting that the trigger warning

Figure 2

Safe Space Notifications Increase Perceptions of Instructor Top-Down Censorship ($BF_{10} > 3$)



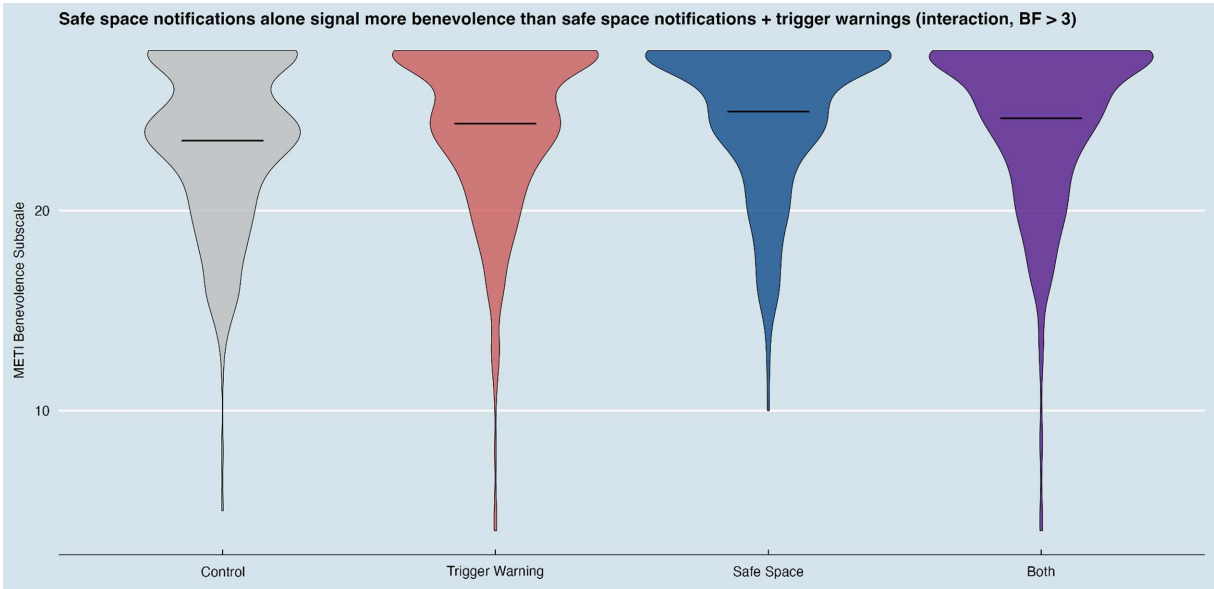
Note. See the online article for the color version of this figure.

(which warned of interpersonal trauma, like sexual violence) was relevant. We also directly asked these 329 participants whether any of the content the instructors discussed reminded them of their trauma. Forty-four percent ($n = 144$) responded "Yes."

These results underscore that contrary to common assumptions, trauma history and PTSD symptoms do not meaningfully shape perceptions of trigger warnings or safe space notifications. The Bayes Factors and effect size estimates for the moderation of trauma

Figure 3

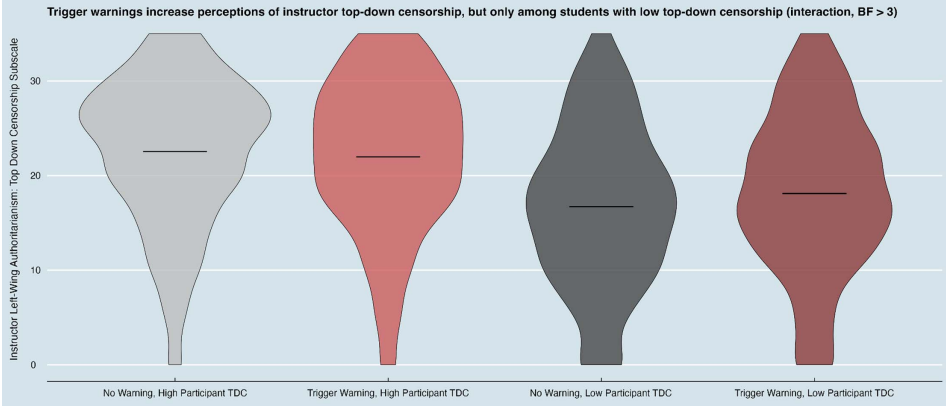
Safe Space Notifications Alone Signal More Benevolence Than When Combined With a Trigger Warning ($BF_{10} > 3$)



Note. METI = Muenster Epistemic Trustworthiness Inventory. See the online article for the color version of this figure.

Figure 4

Trigger Warnings Increase Perceptions of Instructor Support for Top-Down Censorship, but Only Among Students With Low Support for Top-Down Censorship ($BF_{10} > 3$)



Note. See the online article for the color version of this figure.

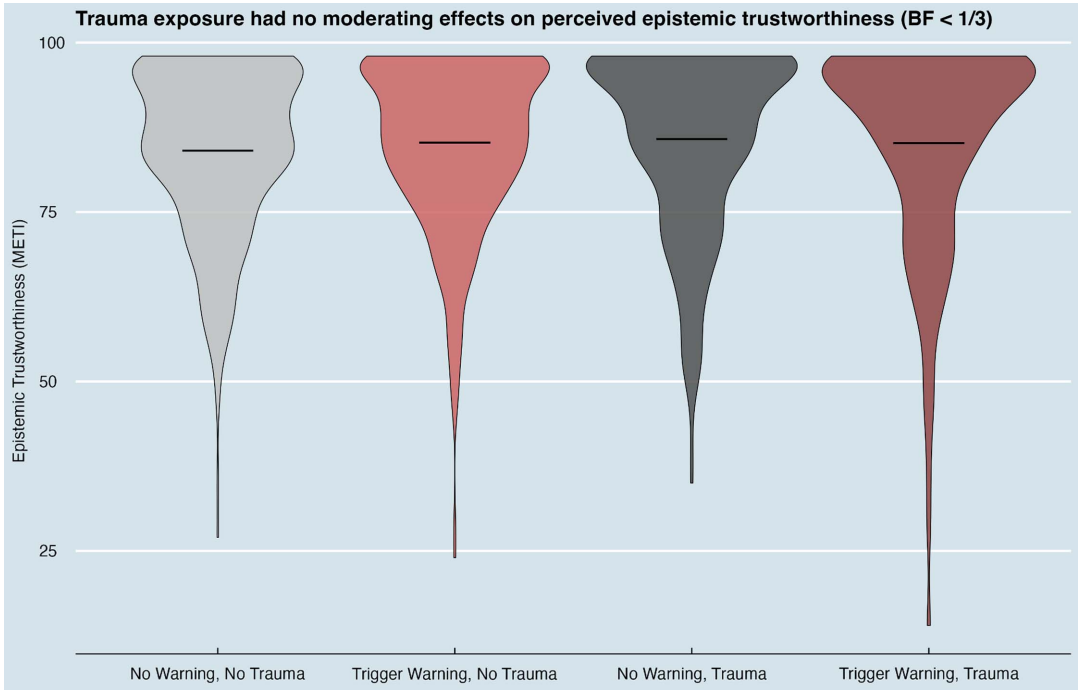
history and PTSD symptoms on each outcome variable are reported in the Supplemental Materials.

Considering additional participant characteristics, we did find several interesting moderation effects. However, we urge caution with these analyses, as the following results are pulled from a large group of moderation tests, and unlike the previously reported effects, were not specifically earmarked in the preregistration as theoretically important moderations. For gender, we found moderation effects on epistemic trustworthiness when trigger warnings were given. Providing a trigger warning increased perceptions of the benevolence ($BF_{10} = 10.14$, $\Delta R^2 = 0.004$) and integrity ($BF_{10} = 6.01$, $\Delta R^2 = 0.003$) subscales among females, but males had lower scores on the same subscales when trigger warnings were given.

We found that race moderated the effect of safe space notifications on perceptions of instructor support for LWA ($BF_{10} = 4.97$, $\Delta R^2 = 0.009$) and TDC ($BF_{10} = 4.27$, $\Delta R^2 = 0.007$). Specifically, giving a safe space notification increased perceptions of the instructors' LWA and TDC for White and multiracial students, but the

Figure 5

Trauma Exposure Had No Moderating Effects on Perceived Epistemic Trustworthiness ($BF_{10} < ⅓$)



Note. METI = Muenster Epistemic Trustworthiness Inventory. See the online article for the color version of this figure.

effect flipped for Black students, who perceived the instructor as *less* authoritarian. Results were inconsistent for Asian students.

Finally, we found several moderating effects depending on the extent to which students endorsed the WCHS. Recall that students high on this scale tended to endorse the belief that words can cause permanent emotional damage, whereas students who scored low on this scale believed the opposite. We found substantial evidence ($BF_{10} > 3$) that students low in WCHS reacted differently—when given trigger warnings, they perceived the instructor as higher in LWA ($BF_{10} = 4.71$, $\Delta R^2 = 0.004$), felt more reluctant to discuss politics in the classroom ($BF_{10} = 5.77$, $\Delta R^2 = 0.003$), and felt more strongly that the instructor would criticize their honest thoughts as being offensive ($BF_{10} = 26.49$, $\Delta R^2 = 0.006$). Participants low in WCHS were also more reluctant to discuss religion when given safe space notifications ($BF_{10} = 7.07$, $\Delta R^2 = 0.004$). In contrast, students scoring high in WCHS perceived instructors who gave safe space notifications as higher in integrity ($BF_{10} = 7.21$, $\Delta R^2 = 0.004$).

Sensitivity Analyses

The TDC subscale of the LWA-13 has an item that relates to the approval of at least some variants of the concept of a safe space ("Classroom discussions should be safe places that protect students from disturbing ideas"). Though related, this item is somewhat distinct from the safe space notification given in the study stimuli. ("Before we begin today's lecture, I want to emphasize to everyone that this classroom is a safe space. If at any point the material becomes too distressing, please feel free to disengage as necessary. It's essential to prioritize your emotional safety.") We conducted sensitivity analyses removing this item to see if it captured important variance relevant to perceptions of instructor LWA. The main effect of safe space notifications on perceptions of instructor LWA-TDC remained substantial ($BF_{10} = 4.23$, $d = 0.10$, $\Delta R^2 = 0.004$), but the effect on perceptions of instructor LWA became ambiguous ($BF_{10} = 1.28$, $d = -0.08$, $\Delta R^2 = 0.003$). In the safe space condition, the moderation of race on the perception of instructor LWA-TDC remained ($BF_{10} = 7.39$, $\Delta R^2 = 0.007$), but the moderation on the perception of instructor LWA became ambiguous ($BF_{10} = 2.97$, $\Delta R^2 = 0.008$). In the trigger warning condition, the moderation of the WCHS on perceptions of instructor LWA remained substantial ($BF_{10} = 4.15$, $\Delta R^2 = 0.004$), as did the moderation of participant LWA-TDC on perceptions of instructor LWA-TDC ($BF_{10} = 4.24$, $\Delta R^2 = 0.004$). In the safe space condition, the moderation of participant LWA-TDC on perceptions of instructor LWA-TDC became null ($BF_{10} = 0.33$, $\Delta R^2 = 0.001$). This seems to indicate that this item captures some of, but not all, the important variance relevant to shifting perceptions of instructor authoritarianism.

Discussion

We experimentally investigated how the presence of a trigger warning and/or a safe space notification impacts students' perceptions of instructors and the classroom environment. We found that trigger warnings had no overall impact on students' perceptions. However, providing safe spaces made most students feel more psychologically safe and report more willingness to discuss controversial topics in the classroom. When an instructor provided a safe space notification, students perceived them as more benevolent and caring but also viewed them as more liberal and left-wing authoritarian. We discuss these results in turn as they relate to conversations about the merits and drawbacks of providing warnings in educational settings.

Although much theoretical discussion surrounds trigger warnings and safe spaces in public and philosophical debate (e.g., Fast, 2019; Lukianoff & Haidt, 2018), empirical research has focused almost entirely on their clinical effects rather than their interpersonal effects (Bridgland et al., 2024). Yet, a strong tradition of research on social perception and signaling theory (Donath, 2007; Fiske et al., 2007) suggests that people's behaviors serve as powerful cues of their intentions, values, and commitments. People routinely manage others' impressions of them by dressing well for job interviews (Spence, 1973), performing costly acts of altruism (Henrich, 2009), or openly expressing their moral values (Tosi & Warmke, 2016). These cues are interpreted by others to form impressions of trustworthiness (Hendriks et al., 2015), social competence (Fiske et al., 2007), and ideology (Tosi & Warmke, 2016). Our study expands the theoretical conversation on trigger warnings and safe spaces by examining the interpersonal signals they send. Our findings suggest that these practices do not merely function as mental health accommodations but also shape how students perceive their instructors and the classroom environment.

Trigger Warnings and Safe Space Notifications as Interpersonal Signals

Proponents of trigger warnings have suggested that these practices signal solicitude and respect (Bloom, 2023; Sastre, 2017), whereas critics have argued that any such effects are mere ideological "virtue signals" (Suk Gersen, 2021; Tosi & Warmke, 2016). In our study, trigger warnings had no overall impact on students' perceptions of the classroom environment or the instructor, suggesting that they do not signal virtue at all—rather, they have very little impact on students' perceptions. This is somewhat surprising given that in our sample, a substantial majority of students endorsed the statement that "trigger warnings should be given prior to potentially distressing material." This suggests that despite students having mostly positive attitudes toward trigger warnings, instructors who offer them are not rated any more highly than instructors who do not. However, several variables moderated the effect of trigger warnings on students' perceptions, including students' beliefs that words can harm and support for TDC. We discuss these moderating variables below.

Though trigger warnings consistently had no overall effect, safe spaces did alter students' perceptions in reliable ways. They caused students to rate the instructor as more epistemically trustworthy and as caring more about the well-being of their students. They also caused students to feel more psychologically safe and report less reluctance to discuss controversial issues in the classroom. This is consistent with the argument that safe spaces can encourage dialogue about difficult issues by creating an environment where students feel supported (Fast, 2019). Of course, more research is needed to conclude whether students' expectations about classrooms that provide safe spaces match their actual emotional and behavioral outcomes when safe spaces are provided (e.g., their actual willingness to discuss controversial topics in real classroom

conversations). Overall, these findings help advance a very small body of work on the signaling effects of safe spaces.

The fact that safe space notifications led students to perceive the instructor as more liberal and left-wing authoritarian (particularly in terms of supporting TDC) is significant to current debates about maintaining ideological neutrality in the classroom. Some have argued against instructors signaling their ideological affiliation (Suk Gersen, 2021), given the high number of students who report self-censoring their political views for fear of backlash (Zhou & Barbaro, 2023). There is some evidence that students prefer politically ambiguous professors to strongly partisan professors, especially when the professor's views conflict with their own views (Giersch, 2020; Liebertz & Giersch, 2021). However, there is also evidence that students across the entire political spectrum prefer mildly liberal professors (Liebertz & Giersch, 2021), perhaps explaining why, in our sample, providing a safe space notification made professors seem more liberal *and* more benevolent and caring, even among conservative students. Overall, students in our study expressed *less* reluctance to voice their honest thoughts when a safe space notification was given. Nonetheless, our results suggest that to the extent that professors aim to be seen as politically neutral by their students, providing a safe space notification may obstruct this goal, even as it advances other goals.

Future research could attempt to mitigate perceptions of partisanship by modifying the safe space notification to make it more politically neutral. For instance, it could avoid politically valenced terms like "safe space" and "emotional safety," remove recommendations that students should avoid distress, and instead more directly signal the instructor's solicitude and goals for the classroom (Palus, 2019). Consider a reworded notification: "Before we begin, I want to remind you that this classroom is a place for open and civil discussion. Your education is incredibly important to me, but so is your mental health and well-being. Let's commit to supporting one another by showing respect and kindness, even though we may have very different life experiences or perspectives." Though empirical validation is needed, it is possible that such a notification could preserve the positive impacts of safe space notifications while avoiding signaling partisanship.

Interestingly, providing a trigger warning prior to a safe space notification tended to dampen the signaling effects of the safe space notification. This is surprising since providing a trigger warning on its own had no effect. Given that the trigger warning mentioned sexual assault, one possibility is that participants who did not feel that this related to them "tuned out" once they received the trigger warning, reducing their attention to the subsequent safe space notification. In videos with both a trigger warning and safe space notification, the order of the trigger warning and safe space notification was not randomized; the trigger warning always appeared first. This is a limitation of our design, and it is unknown whether the outcome would differ if the safe space notification came first.

In terms of moderation effects, it is noteworthy that the effect of trigger warnings and safe spaces on students' perceptions did not depend on students' self-reported political ideology. In other words, liberal and conservative students reacted similarly to these practices. This is surprising given that the debate about using these practices is often portrayed as a political debate (Lukianoff & Haidt, 2018), with conservatives expressing more criticisms (Sevincer et al., 2024; Witherup & Verrecchia, 2020). In our sample, conservatives were less likely than liberals to endorse using trigger warnings and safe spaces, though only very slightly, with very high endorsement rates among both groups (trigger warning endorsement = 92% of liberals, 83% of conservatives; safe space endorsement = 78% of liberals, 75% of conservatives). Of course, we should not necessarily expect political attitudes regarding the somewhat dated "culture wars" to hold in Gen Z college students, who differ from previous generations in their social attitudes (Twenge, 2023). Many factors likely interact to determine one's attitudes toward practices like trigger warnings and safe spaces. Our results suggest that while participants' political orientation did not moderate any effects, participants' belief that words can harm and support for TDC were consistent moderators. As a rule, those who scored low on these variables reacted more negatively to trigger warnings and safe spaces (e.g., viewing instructors as higher in TDC), whereas participants who scored high on these variables generally reacted more positively (e.g., viewing instructors as higher in integrity). This set of findings is consistent with the notion that regardless of political orientation, the degree to which students feel that certain forms of discourse are inherently dangerous and in need of policing (arguably present among both conservatives and liberals in different ways) is more operative in shaping the effects of interventions (such as safe spaces) that seek to shape discourse. Given the theoretical relevance of these variables to discussions of trigger warnings, safe spaces, and free speech, they are ripe for future investigation.

It is important that neither students' trauma history nor their level of PTSD symptoms alter the impact of providing trigger warnings or safe spaces. Originally, the argument for providing trigger warnings in the classroom was framed as enabling trauma survivors to either avoid or cope with potential reminders of their trauma (e.g., Stokes, 2014). However, research on individuals with trauma histories has found no evidence that trigger warnings are helpful, even when the content that is warned about matches the individual's past trauma (e.g., warning a sexual assault survivor about mentions of sexual assault; Jones et al., 2020). Our findings align with this research, indicating that trigger warnings and safe spaces do not hold unique importance in conveying interpersonal signals to trauma survivors compared to students without a trauma history. Although personal experience with trauma is often used to justify the implementation of trigger warnings and safe spaces, research suggests that these practices do not provide distinct benefits nor convey distinct signals to students affected by trauma.

Limitations

Several factors limit the inferences that can be drawn from our results. For example, we used brief prerecorded lecture videos and asked participants to answer questions about the instructors and classroom while imagining that they were in the instructor's course. This limited experimental format may reduce the extent to which participants' ratings generalize to real educational settings. In actual classroom environments, an instructor providing a trigger warning or safe space notification is just one of many factors that might determine students' perceptions of them. For example, a student could form an initial impression of an instructor who provides a safe space notification on the first day, but this impression might be washed away by later interactions throughout

the semester. More research is needed to determine the long-term effects of providing warnings on student perceptions.

At the same time, there is reason to trust participants' ratings based on viewing the lecture videos. Classic psychology research shows that individuals form remarkably accurate perceptions of others based on "thin slices" of their behavior, including brief audiovisual clips (Ambady & Rosenthal, 1992; Ambady et al., 2000; see Murphy & Hall, 2021 for a recent review). In one famous study, participants predicted election outcomes better than chance after viewing pictures of the candidates for just 100 ms and rating their competence (Ballew & Todorov, 2007). Though observing students' judgments in a real classroom setting would be ideal, this research increases our confidence that participants' ratings are reliable.

A second limitation is the potential for political bias in the sample we collected. Online participant recruitment platforms are known to have a user base that is substantially more liberal than the general population (Levay et al., 2016). We tried to mitigate this imbalance by prescreening participants for their political orientation and enforcing a political quota based on a 2024 Gallup poll indicating a roughly 33/33/33 breakdown of liberal, moderate, and conservative views on social issues (McCarthy, 2024). However, though we immediately recruited our quota of liberal participants, we struggled to recruit conservative participants and extend the survey for several weeks to fill this quota, raising concerns about the validity and quality of the conservative participants compared to their liberal counterparts. Additionally, given that we used a convenience sample, some theoretically surprising results lead us to question whether our sample would properly generalize to a more representative population in terms of political orientation. For example, it seems surprising that among conservative students, perceptions of instructors as high in left-wing authoritarian TDC were correlated with perceptions of them as benevolent and caring ($r = 0.23$, LWA-TDC and METI benevolence, $r = 0.31$, LWA-TDC and concern for student well-being). However, it is impossible to know whether this result is a product of inaccurate self-reporting of political ideology, a quirk of a unique online convenience sample, or a genuine trend that would apply to the broader population. We highly recommend that future studies use diverse recruitment methods and employ more complete measures of participants' ideology.

Constraints on Generality

We collected a sample of current American undergraduate students to draw conclusions about the impact of trigger warnings and safe space notifications on this population. One factor constraining the generalizability of our study is that participants completed all tasks individually on a computer. If safe spaces and trigger warnings are often used to regulate emotions and signal respect in group settings (Gainsburg & Earl, 2022), we would want to see how these effects generalize to environments in which participants are learning alongside each other and affected by the imagined and actual emotions of others. Indeed, requests for trigger warnings and safe space notifications are sometimes made on behalf of other students, and many individuals report being more anxious about the reactions of other individuals than they are about their own (Bellet et al., 2018, 2020).

Conclusion

Research on trigger warnings has primarily focused on evaluating their effectiveness as a mental health tool, finding them to be ineffective for that purpose (Bridgland et al., 2024). There is little empirical research on safe space notifications. The present research broadens the conversation about both practices by examining the interpersonal signals they send. Trigger warnings had no impact on students' overall perceptions, despite widespread endorsement among students. Safe space notifications consistently signaled to students that instructors were benevolent and caring. Safe spaces also increased students' feelings of psychological safety and willingness to discuss controversial issues in the classroom. However, they also heightened students' perceptions of instructors as ideologically partisan, specifically as left-wing authoritarian and supportive of TDC. These findings highlight the importance of weighing the potential costs and benefits of these practices in context.

References

Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social Psychology*, *32*, 201–271. https://doi.org/10.1016/S0065-2601(00)80006-4

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*(2), 256–274. https://doi.org/10.1037/0033-2909.111.2.256

Ballew, C. C., II, & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(46), 17948–17953. https://doi.org/10.1073/pnas.0705435104

Bellet, B. W., Jones, P. J., & 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

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

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

Bloom, P. (2023). Why I use trigger warnings [Substack]. *Small Potatoes*. https://smallpotatoes.paulbloom.net/p/why-i-use-trigger-warnings?r=lh5af&utm_campaign=post&utm_medium=web&showWelcomeOnShare=false

Boysen, G. A. (2017). Evidence-based answers to questions about trigger warnings for clinically-based distress: A review for teachers. *Scholarship of Teaching and Learning in Psychology*, *3*(2), 163–177. https://doi.org/10.1037/stl0000084

Boysen, G. A., & Prieto, L. R. (2018). Trigger warnings in psychology: Psychology teachers' perspectives and practices. *Scholarship of Teaching and Learning in Psychology*, *4*(1), 16–26. https://doi.org/10.1037/stl0000105

Bridgland, V. M. E., Jones, P. J., & Bellet, B. W. (2024). A meta-analysis of the efficacy of trigger warnings, content warnings, and content notes. *Clinical Psychological Science*, *12*(4), 751–771. https://doi.org/10.1177/21677026231186625

Burch, G. F., Batchelor, J. H., Burch, J. J., Gibson, S., & Kimball, B. (2018). Microaggression, anxiety, trigger warnings, emotional reasoning, mental

filtering, and intellectual homogeneity on campus: A study of what students think. *Journal of Education for Business*, *93*(5), 233–241. https://doi.org/10.1080/08832323.2018.1462137

Celniker, J. B., Ringel, M. M., Nelson, K., & Ditto, P. H. (2022). Correlates of "Coddling": Cognitive distortions predict safetyism-inspired beliefs, belief that words can harm, and trigger warning endorsement in college students. *Personality and Individual Differences*, *185*, Article 111243. https://doi.org/10.1016/j.paid.2021.111243

Costello, T. H., Bowes, S. M., Stevens, S. T., Waldman, I. D., Tasimi, A., & Lilienfeld, S. O. (2022). Clarifying the structure and nature of left-wing authoritarianism. *Journal of Personality and Social Psychology*, *122*(1), 135–170. https://doi.org/10.1037/pspp0000341

Costello, T. H., & Patrick, C. J. (2023). Development and initial validation of two brief measures of left-wing authoritarianism: A machine learning approach. *Journal of Personality Assessment*, *105*(2), 187–202. https://doi.org/10.1080/00223891.2022.2081809

Donath, J. (2007). Signals in social supernets. *Journal of Computer-Mediated Communication*, *13*(1), 231–251. https://doi.org/10.1111/j.1083-6101.2007.00394.x

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, *44*(2), 350–383. https://doi.org/10.2307/2666999

Ellison, J. (2016). *Dear class of 2016 student*. https://news.uchicago.edu/sites/default/files/attachments/Dear_Class_of_2020_Students.pdf. Retrieved on April 20, 2025.

Engle, J. (2023 May 10). Should teachers provide trigger warnings for 'traumatic content'? *The New York Times*. https://www.nytimes.com/2023/05/10/learning/should-teachers-provide-trigger-warnings-for-traumatic-content.html

Fast, J. (2019). In defense of safe spaces: A phenomenological account. *Atlantis*, *39*(2), 1–22. https://doi.org/10.7202/1064069ar

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

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Gainsburg, I., & Earl, A. (2022). Safe here, but unsafe there? Institutional signals of identity safety also signal prejudice in the broader environment. *Journal of Experimental Social Psychology*, *98*, Article 104232. https://doi.org/10.1016/j.jesp.2021.104232

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

Giersch, J. (2020). Professors' politics and their appeal as instructors. *PS, Political Science & Politics*, *53*(2), 281–285. https://doi.org/10.1017/S104909651900194X

Hendriks, F., Kienhues, D., & Bromme, R. (2015). Measuring laypeople's trust in experts in a digital age: The Muenster Epistemic Trustworthiness Inventory (METI). *PLOS One*, *10*(10), Article e0139309. https://doi.org/10.1371/journal.pone.0139309

Henrich, J. (2009). The evolution of costly displays, cooperation and religion credibility enhancing displays and their implications for cultural evolution. *Evolution and Human Behavior*, *30*(4), 244–260. https://doi.org/10.1016/j.evolhumbehav.2009.03.005

Jones, P. J., Bellet, B. W., & McNally, R. J. (2020). Helping or harming? The effect of trigger warnings on individuals with trauma histories. *Clinical Psychological Science*, *8*(5), 905–917. https://doi.org/10.1177/2167702620921341

Kamenetz, A. (2016). Half of professors in NPR Ed survey have used "Trigger Warnings." *NPR*. https://www.npr.org/sections/ed/2016/09/07/492979242/half-of-professors-in-npr-ed-survey-have-used-trigger-warnings

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of mechanical turk samples. *SAGE Open*, *6*(1), Article 433. https://doi.org/10.1177/2158244016636433

Liebertz, S., & Giersch, J. (2021). Political professors and the perception of bias in the college classroom. *PS, Political Science & Politics*, *54*(4), 755–760. https://doi.org/10.1017/S1049096521000640

Lukianoff, G., & Haidt, J. (2018). *The coddling of the American mind: How good intentions and bad ideas are setting up a generation for failure*. Penguin Press.

McCarthy, J. (2024). *Increase in liberal views brings ideological parity on social issues*. Gallup. https://news.gallup.com/poll/645776/increase-liberal-views-brings-ideological-parity-social-issues.aspx

McNally, R. J. (2016). If you need a trigger warning, you need P.T.S.D. Treatment. *The New York Times*. https://www.nytimes.com/roomfordebate/2016/09/13/do-trigger-warnings-work/if-you-need-a-trigger-warning-you-need-ptsd-treatment

Merriam-Webster. (n.d.). Safe space. *Merriam-Webster.com dictionary*. Retrieved June 5, 2025, from https://www.merriam-webster.com/dictionary/safe%20space

Morey, R., & Rouder, J. (2024). *BayesFactor: Computation of Bayes factors for common designs (R package Version 0.9.12-4.4)* [Computer software]. https://cran.r-project.org/web//packages//BayesFactor/BayesFactor.pdf

Murphy, N. A., & Hall, J. A. (2021). Capturing behavior in small doses: A review of comparative research in evaluating thin slices for behavioral measurement. *Frontiers in Psychology*, *12*, Article 667326. https://doi.org/10.3389/fpsyg.2021.667326

Palus, S. (2019). The latest study on trigger warnings finally convinced me they're not worth it. *Slate*. https://slate.com/technology/2019/07/trigger-warnings-research-shows-they-dont-work-might-hurt.html

Pratt, S., Bellet, B. W., Jones, P. J., & McNally, R. J. (2025). *Testing the coddling hypothesis: Campus safetyism and student resilience*. PsyArXiv. https://doi.org/10.31234/osf.io/zav5g

Pratt, S., Jones, P. J., Bridgland, V., Bellet, B. W., & McNally, R. J. (2025). *Sending signals: The case of trigger warnings and safe space notifications*. [OSF Repository]. Open Science Framework. https://osf.io/9sbnt/?8fd741f038574d3783d8c94f1724c04a

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org

Robinaugh, D. J., & McNally, R. J. (2011). Trauma centrality and PTSD symptom severity in adult survivors of childhood sexual abuse. *Journal of Traumatic Stress*, *24*(4), 483–486. https://doi.org/10.1002/jts.20656

Sastre, A. (2017). Trigger warnings and safe spaces: Reflections and strategies. *Ideas on Fire*. https://ideasonfire.net/trigger-warnings-and-safe-spaces/

Sevincer, A. T., Tenbrueggen, L., & Sokolis, M. (2024). Students' beliefs about trigger warnings. *Psychological Reports*. Advance online publication. https://doi.org/10.1177/00332941241308788

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, *87*(3), 355–374. https://doi.org/10.2307/1882010

Stokes, M. (2014). *In defense of trigger warnings*. The Chronicle of Higher Education. https://www.chronicle.com/blogs/conversation/in-defense-of-trigger-warnings

Suk Gersen, J. (2021). *What if trigger warnings don't work?* The New Yorker. https://www.newyorker.com/news/our-columnists/what-if-trigger-warnings-dont-work

Tosi, J., & Warmke, B. (2016). Moral grandstanding. *Philosophy & Public Affairs*, *44*(3), 197–217. https://doi.org/10.1111/papa.12075

Twenge, J. M. (2023). *Generations: The real differences between Gen Z, Millennials [Boomers, and Silents—And what they mean for America's. In Future]*. Atria Books.

Weathers, F. W., Blake, D. D., Schnurr, P. P., Kaloupek, D. G., Marx, B. P., & Keane, T. M. (2013). *The life events checklist for DSM-5 (LEC-5)*. National Center for PTSD. https://www.ptsd.va.gov

Weathers, F. W., Litz, B. T., Keane, T. M., Palmieri, P. A., Marx, B. P., & Schnurr, P. P. (2013). *The PTSD checklist for DSM-5 (PCL-5)*. National Center for PTSD. https://www.ptsd.va.gov

Witherup, K. L., & Verrecchia, P. (2020). Safe spaces on campus: An examination of student and faculty perceptions. *Journal of Education and Training Studies*, *8*(6), 42–49. https://doi.org/10.11114/jets.v8i6.4874

Zhou, S., & Barbaro, N. (2023). *Understanding student expression across higher Ed.: Heterodox Academy's Annual Campus Expression Survey*. Heterodox Academy

Received February 17, 2025
Revision received May 9, 2025
Accepted May 13, 2025 ■