Pipelines and Master Bedrooms: How Harmful is Harmful Language?

Michael A. Eskenazi[1], Scott W. Semenyna[2], & Christopher J. Ferguson[1]

Stetson University[1] & MacEwan University[2]

Address Correspondence to: Michael Eskenazi, Department of Psychology, Stetson University, 421 N. Woodland Boulevard, DeLand, FL 32723. Email: meskenazi@stetson.edu, Phone: 386-822-7398.

There are no conflicts of interest to report.

Abstract

Harmful language guides are becoming an increasingly common tool used by academic institutions, businesses, and professional organizations such as the American Psychological Association to reduce harmful effects of language, particularly for people from marginalized groups (APA, 2023). These guides provide a list of harmful words or phrases (e.g., pipeline, master bedroom) and alternative phrase (e.g., pathway, primary bedroom) to use in their place. Although these guides are well-intentioned, there is great disagreement as to whether harmful language causes adverse outcomes (Lilienfeld, 2017; Lilienfeld, 2020; Williams, 2020a Williams, 2020b). Thus, the purpose of the current pre-registered studies was to determine (1) how people view harmful language and (2) whether exposure to harmful languages causes adverse outcomes. In Study 1, 616 participants rated 175 harmful or alternative words or phrases. Results indicated that harmful language was viewed less favorably than alternative language; however, the vast majority of harmful language was rated on the favorable side of the scale. In Study 2, 334 participants were randomly assigned to read a short story that included 25 harmful words or phrases or the same story with harmful language replaced by alternative language. After exposure to the experimental or control condition, participants completed surveys measuring their anxiety, affect, and feelings of belonging. Results indicated that there were no differences on any psychological outcomes as a result of being exposed to harmful language. The findings from both studies call into question the concept of harmful language and the utility of harmful language guides.

Keywords: inclusive language; harmful language; personality

Disclosure Statement. All studies, measures, manipulations, participants, and exclusion criteria are included in this manuscript. All data, materials, and code for data analyses are included in this anonymized link.

Pipelines and Master Bedrooms: How Harmful is Harmful Language?

The *Elimination of Harmful Language Initiative* (EHLI) was published by the Stanford Chief Information Officer Council (CIOC) and People of Color in Technology (POC-IT) in 2020 with a goal "…to eliminate many forms of harmful language, including racist, violent, and biased (e.g., disability bias, ethnic bias, ethnic slurs, gender bias, implicit bias, sexual bias) language in Stanford websites and code." The list included more than 150 words or phrases across ten categories (e.g., ableism, ageism) along with recommended "unharmful" alternative words or phrases. For example, in the ableist category, "anonymous review" was proposed as an alternative to "blind review." The EHLI was met with significant criticism resulting in its removal from the Stanford website in 2023 (Gallagher, 2023). However, Stanford is not the first organization to promote the elimination of harmful language. The American Psychological Association (APA) publishes an *Inclusive Language Guide* with a goal "…to dismantle the destructive hierarchies that have marginalized people from equitable representation and participation in society" (APA, 2023). Similar to the EHLI, the APA guide provides a list of words or phrases with suggested alternatives. Other organizations provide similar guides such as The Sierra Club, The American Cancer Society, and the National Recreation and Park Association (Packer, 2023). Given the intense focus on harmful language in academia, industry, and public society, the purpose of the current study was to investigate claims that language can cause harm, in terms of anxiety, affect, or feelings of acceptance and inclusion.

Some words in American English are considered so harmful and offensive that they are rarely spoken or written but instead abbreviated euphemistically, such as the n-word (Henderson, 2003). Only 8% of Americans agree that there is nothing wrong with using this racial slur, with similar ratings across political and racial groups (Frankovic, 2018). Historically, pejoratives such as this one were used with the explicit intention to dehumanize, degrade, or marginalize someone based on their actual or perceived identity (Jeshion, 2021). Thus, it is not surprising that these pejoratives are considered taboo in American English. However, these most offensive pejoratives are not found in harmful language guides, and only a small percentage of words in these guides could be considered pejoratives, ethnic slurs, or racial epithets. Recent research provides a list of 396 taboo words in American English generated by native English speakers (Sulpizio et al., 2024) Only 14 words or phrases from the EHLI are found in that list, which provides further evidence that the list is primarily not composed of pejoratives, slurs, or racial epithets. The potential harm and perceived acceptability of the remainder of the list must be established empirically.

Because many of the words or phrases in these guides do not (presently) stand out as clearly offensive, they may fall better within the controversial concept of microaggressions, defined by the APA as “brief, verbal or nonverbal, behavioral, and environmental indignities that communicate derogatory attitudes or notions toward a different ‘other’” (APA, 2023; Sue et al., 2007). Importantly, intentionality is not a required component of a microaggression (Clay, 2017). The concept of microaggressions has received significant attention in academia and industry through research programs and training interventions to reduce microaggressions. However, the concept is not without controversy (Lilienfeld, 2017; Lilienfeld, 2020; Williams, 2020a Williams, 2020b). An initial comprehensive review of microaggressions raised five questions about the concept, the last of which is most relevant to the current research: do microaggressions cause adverse outcomes (Lilienfeld, 2017). An initial review reported that many studies found associations between microaggressions and adverse outcomes (e.g., anxiety, depression, stress); however, all but two of them were cross-sectional, and none were experimental (Lilienfeld, 2017). Thus, the ability to draw causal conclusions between the experience of microaggressions or harmful language and negative outcomes is limited.

A recent meta-analysis included 78 microaggression studies involving more than 18,000 participants (Lui & Quezada, 2019). This meta-analysis only included two experimental studies, both of which were unpublished dissertations (Prather, 2015; Wilson, 2014). Neither of these experimental studies observed significant differences in emotional reactions between participants who were randomly assigned to experience microaggressions compared to those who were not. In the overall meta-analysis, after controlling for publication bias, the bivariate association between microaggressions and various mental health outcomes was $r = .16$. An association of this size can be interpreted to be just below a level of practical significance (Ferguson, 2009). Lilienfeld (2020) interprets these findings to mean that the association between microaggressions and adverse outcomes is small to medium. In contrast, Williams (2019a) notes that "…it would be a mistake to… conclude that microaggressions cause minimal harm" (p. 8). Thus, much disagreement remains regarding the practical significance of these phenomena as well as the quality of research upon which such claims are made.

A small but informative experimental literature exists for a related concept—trigger warnings. Trigger warnings are cautions provided to viewers, listeners, or readers of various media, which explicitly aim "…to help individuals prepare for or avoid content likely to trigger memories or emotions relevant to past experiences" (Bridgland et al., 2023, p. 2). Similar to identifying microaggressions, and reducing the use of harmful language, the primary goal of trigger warnings is to reduce harm or distress flowing from exposure to harmful phrases or ideas that may provoke upsetting memories. Despite their widespread use, a meta-analysis of experiments show that trigger warnings produce no reduction (or increase) in affective response to the potentially harmful material compared to control groups receiving no such warning (Bridgland et al., 2023). These null effects are punctuated by potential pitfalls, given that trigger warnings tend to significantly elevate *anticipation* of distress. Trigger warnings thus fail to reduce distress when people actually confront difficult subject matter, and instead *produce* unpleasant anticipation (which quickly dissipates). Although people identify the emotions they will experience in future situations with reasonable accuracy, they nonetheless overestimate their realized intensity (Coteț & David, 2016). These findings dovetail with related experiments showing individuals can be primed to perceive ambiguous statements as harmful, especially if they have preexisting tendencies toward negative emotions (e.g., neuroticism; Bleske-Rechek et al., 2023). This may also be true for individuals who are high in trait victimhood, a construct that involves a tendency to focus one's identity on victim status, often feeling justified in subsequently acting aggressively toward others (Gabay et al., 2020). Similarly, those high in ethnocentrism may be quick to perceive slights against their own group, yet indifferent to slights against other groups (e.g., Vallone et al., 1985).

Given the limited experimental research on harmful language and adverse outcomes, it is important to investigate whether people who are exposed to harmful language experience more negative outcomes compared to a control group that is exposed to alternative language. Additionally, it is important to investigate whether individual differences in personality traits moderate this relationship (Bleske-Rechek et al., 2023; Lilienfeld, 2017, 2020). Two pre-registered studies were conducted to investigate this research question. The first study was descriptive in nature and aimed to measure how people rate both the harmful and non-harmful language from the EHLI. The second study was an experimental investigation of whether harmful language causes adverse outcomes. In the second study, participants were randomly assigned to read a story with 25 randomly selected harmful words or phrases or a control story that included alternatives to those harmful words or phrase. Reading comprehension and a series of psychological outcomes were measured along with various personality traits. Based on previous research, it was predicted that harmful language would not produce adverse outcomes except among those with personality traits associated with negative emotionality (i.e., neuroticism would moderate the relationship).

Method

Transparency and Openness

All data, materials, and code for analyses are available at this anonymized link. A pre-registration was created for Study 2, which includes a design plan, sampling plan, variables, and analysis plan. The pre-registration can be accessed at this anonymized link. Public and non-anonymized links will be made available upon publication. Data were analyzed using jamovi and R with the TOSTER package (Lakens, 2017; Caldwell, 2022) and jmv package (The jamovi Project, 2024).

Study 1

Participants

Participants were recruited from both a university participant pool and Amazon Mechanical Turk (MTurk), a crowdsourcing website commonly used for data collection. Data collected through MTurk comes with a risk of inattentive participants or non-human responses through bot accounts that challenge data quality (Chmielewski & Kucker, 2020). To address these challenges, several measures were put in place to ensure data quality. First, an additional fee was paid so that only "Master Workers" participated in this study. Master workers are participants who have submitted high quality responses across a wide range of studies. Second, before participants could begin the study, they were asked to type a word in all capital letters after listening to the word spoken in an audio file. Low-attentive participants who did not completely read the instructions to type the word in all capital letters were automatically filtered from the study. Finally, five made up words (e.g. sponglop) were intermixed with the real words and phrases, and participants who rated these made-up words highly on familiarity were excluded from analyses.

In total, data were collected from 782 participants. Of those participants, 166 were excluded from analyses for either rating the made-up words as highly familiar (*n* = 90) or for only responding to a few words or phrases, none of which were the made-up words (*n* = 76). The final sample size was 616, which included 219 participants from MTurk and 397 participants from the university participant pool. All participants were at least 18 years old, spoke English as their native language, and currently residing in the United States. The average age of the sample was 30 years old (*SD* = 15), and the sample was 65% female, 31% male, 2% non-binary or transgender, and 2% did not respond. The racial stratification of the sample roughly matched that of the United States with 75% of the sample identifying as White, 11% as Black, 16% as Latino, 5% as Asian, 1% as Arab, and 2%

as Native American. MTurk participants were paid $4.00USD for their participation and university participants were compensated with course credit. The research was approved by the Stetson University Institutional Review Board (IRB) and all participants agreed to an informed consent document prior to participating in the study.

Materials

A list of 175 pairs of words or phrases were created that included one harmful word or phrase and its suggested alternative word or phrase. All of the words or phrases from the EHLI were included in this list. Although the EHLI includes 161 pairs, 159 were included in this list because two were repeated (circle the wagons and gyp). An additional 11 pairs were included from the APA Inclusion Guide (e.g., pipeline/pathway, non-consensual sex/rape). Finally, five additional pairs were included based on current cultural controversies (e.g., breastfeeding/chestfeeding, pregnant woman/birthing person). As an attention check, five made up words were included for each participant (e.g., sponglop, flibberts).

Many words on this list have multiple meanings or could be interpreted either as a noun or a verb. To ensure that each word or phrase was interpreted as intended by the participant, an informative sentence frame was provided for each word or phrase. Sentence frames were created using ChatGPT-3 (OpenAI, 2022) with the prompt, "Put the word/phrase ____ in an informative sentence." When necessary, additional information was provided, such as what part of speech to use. When the harmful word was pejorative (e.g., retard), ChatGPT would refuse to create a sentence for that word. In such cases, the alternative word or phrase was used to create the sentence frame. The same sentence frame was used for both the harmful and alternative word. Three undergraduate research assistants read each sentence frame to ensure that it was consistent with the intended meaning of the word or phrase and that both the harmful and alternative word or phrase fit equally well in the sentence frame.

A set of seven questions were created to measures participants' awareness and appraisal of these words. Two questions were related to awareness and asked participants how familiar they are with the word or phrase, and whether they know the meaning. Three questions were related to appraisal, and asked participants how acceptable the word or phrase is, how the word makes them feel (from calm to anxious), and how willing they are to use the word or phrase. Finally, two filler questions were included and asked participants how pronounceable the word or phrase is and what the origin of the word or phrase is (Greek or Latin).

To ensure that each participant would only respond to either the harmful or alternative word or phrase from each pair, two counterbalanced lists were created. Asking participants to answer seven questions for each of the 175 words or phrases would likely have resulted in lower data quality through inattentiveness or participant fatigue. Therefore, each counterbalanced list was further split in half by selecting every other word or phrase so that participants would be exposed to words or phrases from each type of harmful category (e.g., ageism, ableism). Thus, participants were randomly assigned to rate words or phrases from one of four lists which contained both harmful and alternative words from as many categories as possible (the colonialism category only included one pair, so this category was only present in two lists).

Procedure

Participants first read an informed consent document, and if they agreed to voluntarily participate, they were directed to the first attention check. In this attention check, participants listened to an audio file that spoke the word "music" three times and were instructed to type the word that they heard in all capital letters only once. This attention check was used to ensure that participants read instructions completely and were at least initially attentive. Any participant who entered any text other than MUSIC was sent to a page indicating that they were ineligible to participate in the study because of a failure to follow instructions. Participants who passed the first attention check were then provided with instructions about how to complete the task. They were instructed to read each sentence and rate the word or phrase that was bolded in that sentence across each of the seven questions. Participants answered each question by moving a slider button on a scale that ranged from 0 to 100 with appropriate anchor words for each question. The randomizer function in Qualtrics was used to randomly assign participants to one of the four lists of stimuli. Stimuli were presented in random order. Five additional attention checks were presented at random order, which included made up words (e.g., sponglop, flibberts). Participants were expected to rate these words lower on familiarity (< 25) to indicate that they were paying attention. After completing the rating portion of the study, participants answered basic demographic questions. Finally, participants were compensated with $4.00USD through MTurk or with course credit through the university participant pool.

Results

The purpose of this first study was to gather descriptive ratings of words or phrases that are considered harmful and their alternative counterparts. The pre-registration of this study noted that it was descriptive in nature, and thus no inferential statistical analyses were pre-registered. However, to assist in the interpretation of some of the descriptive statistics, one set of unplanned analyses is reported to determine whether ratings of harmful words or phrases differed from their alternative counterparts.

A complete database of ratings on all 350 words and phrases is included in the supplemental materials. Table 1 summarizes the means and standard deviations across five variables on each word or phrase across two categories: awareness and appraisal. A Welch's t-test is also included to determine whether ratings were significantly different between harmful and alternative words. The awareness measure includes the degree of familiarity and knowledge of the meaning of the word or phrase. Participants' familiarity and knowledge ratings of harmful and alternative words or phrases was generally high with average scores ranging from 80 to 87. There were no statistically significant differences between the awareness of harmful and alternative words, and effect sizes were small. For the appraisal variables, acceptability, and willingness to use both the harmful and alternative language was generally high with average scores ranging from 70 to 87. Neither the harmful nor the alternative words or phrases were rated as particularly anxiety-provoking with average scores roughly in the middle ranging from 41 to 46. However, alternative words were rated as significantly more acceptable, less anxiety-provoking, and with more willingness to use them with medium to large effect sizes across all three variables. These results are presented in Table 1.

Table 1

Descriptive and Inferential Results from Study 1

	Harmful **M (*SD*)**	**Alternative** ***M* (SD)**	***t*-test**	**Cohen's *d***
Familiarity	80 (19)	83 (18)	*t*(346) = 1.56, *p* = .119	0.167
Knowledge	84 (18)	87 (16)	*t*(344) = 1.94, *p* = .053	0.207
Acceptability	75 (18)	87 (8)	*t*(250) = 8.21, *p* < .001	0.877
Anxiety	46 (15)	41 (13)	*t*(344) = -3.81, *p* < .001	-0.407
Willing to Use	70 (19)	82 (12)	*t*(296) = 7.17, *p* < .001	0.766

Note: Ratings range from 0-100.

We also created a composite appraisal score by averaging together the three scores on acceptability, anxiety, and willing to use on each word. The difference between the composite score for the alternative and harmful word provides a relative preference for the alternative or harmful word. The 20 pairs with the greatest preference for the alternative and the 20 pairs with the great preference for the alternative are listed in Table 2.

Table 2

Top 10 Pairs with the Greatest Preference for Alternative and Top 10 Pairs with the Greatest Preference for Harmful

Harmful	Alternative	Difference in Appraisal Score
Person with a cognitive disability	Retard	-36
Boring	Retarded	-35
Difficult problem	Tarbaby	-34
Haggled	Jewed	-30
Too Many Competing Ideas	Too many chiefs, not enough Indians	-30
Overworked	Slave labor	-30
Trans or Non-gender conforming folk	Tranny	-28
Person of multiple ethnicities	Half-breed	-27
Person with a disability	Cripple	-25
Rip off	Gyp (v)	-24
Birthing Person	Pregnant Woman	14
Chestfeeding	Breastfeeding	13
Denylist	Blacklist (-ed)	12
Latinx	Hispanic	11
Content note	Trigger warning	9
BIPOC (Black, Indigenous, and People of Color)	People of color	8
DevSecOps team	Yellow team	7
Hooked	Addicted	6
Person who has been impacted by	Victim	6
Relationship with an abusive person	Abusive relationship	5

Note: Scores can range from -100 to +100 with a positive score indicating preference for the harmful and a negative score indicating a preference for the alternative.

Discussion

The purpose of this first study was threefold: (1) to provide descriptive ratings of 175 words or phrases that have been labeled as harmful, (2) to determine whether harmful words or phrases are rated differently than their alternative counterparts, and (3) to provide descriptive norms for Study 2. The descriptive and inferential data should be interpreted wholistically. The inferential data indicate moderate to large differences between how people rate how acceptable, how anxiety-provoking, and how willing they are to use harmful words or phrases compared to their alternative counterparts. Simplistically, these medium to large effect sizes could be interpreted to indicate that harmful language is indeed viewed as harmful by participants. However, the descriptive data provide important context.

On the sliding scale from acceptable (100) to unacceptable (0), the average score for harmful language was well into the acceptable side (M = 75, SD = 17). Only 17 out of 175 harmful words or phrases received average scores into the unacceptable side (below 50). The majority of these words (n = 9) were pejoratives, ethnic, slurs, or racial epithets. Alternatively, 118 out of 175 harmful words or phrases received an average score *above* 70. The same pattern is true for willingness to use harmful language, with 25 out of 175 words or phrases receiving an average score below 50, and 96 out of 175 receiving an average score *above* 70. Thus, participants view the harmful language as generally acceptable and are willing to use them so long as they are not pejoratives. This conclusion is confirmed by a binomial test indicating that the number of response below 50 is much less than would be expected by chance ps < .001.

These findings raise questions about the concept of harmful language as outlined in harmful language guides. It is unlikely that the participants in this sample simply have a high tolerance for harmfulness. Rather, it is more likely that most of these words or phrases are incorrectly classified as harmful. Though harmful language guides are undoubtedly well-intentioned, there are some potential negative consequences to their use. Both situational and dispositional factors can influence how ambiguous language is perceived. Situationally, when people are warned about the potential for phrases to be harmful, they perceive ambiguous phrases as being more harmful even though they carry no specific negative meaning (Bleske-Rechek et al., 2023). Further, dispositional factors are associated with how ambiguous phrases are interpreted. Higher levels of neuroticism are associated with increased anxiety not only from intentionally hurtful statements, but also unintentionally hurtful statements and even positive statements (Bleske-Rechek et al., 2023). Thus, the suggestion that seemingly benign phrases can be harmful can lead to overinterpretation of harmfulness, especially in people with increased levels of neuroticism or cognitive distortions (Celniker et al., 2022).

Study 2

The purpose of the second study was to determine whether exposure to harmful language causes adverse outcomes. As noted in the introduction, there are many studies indicating a relationship between harmful language and negative outcomes; however, only two of them have been experimental, and neither yielded significant effects. Thus, there is a need for further experimental research on whether exposure to harmful language *causes* adverse outcomes.

Method

Participants

Similar to Study 1, participants were recruited from both a university participant pool and MTurk. All participants from the first study were excluded from participating in the second study. As with Study 1, an additional fee was paid so that only Master Workers were recruited from MTurk. The same audio attention check was used in this study along with seven additional attention checks that were embedded into each measure. In total, data were collected from 361 participants. Of those participants 27 were excluded from analyses—three participants who only agreed to the consent form, and 24 participants who failed at least three attention checks. Thus, a final sample of 334 participants were included in analyses. Prior to data collection, a power analysis was conducted, and the specifications of that power analyses were pre-registered. The power analysis was conducted using the TOSTER package in R (Lakens, 2017; Caldwell, 2022) for a power level of .80, an alpha level of .05, and a smallest effect size of interest (SESOI) of $d = .35$. This value was selected because it is roughly similar to the effect size ($r = .16$) observed in other research for the relationship between microaggressions (harmful language) and adverse outcomes (Lui & Quezada, 2019). The power analysis indicated that a minimum for 140 participants would be required in each group, thus the sample size of 334 exceeds the minimum sample size of 280.

The average age of the sample was 35 years old (*SD* = 16), and the sample was 59% female, 39% male, and 1% nonbinary. The racial stratification of the sample roughly matched that of the United States with 77% of the sample identifying as White, 13% as Black, 12% as Latino, 6% as Asian, <1% as Arab, and 2% as Native American. Seventeen percent of the sample identified as LGBTQIA with 64% of this subgroup identifying as bisexual, 31% as homosexual, 4% as pansexual, and 2% as asexual. Ten percent of the sample reported having a disability with 64% of those as a mental disability. A plurality of the sample (48%) identified as atheist, agnostic, or non-religious, 43% identified as some denomination of Christianity, 3% identified as spiritual, 2% as Jewish, 1% as Muslim, and 1% as a non-Abrahamic religious (e.g., Buddhist, Hindu). Finally, 55% of the sample was recruited from MTurk and 45% was recruited from a university participant pool.

Design

A between-subjects post-test only experimental design was used in this study. Participants were randomly assigned to one of two experimental conditions: harmful language or alternative language. To measure adverse outcomes broadly construed, post-test measures included state anxiety, feelings of inclusion and exclusion, and positive and negative affect. Moderating variables included neuroticism, conscientiousness, interpersonal victimhood, ethnocentrism, and the belief that words can harm.

Materials

Experimental Conditions

Twenty-five words or phrases were randomly selected from the Study 1 list. Each pair was assigned a number from 1 to 175 and the *sample* function in *R* was used to randomly select 25 numbers using the script, sample(1:175, 25, replace=FALSE). The norming data from Study 1 showed that these two conditions did not differ in mean familiarity (*t*(47) = 1.01, *p* = .32, *d* = 0.288) knowledge of the meaning (*t*(47) = 1.31, *p* = .20, *d* = 0.375), or the degree to which they were anxiety-provoking (*t*(47) = -0.94, *p* = .350, *d* = -0.270). However, the harmful language was rated as less acceptable (*t*(47) = 3.11, *p* = .003, *d* = 0.889), and people were less willing to use it (*t*(47) = 2.59, *p* = .012, *d* = 0.741). The 25 words or phrases came from a variety of harmful language categories including ableism (*n* = 3), APA inclusion guide (*n* = 2), cultural appropriation (*n* = 3), gender bias (*n* = 4), imprecise language (*n* = 4), institutionalized racism (*n* = 5), violence (*n* = 3), and additional considerations (*n* = 1).

A short story was created containing all 25 harmful words or phrases using ChatGPT 3.5 (OpenAI, 2023). Ten multiple-choice comprehension questions were also created using ChatGPT 3.5. Some questions were confusing, misleading, or contained no correct answers and were thus edited further. The control condition was created by eliminating all 25 harmful words or phrases and replacing them with their alternative counterparts[1]. The rest of the text of the story remained exactly the same. The harmful story included 482 words and the alternative story included 470 words.

State Anxiety

State anxiety was measured using the State Trait Anxiety Inventory (STAI; Spielberger et al., 1983). This measure contains 20 items each to measure state anxiety and trait anxiety; however, only state anxiety was measured in this study as we were interested in where exposure to harmful language affected the state of the participants' anxiety, not their trait. Participants responded to

[1] The harmful word "brave" had no proposed alternative, so the alternative condition included 24 words or phrases.

statements such as “I feel upset” on a 1 (“not at all”) to 4 (“very much so”) scale. This measure had excellent internal reliability α = .95 ω = .95.

Inclusion and Exclusion

The General Belongingness Scale (GBS) was used to measure feelings of acceptance/inclusion and rejection/exclusion, the two subscales of this measure (Malone et al., 2012). This measure has 12 items such as “I feel accepted by others” or “I feel like an outsider.” Participants respond on a 7-point Likert scale from “strongly disagree” to “strongly agree.” This measure had excellent internal reliability in both the acceptance/inclusion α = .93 ω = .93 and rejection/exclusion subscales α = .92 ω = .92.

Affect

Affect was measured using the Positive and Negative Affect Schedule (PANAS; Watson et al., 1988). The PANAS contains 20 items that are measured on a five-point scale from “very slight or not at all” to “extremely.” The subscale of positive affect includes items such as “proud: and “excited,” and the subscale of negative affect includes items such as “scared” and “distressed.” This measure had excellent internal reliability in both the positive α = .93 ω = .93 and negative subscales α = .91 ω = .92.

Personality Traits

The personality traits of neuroticism and conscientiousness were measured using the Big Five Inventory (BFI; John & Srivastava, 1999). This measure contains eight items to measure neuroticism and nine items to measure conscientious. Participants respond on a five-point Likert scale from strongly disagree to strongly agree. An example of an item to measure neuroticism is “gets nervous easily,” and an example of an item to measure conscientiousness is “tends to be organized.” This measure had good internal reliability in both the neuroticism α = .86 ω = .87 and conscientiousness subscales α = .89 ω = .89.

Ethnocentrism

The Multiethnic Climate Inventory (MCI) was used to measure ethnocentrism (Johnson & Johnson, 1996). The MCI is a 20-item scale with items such as “I don’t trust people of some other races or cultures.” Participants respond to each statement on a five-point Likert scale from strongly disagree to strongly agree. This measure had good to excellent internal reliability, α = .89 ω = .91.

Words Can Harm

The belief that words can harm was measured using the Words Can Harm Scale (WCHS; Bellet et al., 2018). This is a 10-item scale with items such as “a person might develop post-traumatic stress disorder or at least some of its symptoms from something they read.” Participants respond to each question on a 100-point scale from strongly disagree to strongly agree. This measure had excellent internal reliability, α = .92 ω = .92.

Interpersonal Victimhood

Interpersonal victimhood is defined as “an enduring feeling that the self is a victim across different kinds of interpersonal relationships” and was measured using the Tendency for Interpersonal Victimhood (TIV) scale (Gabay, et al., 2020). This is a 22-item measure that includes four subscales: need for recognition (e.g., “it is important to me that people who hurt me acknowledge that an injustice has been done to me”), moral elitism (e.g., “I give others much more than I receive from them”), lack of empathy (e.g., “people who are offended by me are only thinking of themselves”), and rumination (e.g., “days after the offense I am very preoccupied by the injustice done to me”). Participants respond on a seven-point Likert scale from strongly disagree to strongly agree. This measure had good to excellent internal reliability across all sub-scales: need for recognition α = .91 ω = .91, moral elitism α = .83 ω = .84, lack of empathy α = .83 ω = .84, and rumination α = .85 ω = .85.

Demographics

A large range of demographic variables were measured in an attempt to capture all identities that might be represented in harmful language guides. This included age, gender identity, race, sexuality, disability status, and religion.

Procedure

After reading a consent form, participants voluntarily agreed to participate in this study. The same audio attention check from Study 1 was used in this study, which prevented at least initially inattentive participants from participating. Participants who passed the first attention check were then randomly assigned to one of two experimental conditions: harmful language or alternative language. Participants were instructed to read the story for comprehension. To ensure that participants did not proceed without reading the story, a timer was used to prevent participants from proceeding to the next page until 60 seconds had passed. After reading the story, participants filled out each of the remaining surveys including the comprehension questions, STAI, GBS, PANAS, MCI, WCHS, TIV, and demographics. Each of those scales except for demographics contained another attention check. After completing all measures, participants were compensated with either $2.00USD if recruited from MTurk or 1 SONA credit if recruited from the university participant pool.

Results

Two sets of analyses were pre-registered to answer the two primary research questions: (1) does exposure to harmful language cause adverse outcomes, and (2) do certain personality traits moderate the relationship between harmful language and adverse outcomes. To answer the first research question, a series of Two One-Sided t-Tests (TOSTs) were conducted (Lakens, 2017). TOSTs are an alternative to the typical frequentist hypothesis testing approach and allow the researcher to draw the conclusion that two groups are equivalent. In this approach, a smallest effect size of interest (SESOI) is set, and two one-sided t-tests are conducted against the upper and lower bound SESOI. If both analyses fail, then it can be concluded that the value of the outcome variable for the two groups is essentially equivalent. This approach is particularly beneficial when the a priori hypothesis is that there will be no differences between the two groups because the frequentist hypothesis testing approach only allows for the failure to reject the null hypothesis, not the acceptance of the null hypothesis. Thus, TOSTs were used to determine whether exposure to harmful language results in any differences between the two conditions. The second set of analyses were conducted to determine if each of the moderating variables changed the relationship between the experimental condition and each outcome. A series of linear regressions were conducted with three predictors: the experimental condition, the moderating variable, and the interaction of the two.

TOST Analyses

Table 2 lists the TOST analyses for each of the outcome variables. As can be seen, the only outcome that was *not* equivalent was reading comprehension such that those in the harmful condition (M = 9.42, SD = 1.78) had higher comprehension than those in the alternative condition (M = 9.02, SD = 1.85), $t_{TOST}(332) = 1.188$, p = .118, 90% CI [-0.7282, -0.0718], d = -0.22. All other analyses resulted in significant TOSTs, which indicates equivalence between the two groups on that outcome variable.

Table 2

TOST Analysis Outcomes and Interpretations.

Outcome Variable	TOST	Interpretation
Comprehension	$t(332) = 1.188, p = .118$	Not Equivalent
Harmful	$M = 9.42, SD = 1.78$	
Alternative	$M = 9.02, SD = 1.85$	
State Anxiety	$t(332) = -3.125, p = .001$	Equivalent
Harmful	$M = 37.23, SD = 12.73$	
Alternative	$M = 37.34, SD = 12.37$	
Inclusion	$t(332) = -3.129, p = .001$	Equivalent
Harmful	$M = 23.60, SD = 5.46$	
Alternative	$M = 23.64, SD = 5.11$	
Exclusion[2]	$t(327.13) = -3.081, p = .001$	Equivalent
Harmful	$M = 12.95, SD = 6.61$	
Alternative	$M = 13.03, SD = 5.85$	

[2] This assumption of homogeneity of variance was violated, so this analysis was run with var.equal set as false.

Positive Affect	$t(332) = 2.118$, $p = .0175$	Equivalent
Harmful	$M = 28.59$, $SD = 10.26$	
Alternative	$M = 27.43$, $SD = 9.34$	
Negative Affect	$t(332) = -2.882$, $p = .002$	Equivalent
Harmful	$M = 14.26$, $SD = 6.45$	
Alternative	$M = 14.48$, $SD = 6.25$	

Moderation Analyses

Three significant moderation analyses were observed out of 42 moderation analyses conducted (six outcomes and seven moderators), thus these results should be interpreted as suggestive but in need of replication (Benjamin et al., 2018). The first significant moderation was between the interpersonal victimhood subscale of moral elitism and condition on negative affect, $R^2 = .095$, $F(3, 330) = 11.50$, $p < .001$, $B = 0.20$, $p = .028$. As moral elitism increased, negative affect also increased, but this positive relationship was stronger in the harmful language condition. In other words, exposure to harmful language exacerbates the relationship between moral elitism and negative affect. The second significant interaction demonstrated the same pattern of results but with state anxiety as the outcome, $R^2 = .121$, $F(3, 330) = 15.10$, $p < .001$, $B = 0.37$, $p = .037$. In this interaction, the positive relationship between moral elitism and state anxiety was also exacerbated by being in the harmful language condition. Finally, the third significant interaction was between the interpersonal victimhood subscale of need for recognition and condition on positive affect, $R^2 = .031$, $F(3, 330) = 3.46$, $p < .017$, $B = -0.26$, $p = .037$. In the alternative condition, there was no relationship between need for recognition and positive affect; however, in the harmful condition, as need for recognition increased, positive affect decreased.

Unplanned Analyses

One set of unplanned analyses was conducted that was not part of the pre-registration or central to the research question but can provide additional insights after the planned analyses. The purpose of these analyses was to determine whether participants who are part of a marginalized group were more affected by the harmful language condition than participants who are not part of a marginalized group. One might hypothesize that the lack of differences observed in the TOST analyses was because adverse outcomes would only be present for participants from a marginalized group. To conduct this analysis, a new variable was created to classify participants into a marginalized or non-marginalized group. Though this group could be defined in many ways, we included the following participants in this group to reasonably capture what marginalization means in American society: women, any participant who identified as gay, lesbian, bisexual, transgender or non-binary, any participant who identified as a race other than white, any participant with a physical or mental disability, and any participant from a religious minority (e.g., Jewish, Muslim, Hindu, Buddhist). This classification process resulted in 256 participants in the marginalized condition and 77 participants in the non-marginalized condition. A moderation analysis was conducted using linear regression that included three predictors: condition, marginalized group, and the interaction between the two for each of the outcome variables. The interaction between these two variables was not significant for any outcome.

Discussion

The purpose of Study 2 was to determine whether exposure to harmful language causes adverse outcomes and whether the relationship between exposure to harmful language and those outcomes is moderated by various personality traits. Results indicated that exposure to harmful language does not cause any adverse outcomes. There was a small benefit in comprehension in the harmful language condition, but this effect was quite small. Personality traits associated with interpersonal victimhood (e.g., moral elitism and need for recognition) interacted with the experimental condition. For those in the harmful language condition, the relationship between interpersonal victimhood and adverse outcomes was stronger than for those in the alternative condition. Overall, results are consistent with existing experimental research indicating null effects from exposure to harmful language. Implications of these results are discussed in the general discussion.

General Discussion

Across two studies, we provide evidence that harmful language is not viewed as particularly harmful, nor does it result in adverse outcomes. We also find no evidence that participants from marginalized groups are more negatively impacted by exposure to harmful language than participants who are not from marginalized groups. The only participants who indicated any evidence of an increase in adverse outcomes from exposure to harmful language were those with higher levels of moral elitism and need for recognition. The results raise questions about the concept of harmful language and the tangible benefits of this language reform.

Concept Creep and Harmful Language

The findings of these studies are in line with a recent trend in psychology referred to as "concept creep" (Haslam, 2016). Concept creep refers to the process by which the definition of a negative concept (e.g., harmful language) expands to include elements that were not part of the concept's original definition. This form of semantic inflation has occurred for a variety of topics in psychology. For example, the concept of addiction initially included necessary components of powerlessness, dependency, and compulsion to use a specific substance despite negative consequences for the addict. The first semantic expansion of addiction included addiction to behaviors such as gambling, which dropped the necessity of a substance. The second semantic expansion of addiction included "soft" addictions such as procrastination or gossiping (Wright, 2006), which dropped the necessity of powerlessness, dependency, and compulsion. Some forms of concept creep may be beneficial in that they simply represent societal progress or technological advancements that necessitate an expansion of the original definition. For example, the semantic expansion of addiction to remove substance use has contributed to research on effective treatments for addictive behaviors such as gambling (Haslam et al., 2020). However, concept creep can also come with negative consequences such as the reconceptualization of mildly bad behaviors as pathological or the creation of controversial new disorders with shaky foundations, such as addiction to video games (Bean et al., 2017) or pornography (Ley et al., 2014).

The concept of harmful language is likely going through its own form of concept creep. A narrow original definition of harmful language is language that is intentionally designed to dehumanize, degrade, or marginalize someone based on their actual or perceived identity. This definition includes pejoratives, slurs, and epithets used intentionally towards members of a particular group. Modern harmful language guides have semantically expanded the concept to remove intent, dehumanization, degradation, and marginalization. The reconceptualized and expanded concept of harmful language can be defined as a word or phrase that has violent, racist, sexist, or other discriminatory origins or words that can be loosely interpreted to have those meanings even if there is no evidence to support that conclusion. For example, the term "master bedroom" is interpreted as harmful because it is claimed that it originally referred to a slave master's bedroom. However, there is no evidence of this connection as the word first appeared in a 1926 Modern Homes Catalog well after the end of slavery in the United States (Franklin, 2020, NYT).

As with the concept creep of addiction, there may be some positive outcomes of the concept creep of harmful language. As society becomes more diverse and inclusive, we might realize that the use of certain words does not make everyone feel welcome in that society and choose to stop using them. In 2009, the *Spread the Word to End the Word* campaign began with the goal to end the use of the words retard and retarded. Fifteen years later, data from our own study indicate that this campaign was a success given that these words evidenced the lowest acceptability rating. The inclusion of retard and retarded as harmful words can be considered concept creep because most of the time when people use these words, they are not referring to people with intellectual disabilities (Siperstein et al., 2010) as would be required in the original narrow definition of harmful language. However, we are not arguing that this form of concept creep is bad. This example is definitively good and represents a positive outcome of this expanded definition.

Other forms of harmful language concept creep are more questionable with the potential for negative outcomes. The APA's *Inclusive Language Guide* lists the word pipeline as "offensive to Indigenous communities as a result of oil companies transporting crude oil through the sacred lands of Native Americans or Alaska Natives and contaminating their water supply" (APA, 2023). Others have gone so far as to say that the word pipeline is a pejorative (Stern et al., 2023). This categorization of pipeline as harmful further expands the concept of harmful language to include metaphors whose meaning (in this case, pathway) has no direct connection to the group that is supposedly harmed. This further semantic expansion raises the likelihood for negative outcomes such as social conflict, moral typecasting, and polarization (Haslam et al., 2020).

In this sample, the average acceptability score for the word pipeline was 89 (*SD* = 16) and 88% of participants gave it a rating above 70. Thus, the word pipeline is not yet viewed negatively. However, the concept creep that turns pipeline into a pejorative or a harmful word has the possibility to create social conflict between people who have a broad or narrow view of harmful language. If the word pipeline is a harmful pejorative, how is someone supposed to respond when they hear it? As Haslam et al. (2020) note, "These disagreements are likely to lead to conflicting opinions about whether problematic behavior has occurred, how severe it is, what if anything should be done to punish 'perpetrators,' whether 'victims' have legitimate standing as such, and whether institutional intervention is required to ensure justice for them" (p. 272). It is not clear whether marginalized communities benefit from semantic inflation, but the potential for social conflict abounds.

The potential for social conflict only grows as one considered how many words or phrase could be considered harmful under an expanded definition. One does not need to spend much time thinking to realize that current harmful language guides only include a small fraction of potentially harmful language. Using the word "hysterical" to mean humorous could be considered harmful because of the misogynist origins of hysteria. Saying "I hear you" to mean "I understand" could be considered ableist as it excludes deaf or hard of hearing people. The numbers 14 and 88 could be considered harmful to Jews as they are used by neo-nazis to represent antisemitic phrases. The first author's own last name could be considered harmful because it contains the word "nazi." Harmful language guides often include a category of violent language to avoid, which would exponentially expand to include "shoot yourself in the foot," "when push comes to shove," "get away with murder," "punchline," "shoot from the hip," "twist your arm," "take a stab at it," "give it your best shot," "slap in the face," "add insult to injury," "roll with the punches," "under the gun," and so many others. As one can see, an expanded definition of harmful language can easily turn everyday speech into a moral minefield.

Moral typecasting offers another area of concern that could result from harmful language concept creep. Moral typecasting involves moral intuitions/reasoning reflexively viewing those who cause (perceived) harms as having agency and acting with intention, but recipients of those actions as victims (Gray & Wegner, 2009; Reynolds et al., 2020). Harmful language concept creep may lead to moral typecasting that denigrates those who use this language, viewing them as worthy of blame while simply presuming harm to victims. This is warranted in instances where individuals use pejorative invectives, as noted above, but our data show that most of the harmful terminology was viewed as acceptable (Study 1) and did not create measurable harm (Study 2). As such, there is little need to cast those who use terms such as pipeline, master bedroom, breastfeeding, and blind review, as seeking to cause harm and hence needing to be punished socially.

Semantic Change and Harmful Language

Concept creep's semantic inflation is a subtype of a broader category called semantic change. Semantic change is the process by which a word sheds its former meaning and adopts an entirely new one (Blank, 1999). The word "gay" has gone through several semantic changes from its original meaning of happy, to a more negative meaning of hedonistic, to a categorical term for homosexuals, and most recently to an insult to mean lame or boring (Lalor & Rendle-Short, 2007). These semantic changes mean that "gay" can be considered polysemous, because the word maintains more than one of these meanings. Which meaning a hearer infers will depend on context, speaker, assumed intent, and other features of an interaction. However, other words that have gone through semantic changes are now interpreted completely independently of their original meaning. For example, the word "nice" originally meant simple or ignorant but now means friendly or approachable (Fortson, 2003).

The concept of semantic change can be applied to many words or phrases in harmful language guides and can help to explain why we did not observe adverse outcomes from exposure to harmful language. One of the words in the EHLI is "lame," which initially referred to physical disability but is now colloquially used to mean uncool (Aaron, 2010). The EHLI notes that using the word lame can "trivialize the experience of people living with disabilities" in an apparent attempt to undo the semantic change that has already occurred and reinstate its original meaning. When a word goes through semantic change, it can continue to change its meaning over time, but there is no evidence that it ever reverts to its original meaning. If the EHLI were to be successful, this would represent the first semantic reversion. The lack of semantic reversion can also partially explain why exposure to harmful language did not cause adverse outcomes. When someone hears or reads the word lame, the semantic concepts that are activated in mind are tied to the current understanding of the word rather than its original meaning as word associations change across time based on current usage (Laurino et al., 2023). Thus, there is no causal mechanism to predict adverse outcomes from words that formerly had a harmful association but no longer do.

Even if this semantic reversion were to be successful, it is not clear what the benefit would be. Data from Study 1 indicate that English speakers do not view the word lame as unacceptable, anxiety-provoking, or that they would be unwilling to use it. Data from Study 2 further indicate that harmful language does not cause any adverse outcomes for people in marginalized communities. Thus, the attempt to reassociate "lame" with "disability" at best does not appear to offer any benefits, especially for people living with disabilities, and at worst may cause harm to some people based on their personality traits.

The only participants in Study 2 who experienced any adverse outcomes from the harmful language condition were those who had increased levels of moral elitism and need for recognition. Moral elitists hold the view that they have "immaculate morality" while other people are immoral (Gabay et al., 2020). People with a higher need for recognition feel an increased desire for their own perceived victimhood to be acknowledged and empathized with. When exposed to harmful language, people with higher levels of these traits may be more negatively impacted through perceived moral violations and an absence of recognition. Thus, as seemingly benign language (e.g., pipeline) is promoted as being harmful, people with certain personality traits may experience more negative outcomes. This conclusion is in line with other research indicating that people with higher levels of neuroticism experienced greater anxiety from reading ambiguous or positive statements after the suggestion that they could be harmful (Bleske-Rechek, 2023).

Limitations and Context

The findings of these studies must be interpreted with important contextual constraints. First, harmful language in this study refers only to types of words that are commonly found in harmful language guides. Though there are some pejoratives in these guides, our conclusions refer more to words and phrases with more ambiguous interpretations (e.g., pipeline, master bedroom). We are in no way drawing any conclusions about the harmfulness or impact of explicitly racist language. The results of this study should not be used as an excuse for those who seek to use racial slurs, ethnic epithets, taboo language, or other pejoratives. Our findings do not apply to that category of words or phrases. Second, the results of Study 2 are cross-sectional and refer to a single exposure to 25 harmful words. Although this study was not directly about microaggressions, one could argue from a microaggression perspective that no adverse outcomes were observed because the negative effects build up cumulatively over multiple exposures (Sue et al., 2008). This is a valid critique, and we encourage the use of longitudinal experimental methods to investigate this empirical question. Nevertheless, the results provide important baseline evidence about the effects of a single exposure to harmful language specifically in the context of harmful language guides. Third, a broad approach was used in Study 2 that included a diverse sample and harmful words from a wide range of categories (e.g., ableism, ageism, sexism, etc.). It is possible that harmful language effects only emerge with a more homogenous sample and harmful language that is specifically about that group. Thus, future research should investigate whether adverse outcomes occur within specific harmful language categories with samples of participants from that category.

Conclusions

Language is a dynamic system with features that change across time. Our phonological forms, orthographic forms, syntactic structures, and semantics are not only different from Middle and Old English of 500 and 1,000 years ago but can change within our own lifetimes. Many languages have official academies that oversee and make rules about their language such as the Académie Française (French Academy) and the Real Academia Española (Royal Spanish Academy). However, English is unique in that it has no official governing body. Thus, it is up to us collectively to decide how our language works and what our words mean. As our society has progressed, our language has progressed along with it. We have shunned offensive words to euphemism (e.g., the n-word), we have abandoned words whose phonological form seems offensive even though their semantic form is not (e.g., niggardly; McWhorter, 2000), and we are gradually adjusting our grammatical forms to respect people with different identities (e.g., singular they/them; Gernsbacher, 1997). While the initial concept creep of harmful language is undoubtedly good and mirrors societal progress, further top-down attempts at advanced semantic inflation of harmful language offers few benefits and many risks.

Open Practices

All data, materials, and code for analyses are available at this anonymized link. A pre-registration was created for Study 2, which includes a design plan, sampling plan, variables, and analysis plan. The pre-registration can be accessed at this anonymized link. Public and non-anonymized links will be made available upon publication. Data were analyzed using jamovi and R with the TOSTER package (Lakens, 2017; Caldwell, 2022) and jmv package (The jamovi Project, 2024).

References

Aaron, J. E. (2010). An awkward companion: Disability and the semantic landscape of English lame. *Journal of English Linguistics*, *38*(1), 25-55.

American Psychological Association. (2023). *Inclusive language guide*. https://www.apa.org/about/apa/equity-diversity-inclusion/language-guidelines

American Psychological Association. (2023). Microaggression. In *APA dictionary of psychology.* https://dictionary.apa.org/microaggression

Baron, R., & Richardson, D. (1994). *Human aggression.* New York: Plenum Press.

Bean, A. M., Nielsen, R. K., Van Rooij, A. J., & Ferguson, C. J. (2017). Video game addiction: The push to pathologize video games. *Professional Psychology: Research and Practice*, *48*(5), 378-389.

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.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6-10.

Blank, A. (1999). Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. *Historical Semantics and Cognition*, *13*(6).

Bleske-Rechek, A., Deaner, R. O., Paulich, K. N.., Axelrod, M., Badenhorst, S., Nguyen, K., Seyoum, E., & Lay, P. S. (2023). In the eye of the beholder: Situational and dispositional predictors of perceiving harm in others' words. *Personality and Individual Differences*, *200*, 111902. https://doi.org/10.1016/j.paid.2022.111902

Bridgland, V. M. E., Jones, P. J., & Bellet, B. W. (2023). A meta-analysis of the efficacy of trigger warnings, content warnings, and content notes. *Clinical Psychological Science*. Advanced online publication. https://doi.org/10.1177/21677026231186625

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*, 111243.

Caldwell, A. R. (2022). Exploring equivalence testing with the updated TOSTER R Package. *PsyArXiv*. doi:10.31234/osf.io/ty8de

Chmielewski, M., & Kucker, S. C. (2020). An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. *Social Psychological and Personality Science*, *11*(4), 464-473. https://doi.org/10.1177/1948550619875149

Clay, R. (2017). Did you really just say that? *Monitor on Psychology, 48*(1), 46.

Coteț, C. D., & David, D. (2016). The trust about predictions and emotions: Two meta-analyses of their relationship. *Personality and Individual Differences*, *94*, 82–91. http://dx.doi.org/10.1016/j.paid.2015.12.046

Ferguson, C. J. (2016). An effect size primer: A guide for clinicians and researchers. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (4th ed., pp. 301–310). American Psychological Association. https://doi.org/10.1037/14805-020

Fortson, B. W. (2003). An approach to semantic change. In Joseph, B. D., & Janda, R. D. (Eds.), *The handbook of historical linguistics* (pp. 648-666). Blackwell.

Franklin, S. (2020). The biggest bedroom is no longer a 'master.' *New York Times.* https://www.nytimes.com/2020/08/05/realestate/master-bedroom-change.html

Frankovic, K. (2018). When it comes to the n-word, most see it as offensive whoever says it. *YouGov*. https://today.yougov.com/politics/articles/21420-when-it-comes-n-word-most-see-it-offensive-whoever

Gabay, R., Hameiri, B., Rubel-Lifschitz, T., & Nadler, A. (2020). The tendency for interpersonal victimhood: The personality construct and its consequences. *Personality and Individual Differences*, *165*.

Gallagher, S. (2023). *Update on Elimination of Harmful Language Initiative in Stanford's IT Community*. Stanford University. https://itcommunity.stanford.edu/news/update-elimination-harmful-language-initiative-stanfords-it-community

Gernsbacher, M. A. (1997). Generic pronominal anaphora: The case of the English singular they. *Verbum (Nancy, France)*, *19*(1-2), 67-84.

Gray, K., & Wegner, D. M. (2009). Moral typecasting: divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, *96*(3), 505-520.

Haslam, N. (2016). Concept creep: Psychology's expanding concepts of harm and pathology. *Psychological Inquiry*, *27*(1), 1–17. https://doi.org/10.1080/1047840X.2016.1082418

Haslam, N., Dakin, B. C., Fabiano, F., McGrath, M. J., Rhee, J., Vylomova, E., ... & Wheeler, M. A. (2020). Harm inflation: Making sense of concept creep. *European Review of Social Psychology*, *31*(1), 254-286.

Henderson, A. (2003). What's in a slur? *American Speech, 78*(1), 53-74.

Jeshion, R. (2021). Varieties of pejoratives. In Khoo, J., Sterken, R. K. (Eds.), *The Routledge handbook of social and political philosophy of language,* Routledge.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), Handbook of personality: Theory and research (2nd ed., pp. 102-138). New York: Guilford.

Johnson, P. E., & Johnson, R. E. (1996). The role of concrete-abstract thinking levels in teachers' multiethnic beliefs. *Journal of Research & Development in Education*, *29*, 134-140.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355-362.

Lalor, T., & Rendle-Short, J. (2007). 'That's so gay': a contemporary use of gay in Australian English. *Australian Journal of Linguistics*, *27*(2), 147-173.

Laurino, J., De Deyne, S., Cabana, Á., & Kaczer, L. (2023). The pandemic in words: Tracking fast semantic changes via a large-scale word association task. *Open Mind*, *7*, 221-239.

Ley, D., Prause, N., & Finn, P. (2014). The emperor has no clothes: A review of the ‘pornography addiction’ model. *Current Sexual Health Reports*, *6*, 94–105. https:doi.org/10.1007/s11930-014-0016-8

Lilienfeld, S. O. (2017). Microaggressions: Strong claims, inadequate evidence. *Perspectives on Psychological Science*, *12*(1), 138–169.

Lilienfeld, S. O. (2020). Microaggression research and application: Clarifications, corrections, and common ground. *Perspectives on Psychological Science*, *15*(1), 27-37.

Lui, P. P., & Quezada, L. (2019). Associations between microaggression and adjustment outcomes: A meta-analytic and narrative review. *Psychological Bulletin, 145*(1), 45-78. https://doi.org/10.1037/bul0000172

Malone, G. P., Pillow, D. R., & Osman, A. (2012). The general belongingness scale (GBS): Assessing achieved belongingness. *Personality and Individual Differences*, *52*(3), 311-316.

McWhorter, J. H. (2000). *Losing the race: Self-sabotage in Black America*. Simon and Schuster.

OpenAI. (2022). *ChatGPT-3* [Large language model]. https://chat.openai.com

OpenAI. (2023). *ChatGPT-3.5* [Large language model]. https://chat.openai.com

Packer, G. (2023). The moral case against equity language. *The Atlantic*. https://www.theatlantic.com/magazine/archive/2023/04/equity-language-guides-sierra-club-banned-words/673085/

Prather, C. (2015). *Nice dissertation, for a girl: Cardiovascular and emotional reactivity to gender microaggressions.* (Doctoral dissertation). Retrieved from ProQuest.

Reynolds, T., Howard, C., Sjåstad, H., Zhu, L., Okimoto, T. G., Baumeister, R. F., Aquino, K., & Kim, J. (2020). Man up and take it: Gender bias in moral typecasting. *Organizational*

Behavior and Human Decision Processes, *161*, 120–141. https://doi.org/10.1016/j.obhdp.2020.05.002

Spielberger, C. D. (1983). *State-Trait Anxiety Inventory for Adults (STAI-AD)* [Database record]. APA PsycTests. https://doi.org/10.1037/t06496-000

Siperstein, G. N., Pociask, S. E., & Collins, M. A. (2010). Sticks, stones, and stigma: A study of students' use of the derogatory term "retard". *Intellectual and Developmental Disabilities*, *48*(2), 126-134.

Sue, D. W., Capodilupo, C. M., & Holder, A. (2008). Racial microaggressions in the life experience of Black Americans. *Professional Psychology: Research and Practice*, *39*(3), 329-336.

Sue, D. W., Capodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A. M. B., Nadal, K. L., & Esquilin, M. (2007). Racial microaggressions in everyday life: Implications for clinical practice. *American Psychologist, 62*(4), 271–286. https://doi.org/10.1037/0003-066X.62.4.271

Stern, N., Vela, M., & Nakae, S. (2023). Eliminating the pipeline metaphor in framing workforce equity for American Indian and Alaska Native communities. *JAMA Network Open*, *6*(7), e2321926-e2321926.

Sulpizio, S., Günther, F., Badan, L., Basclain, B., Brysbaert, M., Chan, Y. L., ... & Marelli, M. (2024). Taboo language across the globe: A multi-lab study. *Behavior Research Methods*, *56*, 3794–3813.

Vallone, R. P., Ross, L., & Lepper, M. R. (1985). The hostile media phenomenon: Biased perception and perceptions of media bias in coverage of the Beirut massacre. *Journal of Personality and Social Psychology, 49*(3), 577–585.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063-1070.

Williams, M. T. (2020a). Microaggressions: Clarification, evidence, and impact. *Perspectives on Psychological Science*, *15*(1), 3-26.

Williams, M. T. (2020b). Psychology cannot afford to ignore the many harms caused by microaggressions. *Perspectives on Psychological Science*, *15*(1), 38-43.

Wilson, M. (2014). *Neurocognitive and whole-system reactions to covert racial discrimination-induced stress.* (Doctoral dissertation). Retrieved from ProQuest.

Wright, J. (2006). *The soft addiction solution: Break free of the seemingly harmless habits that keep you from the life you want*. New York: J. P. Tarcher/Penguin.