Prepublication version. This paper is now published in Feminism & Psychology
https://doi.org/10.1177/09593535221104874

Contributions to reducing online gender harassment:

Social re-norming and appealing to empathy as tried-and-failed techniques

Lilith A. Whiley[1], Lukasz Walasek[2], and Marie Juanchich[3]

1 Birkbeck University of London, UK

2 University of Warwick, UK

3 University of Essex, UK

Corresponding author: Dr. Lilith A. Whiley, Birkbeck University of London,

l.whiley@bbk.ac.uk

Abstract

Inspired by similar methods shown to be effective in reducing online racist harassment, we designed two tweets aimed at reducing online gender harassment. Our interventions were based on the principles of social re-norming and appealing to harassers' empathy. In a sample of 666 Twitter users engaging in sexist or misogynist tweeting, we found that our intervention tweets did not reduce the number of sexist slurs or sexist users, either 7 days or 31 days after being sent. Our attempts also affected neither the valence nor the arousal of subsequent tweets posted by our sample of Twitter users. We discuss the conceptual, methodological, and ethical challenges associated with activist research aimed at reducing online gender harassment and discuss some of the implications of our attempts to do so.

Keywords:

Online gender harassment; sexism; misogyny; social norms; empathy; social media.

Despite being unlawful under the Equality Act (2010), 71% of women in the United Kingdom (UK) still experience gender harassment in shared spaces (UN Women UK, 2021). Fifty-four percent of women hear men wolf-whistling at them and 39% are called derogatory names; 23% are even groped (Action Aid, 2016). Online behaviour mirrors these experiences: up to 45% of gender harassment is done virtually (Rights of Women, 2021). Indeed, online gender harassment reflects the wider misogynist treatment of women; it is "firmly grounded in the material realities of women's everyday experiences of sexism in patriarchal society" (Megarry, 2014, p. 49). Online space, generally speaking, is a heteronormative and hegemonically masculine space (Drakett et al., 2018). Han (2018) describes it as a space of *toxic* masculinity, "of technological privilege where the masculine elite dominates the archetypical passive sexualised woman" (Lock et al., 2018, p.7). To illustrate, tweets that blame victims and shame rape survivors have more followers and retweets than those supporting the women who have experienced sexual violence (Stubbs-Richardson et al., 2018). Over 400 000 sexist slurs are posted on Twitter daily (Felmlee et al., 2020). Harassment is often based on accusing women of 'failing' to meet dominant patriarchal norms of femininity (i.e., thin, young, innocent, passive). Offenders typically attack women's physical appearance, their intelligence, and their age. Women consistently receive death threats and even calls for rape (Chen et al., 2020). They are harassed on online dating sites (Thompson, 2018) and are re-victimised in image-based sexual abuse, while abusers hide behind the protective cloak of online anonymity (Uhl et al., 2018).

The shift to flexible and remote working has exacerbated women's suffering in this regard; for example, in a recent report by Rights of Women (2021) one woman shared her perception that offenders are now in one's home and bedroom: "I feel my privacy has been invaded and nowhere is safe". Indeed, the purpose of sexual harassment is to violate someone's dignity, to intimidate, degrade, and humiliate them, and create a hostile environment (Citizens Advice, 2021). Up to 61% of women who experience online gender harassment have trouble sleeping afterwards, 55% experience anxiety, and 67% feel apprehensive about using social media again (Amnesty International, 2017). Women become more cautious about what they post and “keep quiet so as to reduce abuse” (Adams, 2017, p. 7), actively avoid voicing their opinions in online discussions (Chadha et al., 2020), and “[watch] over [their] shoulder in cyberspace” (Chen et al., 2020, p. 887). For these reasons, some women decide to leave social networking altogether (Citron, 2014).

Consequently, online gender harassment limits women’s equal participation in online communities and social networks (Megarry, 2014). Constraints and limitations are imposed on women’s freedom in both the physical world and the online one (Vera-Gray, 2017), which can also have a profound impact on women’s livelihood either directly or indirectly impacting on women’s professional lives in what Jane (2018) terms ‘economic vandalism’. Online gender harassment is insidious and proliferates in almost every aspect of women’s lives; it is almost “the destruction” of personhood (Chen et al., 2020, p. 884). In these ways, online gender harassment becomes another means by which women’s behaviour is monitored, policed, and contained, especially when they are perceived to be breaching patriarchal hegemonic social norms.

Dealing with and responding to online gender harassment

What channels exist for dealing with online harassment? Online gender harassment can be reported to the UK police as either ‘harassment’ or ‘malicious communications’ (Met Police, 2021). It can also be reported directly to the social media platform, but despite *Facebook*, *Twitter*, and *YouTube* agreeing a Code of Conduct on Countering Illegal Hate Speech Online with the European Commission, 43% of women in the UK still think that social media giants’ responses are inadequate in addressing online gender harassment (Amnesty International, 2020). Indeed, despite several complaints, few if any posts are deleted. Responses also range from automated emails to speedy investigations exonerating the offenders. Certainly, social media giants lament the difficulty of regulating 'hate speech', while citing 'freedom of speech' as one reason for their limited intervention (House of Commons, 2017). Interestingly, male Internet users think that 'censorship' is their greatest threat, whereas women believe it to be 'privacy' (Herring, 2003). Many of the recommended courses of action such as 'unfriend the person', 'block the person', and 'don't retaliate' (House of Commons, 2017), do little in the way of giving women resources to *actually* respond to offenders. Indeed, the advice to ignore the problem is harmful. Mallett et al. (2019) found that when women did not confront instances of harassment, it desensitised them and increased their tolerance for future abuse.

There are, however, significant risks to resisting harassment. Women can experience psychological harm such as increased anxiety and depression, and decreased wellbeing (Cortina & Magley, 2003). They can be further victimized by "doxxing" whereby their personal information is distributed online to cybermobs (e.g., Gamergate) (Eckert and Metzer-Riftkin, 2020). Harassers may even solicit actual physical violence from their communities and followers (e.g., INCELS) (Regehr, 2020). Rebecca Watson, for example, was horrendously gender "trolled" in retaliation for posting a video about being propositioned. Insults were posted on her *YouTube* account, her *Wiki* page was vandalised, and fake *Twitter* accounts were created in her name and used to post vile messages (Mantilla, 2013). Indeed, rebuking online gender harassment, as with in-person harassment, can be dangerous for women.

Nevertheless, women should not have to put up with online gender harassment and have the right to speak up without further victimisation. Budding feminist scholarship on interpersonal management of harassment and resistance shows that women do have a range of (tentative) strategies in their repertoire for responding to abuse. Roberts, Donovan, and Durey (2019, p. 334) assert, and we agree, that "by acting agentically women challenge patriarchal ideals because they are both problematising the perpetrator's behaviour and their strategies of resistance can be seen as examples of social change". In their analysis of 1034 survey responses, they found examples of women fighting back verbally, physically, and collectively in offline spaces by intervening to protect other would-be targets of harassers. Others examples of agentic behaviours include changing routines and actively leaving uncomfortable situations. Similarly, women behave agentically in online spaces to resist online gender harassment. In online games, for example, women may choose to conceal their gender and also avoid playing with strangers (Cote, 2017). They might cultivate high levels of expertise to be known for their skill and adopt an especially aggressive persona in groups.

Collective forms of online resistance have also been found by Pei, Chib, and Ling (2021), for example, when women intervene and confront men who are sexually harassing other women users on group chats. In this vein, satire is another tool that can be collectively weaponized by women (Ringrose & Lawrence, 2018). For example, Vitis and Gilmour (2017) describe how the *Instagranniepants* art project brings together art and humour 'objectifying back' and satirizing male harassers, while the website savingroomforcats places cats in "manspreading" photographs [i.e., sitting with legs wide apart] (Ringrose & Lawrence, 2018). *TrollBusters*, described as "online pest control", provides immediate online 'rescue services' (Ferrier & Garud-Patkar, 2018, p. 316). More retaliatory tactics can also be employed; the campaign #OutThem actively denounces male harassers and #MenCallMeThings gives women voice by revealing and re-tweeting sexist comments. As

Jane (2019, p. 1) astutely observes, "shutting down sexual harassment shouldn't make you shake in fear or feel like your stomach just fell 10 stories. We don't just need to be empowered, but released from the burden of protecting men's comfort at the expense of ourselves".

Reducing online gender harassment

While the studies cited above show how women may *cope*, they do not actually provide answers as to how online gender harassment might be *reduced*. Against this backdrop, we sought to provide women and their allies resources to stand up against online gender harassment. Our study is inspired by Munger's (2016) activist work on racial harassment, which shows that targeted message-based interventions can effectively reduce prejudice in online communication.

In a sample of 242 *Twitter* users, Munger (2016) found that participants who were 'told off' on Twitter statistically significantly reduced the number of racist slurs in their future posts. The intervention tweet was always the same (though who sent it was experimentally manipulated): "@[subject] Hey man, just remember that there are real people who are hurt when you harass them with that kind of language". The tweet reminds the recipient that tweets including racial slurs are hurtful and constitute a form of harassment. Munger's intervention suggests that appealing to offenders' emotions and inviting them to take the perspective of those that are discriminating against could be a way of reducing online harassment (Dovidio et al., 2004).. Although the results indicate that the identity of the tweeter plays a role. The tweet was effective at reducing racial slurs when it was sent by a White twitter user with a higher number of followers up to 2 weeks after being posted, but did not reduce the rate of racial slur when it was sent by either a person of colour, even with many followers, or a White person with few followers.

Other interventions that encourage people to focus on the feelings of another have been shown to arouse feelings of empathy and reduce prejudice toward outgroup members (Batson et al., 2002; Galinsky & Moskowitz, 2000). For example, Batson et al. (1997) found that perspective taking increased feelings of empathy toward members of stigmatized groups. In a study with 96 participants, Batson et al. (1997) found that eliciting empathy for Julie, a young woman with HIV, increased broader feelings of empathy towards people living with HIV in general. The researchers replicated their findings with Harold, a homeless man. In both cases, inducing empathy improved attitudes towards the stigmatized group as a whole. Further studies show that this effect remains regardless of stereotype beliefs (Vescio et al., 2003). Empathetic perspective taking can also reduce in-group favouritism (Galinsky and Moskowitz, 2000) and increase helping behaviours (Mallett et al., 2008). One way of tackling online gender harassment could therefore be to encourage harassers to take women's perspective and to inform them about the negative emotional consequences of their misogynist tweets.Another way to tackle online gender harassment is by spotlighting social norms. There is certainly an extensive amount of research on the importance of social norms in promoting behavioural change (see the review by Paluck & Green, 2009). The idea is to encourage people to change their behaviour without external incentives by simply communicating information about 'what is commonly done' (Schultz et al., 2018). People begin to realise that others do not engage in the same behaviour as much and would disapprove of them.

Social norm campaigns have been successful in reducing alcohol consumption (Perkins & Craig, 2006) and smoking (Hancock and Henry, 2003). They have also had some success online, for example, in a community group with 13 million subscribers, Matias (2019) found that announcing socially normative expectations of members' behaviours increased compliance and reduced harassment. Dai et al. (2021) also found that sending small nudges via text messages can mobilise action. Given that one purpose of online abuse is to harass women into conforming to patriarchal social norms (e.g., Felmlee et al. 2020), what would happen if women attempted to 're'-norm offenders' beliefs? Accordingly, we reasoned that if we informed misogynist offenders that most people disapprove of their sexist language, that this could reduce the number and frequency of their sexist Tweets. Indeed, most men over-estimate others' sexism and educating them about this could be the first step (Kilmartin et al., 2008).

Following Munger's (2016) example, and the insights gleaned from the other interventions discussed above, we created two simple messages that could be tweeted in response to online gender harassment with the aim of reducing sexist slurs in subsequent tweets. In doing so we are both engaging in activism to challenge online misogyny via feminist action (e.g., Turley & Fisher, 2018) and responding to academic calls to design interventions to strengthen women's voices in online spaces (e.g., Jane, 2014).

Method

Overview

Small nudges might work to change behaviours (i.e., Dai et al., 2021; Munger, 2016). For this reason, when we designed our study we had reason to believe that our interventions could potentially reduce online gender harassment. Like Munger (2016), who first identified a group of racist tweeters, we identified a sample of misogynist *Twitter* users who frequently tweeted sexist slurs and then posted two tweets using *Twitter's* mention function in which a person can be mentioned or directly addressed by 'tagging' them (e.g., Hey, @Username!). One tweet aimed to socially re-norm sexist users and the other called for empathy. We then reviewed the pre-and-post streams of tweets to assess if our interventions had any effect. To foreshadow our results, they did not. We transparently share our methodological approach and decision-making below.

Step 1: Identify misogynist Twitter users

Identifying *Twitter* users posting misogynist tweets was not that difficult given the large population group, but we still needed to identify a sample of users with whom we could try our interventions. Since tweets are only 280 characters long, we needed a very concise and precise way to identify online gender harassment. To do so, we operationalised online gender harassment via posting of either one of two sexist slurs: “fucking bitch” and “fucking cunt” in users’ tweets. The most popular expletive on *Twitter* is “fuck”, which accounts for 34.73% of all occurrences (Wang et al., 2014). “Bitch”, “cunt”, and “slut” are common gendered slurs that target women specifically (Felmlee et al., 2020). Initially, we combined “fuck” with all three terms given their prevalence, however the term “fucking slut” brought up mostly pornographic contents and, while we acknowledge that some porn can be misogynist, this was beyond the scope of our study. We therefore selected only the sexist slurs “fucking bitch” and “fucking cunt”.

Our first objective was to obtain a sample of tweets that featured these sexist slurs. To do so, we used the StreamR package in R (Barbera, 2014) to connect to *Twitter’s* official application programming interfaces (API) and collected tweets (i.e., scraped) over a period of six days. Our initial sample consisted of whopping 89,939 tweets. We proceeded to automatically remove non-alphanumeric symbols, links, excessive white space, numbers, and usernames (e.g., “@username” inside the tweet’s body). We screened out all retweets and removed duplicates. The initial filtering process left us with 6,024 tweets (out of the initial 89,939) that featured at least one of our two sexist slurs (i.e., “fucking bitch” and “fucking cunt”). We still found that a large proportion of these tweets were pornographic content (e.g., advertisements), and we therefore strengthened our exclusion criteria (see fig. 1). To do so, we removed tweets with more than three hashtags (i.e., #word) and from users who tweeted most often (the upper quartile of average activity in our sample [75th to 100th]) since these users are predominantly advertisers. The remaining sample included 2,970 misogynistic tweets that featured at least one of our two sexist slurs; these were posted from 2,844 Twitter users.

We then manually coded each tweet to confirm that it was indeed aimed at harassing women. This process was arduous, time consuming, and shocking. We were disappointed and saddened at the vehement violence directed at women on Twitter. We also experienced several methodological challenges. As shown in the inclusion and exclusion criteria listed below (Figure 1), we had to identify tweets where the sexist slur was used in an unambiguous derogatory way (inclusion criterion 1) and targeted a woman/women (inclusion criterion 2). We had to weed out tweets where the sexist slur was negated (exclusion criterion 3) or not intended to be derogatory/used in a power affirming way (e.g., “Well done you fucking bitch! You nailed it!”; exclusion criterion 5), but the intent was not always easy to decipher. Our coding framework (Figure 1) emerged iteratively by toing-and-froing between the tweets and discussions between authors to assess their relevance.

INSERT FIGURE 1 HERE

We then assessed inter-rater reliability (Kappa = .634) and selected tweets above the chance threshold, with an agreement rate of 65% or more. In this way, we arrived at a sample of 1,000 tweets containing sexist slurs harassing women. Given that manually coding such a large sample of tweets was time consuming, by the time we had accomplished our goal, only 847 offending users were still active on *Twitter*. The sample attrition may have occurred because users closed their accounts, changed their privacy settings, or changed their username.

Step 2: Designing and delivering our interventions

We created two tweets based on re-norming and encouraging empathy, respectively: (1) @_____________ *Most people believe that some of your tweets against women are simply unacceptable* and (2) @_____________*Women are hurt by some of your tweets. Take a minute to think about how they feel.* To assess their suitability for our purpose and determine whether people would indeed interpret these statements as communicating social norms and appealing to empathy, we presented both tweets to an independent and unrelated sample of 272 participants (136 participants evaluated each tweet). We asked whether these tweets were believable (yes/no), if their presumed goal was to stop online gender harassment (yes/no), and might they allude to social norms or empathy. For both interventions, we also asked a "check" question that stated an erroneous goal (that the tweet was aimed at encouraging people to recycle) to avoid capturing acquiescence as evidence of understanding. The results showed that participants believed that both tweets were realistic (90% and 85% for the re-norming and empathy interventions respectively) and that their aim was to stop the recipient from harassing women online (93% and 79%). Participants also correctly identified the re-norming tweet as communicating social disapproval from most people (89%) and the empathy tweet as appealing to the recipient's emotions (82%).

We then proceeded to randomly allocate our sample of 847 users into two experimental groups (n=282 in the re-norming tweet condition and 282 in the empathy one) and a control group (n = 283) to whom we did not send any intervention tweet. We sent the intervention tweets at regular intervals to abide by Twitter's rules and regulations concerning the limited number of tweets that can be sent to other users in any given hour. All tweets were sent from a research account that we had named "Lizzy ________" belonging to a fictional woman named Elizabeth ______. We addressed each unique user specifically via the

"@username [intervention message]" format. Two users replied: "Hiya Lizzy he just dmed [sent a direct message to] me telling you to lick his bald head" and "#Balded". Someone retweeted our intervention tweet and someone liked it. The "Lizzie's" account was populated with neutral tweets prior to sending the intervention and had just over 50 followers at the time of data collection.

The tweets were continuously monitored, and data were collected for a period of 62 days, 31 days before and after the intervention Tweets. Afterwards, we individually tweeted the messages below to our sample to debrief them and give them the opportunity to withdraw their data. No one requested to withdraw their data.

@______________*You have been part of a study on online behaviour towards women. We are interested in finding solutions to reduce poor online behaviour such as being derogatory against women.*

@______________*We hope you value our interest in improving girls and women's lives. If you would like to withdraw your participation from our study, please let us know by emailing: withdrawresearch@gmail.com with your Twitter username.*

Data analysis

We extracted each user's *Twitter* activity exactly 31 days prior and 31 days after the intervention tweets. For the control group, we used a 62-day window of activity that we split in two 31-day periods: prior to and after *non*-intervention to make the number of tweets comparable across conditions. There was further sample attrition because some people did not tweet at all or tweeted very rarely during this time frame. (Accounts with fewer than five tweets either before or after the intervention were excluded.) After sample attrition, our final sample included 487,659 tweets from 666 users; 218 were in the re-norming condition, 214 were in the empathy condition, and 234 were in the control condition. Table 1 provides the descriptive statistics for our final sample.

INSERT TABLE 1 HERE

We assessed the frequency with which users posted sexist slurs by developing a list of words and expressions commonly used to denigrate women from urbandictionary.com (the full list appears in appendix I). This approach allowed us to form a ‘big’ picture overview and to assess changes in public discourse on *Twitter* more generally. To code the data, raw tweets containing any of the terms from appendix I were pooled into *Excel*. All authors then applied the coding framework (figure 1). As we had already encountered in step 1, some of the terms were used in a literal sense and not as slurs (e.g., ‘tart’ to refer to a pie).

We also noted that specific sexist slurs are somewhat limited in capturing subtler forms of gender harassment (e.g., “this woman was so fucking stupid that it was actually fun to see her fail”) or threats (e.g., “I would like to kill this woman”). We therefore complemented the focus on frequencies of specific sexist slurs by assessing the valence (i.e., positive or negative) and arousal (i.e., low or high intensity) of the words, based on the premise that words carry and evoke emotions (e.g., happy, unhappy etc.) (Warriner, Kuperman, and Brysbaert, 2013).. Some words can have both a positive valence and high arousal (e.g., “excited”) and others can have a neutral valence and low arousal (e.g., “table”). Offensive words have both high negative valence and high arousal (e.g., “bitch”). They imply very negative feelings and high levels of intensity. Using Warriner et al.’s (2013) coding of 13,915 words and matching them to our sample of tweets, we were also able to explore if there were any changes in the valence and arousal of users’ tweets following our intervention tweets.

Results

Effects of the intervention on sexist tweets and users

To compare a user's propensity to tweet a sexist slur, we focused on the normalised variables: (a) the frequency of tweets featuring a sexist slur out of the total number of tweets sent by a given user, and (b) the number of users who tweeted a sexist slur (at least once) out of the total number of users in each condition. We also checked the transience of our interventions on both the short-term (i.e., 7 days after our intervention) (see table 2) and longer-term (i.e., 31 days after our intervention) (see table 3). To compare the rate of sexist tweets and sexist users before and after the intervention, we computed: (a) the number of sexist tweets after our intervention and deducted from this the number of sexist tweets that were being posted before our intervention (scores ranged from -9.09% to +14.29%), and (b) the number of sexist users after our intervention minus the number of sexist users before our intervention (ranges from -1 to 1). A difference score of 0 meant that the intervention did not have an effect, whereas a positive difference meant that the rate of sexist tweets and sexist users increased after the intervention, and finally, a negative difference meant a decrease in the rate of sexist tweets and sexist users.

As shown in Table 2, the rate of sexist slurs and users did not vary greatly before and after the intervention (see rows in bold). In the social re-norming condition, there was an increase in the number of sexist slurs while the number of Twitter users who tweeted a sexist slur remained stable. In the empathy condition, we noticed both an increased trend in the number of sexist slurs and an increase in the number of users who tweeted a sexist slur. However, the most important increase in sexist slurs and users occurred in the control condition. To assess significance, we used a non-parametric Kruskal-Wallis test, which showed that the effects were not statistically significant in either the short (7 days) or the long term (31 days), *Kruskal-Wallis* (2) = 2.98, *p* = .225 and *Kruskal-Wallis* (2) = 1.15, *p* = .564. A chi square comparing the change in proportion of sexist users across conditions was not statistically significant either, whether we considered the short-term effect (7 days) or the longer one (31 days), $\chi^2(4, N = 578) = 7.53$, *p* = .110, *Cramer's V* = .08 and $\chi^2(666) = 2.56$, *p* = .634, *Cramer's V* = .05.

INSERT TABLE 2 AND TABLE 3 HERE

Effect of the intervention of the valence and arousal of tweets

Figure 2 illustrates the valence and arousal averages for all tweets across the study's 62-day research window. We evaluated the valence and arousal of tweets that included a sexist slur and those that did not as to assess whether the former was indeed more negative and arousing than the latter. . Our preconceptions were correct and we found that tweets with sexist slurs were markedly less positive and generated stronger arousal than tweets that did not include sexist slurs. However, as is clear from the flat pattern over time, our intervention tweets did not have any effect on the valence and arousal of the words being used. Users' tweets following our interventions were neither less negative nor less emotionally loaded.

INSERT FIGURE 2 HERE

General Discussion

Responding to calls by Turley and Fishers (2018) and Jane (2014) to empower women in online spaces, we designed two straightforward responses that women could tweet in reply to online gender harassment. Our preconceptions were that (1) social re-norming, and (2) appealing to empathy could decrease sexist slurs—in the same way that these types of messages were found to reduce online racist harassment (Munger, 2016). We also tested to see if the valence and arousal of tweets posted before and after our interventions changed. Regrettably, our interventions did not reduce the frequency of sexist tweets nor the number of sexist users either 7 or 31 days after. We did not observe a change in the valence or arousal of users' tweets, or a reduction in the overall rate that users tweeted (with or without sexist slurs). Although these findings are disappointing, it is important to reflect on the possible reasons for our interventions' lack of effect and discuss the conceptual, technical, and ethical challenges associated with reducing online gender harassment.

Our first tweet was an attempt at socially 're'-norming by reminding offenders that most people found their tweets against women unacceptable. Our second intervention was based on highlighting the affective consequences of using misogynistic language and appealing to users' empathy. The apparent failure of these interventions could be a type II error: the effect exists, but we were not able to statistically capture it in this sample. Our study focused on a sample of 666 Twitter users who posted before and after our interventions, and they were split across three conditions: social re-norming, empathy, and control. When comparing one of our two experimental conditions to the control condition, we had a 90% power (with a 5% alpha) to detect a small to medium between-subject mean difference in the number of tweets including a slur (Cohen's d = .29). We could argue that even a difference of 0.5% could actually be meaningful and represent a large number of tweets (we found 89,939 tweets featuring "fucking cunt" or "fucking bitch" in only 6 days).

Indeed, research shows that both these strategies have been successfully used in changing antisocial or undesirable behaviours in several different interventions. Communicating social norms has been shown to limit alcohol consumption (Perkins & Craig, 2006) and even reduce intentions to harass in *Facebook* groups (Van Royen et al., 2017). Similarly, encouraging people to take the perspective of others and develop empathy can decrease prejudice (Batson et al., 2002; Galinsky and Moskowitz, 2000; Vescio et al., 2003). Intervention messages that highlight the negative consequences of online harassment (e.g., "This comment may be hurtful for the receiver. Are you sure to post it?") were found to be successful in reducing the intention to harass on *Facebook* (Van Royen et al., 2017). Yet again, this was not the case in our study, and we did not find that our intervention tweets influenced the frequency of sexist slurs tweeted or the number of users tweeting derogatory material. Why might our intervention not have proved as successful in reducing the use of sexist slurs? The possible reasons for our lack of success provide some insights for designing ones that might work better.

Light nudges might not be enough to reduce online gender harassment

Communicating social norms has been shown to be an effective way to nudge behaviour change (Paluck & Green, 2008). Light nudges have proven successful via text messages (Dai et al., 2021) and on *Twitter* (Munger, 2016). However, it may be that single tweets are simply not powerful enough to prompt misogynist behaviour change. We are exposed to such a vast amount of content on social media that a single tweet may have been drowned in masses of other emotion-rich contents. It may be that a greater number of intervention tweets could actually have an impact; for example, multiple similarly worded messages from several different accounts could reinforce the re-norming message that the harassing behaviour is not acceptable, but from an ethical perspective this would constitute 'harassing the harassers'. Notwithstanding, we do know that at least some offenders in our sample received and noted our messages because we received a few reactions to our tweets (e.g., likes, retweets, replies). Nevertheless, a tweet is only a micro-intervention in a macro-level system of entrenched in sexism.

The nature and extent of sexism on- and off-line

Further, sexism is deeply ingrained in society (e.g., #MeToo, Time's Up) and online gender harassment is normalised on the Internet (e.g., Felmelee et al., 2020). We were perhaps overtly optimistic in attempting to reduce it via a couple of tweets, despite this approach being successful at reducing racist harassment in other online studies (e.g., Munger, 2016; Pennycook et al., 2021). Reflecting on why this might be, research shows that racism is generally believed to be more offensive than sexism (Woodzicka et al., 2015). Individuals who are publicly confronted for using racist slurs might feel more embarrassed than people who are using sexist slurs given the prevalence of sexist attitudes are very common (Georgeac et al., 2019). To illustrate, Felmelee et al. (2020) found over 2.9 million tweets in just one week that contained sexist slurs. This shocking rate maintains the online gender harassment cycle because the sheer number of tweets reinforce the idea that 'everybody does it'. Certainly, sexist harassers might feel less chastised in online spaces than in offline spaces, especially given the protection of anonymity that social media affords them. Disclosing one's true identity has been shown to reduce the use of offensive language (Cho and Acquisti, 2013). For example, Lapidot-Lefler and Barak (2012) found that participants assigned to an eye-contact condition via webcam were twice less likely to engage in "flaming" behaviours [i.e., personally attacking] than those assigned to the no-eye-contact condition in their online experiment on "toxic" online disinhibition. In our case, although some users did display demographic data, it was impossible to discern its authenticity and thus whether our tweet could expose them in any meaningful way, as in campaigns such as #OutThem do successfully.

The social identity of the tweeter

Yet another reason, grounded in patriarchy and misogyny, might be that our intervention tweet was posted by someone who clearly appeared to be a woman: "Elizabeth_______". Women who confront sexism are often denigrated as hysterical 'whiners' (Doyle, 2011) and 'overreactors' (Czopp et al., 2006), enabling their views to be more easily discounted. Men, of course, are taken more seriously than women when they confront sexism (Drury and Kaiser, 2014). We tried to mitigate this by using the gender neutral 'most people' as our reference group in the re-norming condition, but we nevertheless recognise that the confronter's apparent gender might have played a role in the intervention's lack of effect.

Methodological and ethical considerations when studying (and trying to change) online behaviour on social media

Efficiently identifying online gender harassment for research purposes is difficult on social media. Despite using stringent filtering criteria on raw tweets, we had to resort to manual coding. Our initial sample contained an overwhelming amount of pornography demeaning women (i.e., content promoted as being about 'fucking bitch' and 'fucking cunt'). We managed to exclude a substantial number of those tweets by filtering out those featuring web links and more than three hashtags; nonetheless, we still found a significant number of pornographic tweets including the two sexist slurs "fucking bitch" and "fucking cunt" while manually inspecting our data.

Should social media platforms take more responsibility and actions for policing their content? Despite several high-profile cases and activism by groups such as *Amnesty International*, social media giants are largely only meekly 'policing' themselves, with little to no impact on harassed women's actual experiences (e.g., Chadha et al., 2020; Amnesty International, 2020). Moreover, the Home Affairs Committee (2017: 31) in the UK has criticised social media companies' reliance on users to report abuse as "outsourcing the vast bulk of their safeguarding responsibilities at zero expense". This is simply one example of a larger issue around social media companies failing to adequately address hate speech and misinformation on their platforms.

A second reason efficiently identifying online gender harassment is challenging for research purposes is because it is not possible to automatically detect slurs that are used in an empowering way. For example, marginalised groups often ‘take ownership’ of derogatory words that have been historically used against them (Galinsky et al., 2013) (e.g., the adoption of the word ‘queer’ by gender non-conforming persons), but manually coding such a large dataset is resource intensive (see Schwartz and Ungar, 2015 for further guidance on how to review social media posts). The creation of algorithms to automatically detect a range of negative content online is currently a pressing topic to tackle all forms of harassment including hate speech (Schmidt & Wiegand, 2017) and cyberbullying (Van Hee et al., 2018) – see Zimmerman et al. (2018) for discussions on how to improve detection. Other avenues for research include how women might take ownership of sexist discourse in online spaces in an empowering way, and how it is precisely the *femininity* in sexist slurs that is perceived to be offensive (see Hoskin's 2019 work on femmephobia), for example, by ‘insulting’ a male footballer in saying that he plays like a ‘bitch’.

Our involvement in this study also brought up interesting questions about conducting ethical research online. We were engaging with publicly available data and did, of course, acquire ethics approval from our university. It is, however necessary, that any users whose behaviours are monitored, are made aware that they are participating in research, but in some instances, like ours, the study premise relies on being covert. We felt it important to debrief users after the study and give them the opportunity to withdraw their tweets. Yet, this brings up another uncomfortable dilemma for researchers. Do we want to open ourselves to harassment by people who are clearly prone to harassing? Vera-Gray (2017) has already documented the dangers of women academics being trolled for simply doing research online. Despite the time-consuming nature of the activity, we manually sent individual debrief @tweets to everyone in our sample, as discussed earlier, but we chose not to disclose our identities and used an anonymous email research account. (See the BPS Ethics Guidelines for Internet-Mediated Research, 2017 or the AoIR Internet Research: Ethical Guidelines 3.0 (2019) for further information.)

Conclusion

Online gender harassment is an extension of the violence done to women in the offline world (Lindsay et al., 2016). Several scholars have called for measures to tackle this problem (e.g., Turley & Fisher, 2018; Jane, 2014). We therefore designed two straightforward tweets based on principles of social re-norming and empathy and tested these on a sample of 666 *Twitter* users. Our intervention tweets did not, regrettably, reduce the number of sexist slurs or sexist users in our sample or affect the valence or arousal of subsequent tweets. Disappointing, but perhaps not altogether surprisingly given how prolific sexist slurs are on social media and how normalised online gender harassment has become. Nonetheless, our results provide some insights and points for further discussion about how women might be empowered to respond to online gender harassment. We have pointed to the insufficiency of strategies, such as light nudges to promote pro-social norms and empathy, that work with other kinds of prejudice for intervening in online sexism, as well as the difficulties for women as victims of abuse to enact such strategies. We add our voices to calls for further work in this area.

Declaration of conflicting interests

No conflict of interests to declare.

Funding

We received no financial support for the research.

References

Action Aid. (2016). *Three in four women experience harassment and violence in UK and global cities*. https://www.actionaid.org.uk/latest-news/three-in-four-women-experience-harassment-and-violence

Amnesty International. (2017). *More than a quarter of UK women experiencing online abuse and harassment receive threats of physical or sexual assault - new research*. https://www.amnesty.org.uk/press-releases/more-quarter-uk-women-experiencing-online-abuse-and-harassment-receive-threats

Amnesty International. (2020). *Violence against women*. https://www.amnesty.org.uk/violence-against-women

Barbera, P. (2014). *streamR: Access to Twitter Streaming API via R. R package version 0.2.1*. https://cran.r-project.org/package=streamR

Batson, C. D., Chang, J., Orr, R., & Rowland, J. (2002). Empathy, attitudes, and action: Can feeling for a member of a stigmatized group motivate one to help the group? *Personality and Social Psychology Bulletin*, *28*(12), 1656–1666. https://doi.org/10.1177/014616702237647

Chadha, K., Steiner, L., Vitak, J., & Ashktorab, Z. (2020). Women's Responses to Online Harassment. *International Journal of Communication*, *14*(0), 19.

Chen, G. M., Pain, P., Chen, V. Y., Mekelburg, M., Springer, N., & Troger, F. (2020). 'You really have to have a thick skin': A cross-cultural perspective on how online harassment influences female journalists. *Journalism, 21*(7), 877–895. https://doi.org/10.1177/1464884918768500

Cho, D., & Acquisti, A. (2013). The More Social Cues , The Less Trolling? An Empirical Study of Online Commenting Behavior. *The Twelfth Workshop on the Economics of*

Information Security, Weis.

Citizens Advice. (2021). *Sexual Harassment*. https://www.citizensadvice.org.uk/law-and-courts/discrimination/what-are-the-different-types-of-discrimination/sexual-harassment/

Citron, D. (2014). *Hate crimes in cyberspace*. Harvard University Press.

Cortina, L. M., & Magley, V. J. (2003). Raising voice, risking retaliation: Events following interpersonal mistreatment in the workplace. *Journal of Occupational Health Psychology, 8*(4), 247. https://doi.org/10.1037/1076-8998.8.4.247

Cote, A. C. (2017). "I Can Defend Myself": Women's Strategies for Coping with Harassment while Gaming Online. *Games and Culture, 12*(2), 136–155. https://doi.org/10.1177/1555412015587603

Czopp, A. M., Monteith, M. J., & Mark, A. Y. (2006). Standing up for a change: Reducing bias through interpersonal confrontation. *Journal of Personality and Social Psychology, 90*(5), 784. https://doi.org/10.1037/0022-3514.90.5.784

Dovidio, J. F., Ten Vergert, M., Stewart, T. L., Gaertner, S. L., Johnson, J. D., Esses, V. M., Riek, B. M., & Pearson, A. R. (2004). Perspective and prejudice: Antecedents and mediating mechanisms. *Personality and Social Psychology Bulletin*, *30*(12), 1537–1549. https://doi.org/10.1177/0146167204271177

Drakett, J., Rickett, B., Day, K., & Milnes, K. (2018). Old jokes, new media -online sexism and constructions of gender in Internet memes. *Feminism and Psychology*, *28*(1), 109–127. https://doi.org/10.1177/0959353517727560

Drury, B. J., & Kaiser, C. R. (2014). Allies against sexism: The role of men in confronting sexism. *Journal of Social Issues, 7*0(4), 637-652. doi: 10.1111/josi.12083

Eckert, S., & Metzger-Riftkin, J. (2020). Doxxing, Privacy and Gendered Harassment. The Shock and Normalization of Veillance Cultures. *M&K Medien & Kommunikationswissenschaft, 68*(3), 273-287. doi.org/10.5771/1615-634X-2020-3-273

Felmlee, D., Inara Rodis, P., & Zhang, A. (2020). Sexist Slurs: Reinforcing Feminine Stereotypes Online. *Sex Roles*, *83*(1–2), 16–28. https://doi.org/10.1007/s11199-019-01095-z

Ferrier, M., & Garud-Patkar, N. (2018). TrollBusters: Fighting Online Harassment of Women Journalists. In J. Vickery & T. Everback (Eds.), *Mediating Misogyny* (Issue February, pp. 311–332). Palgrave Macmillan. https://doi.org/10.1007/978-3-319-72917-6

Galinsky, A., & Moskowitz, G. (2000). Perspective-Taking: Decreasing Stereotype Expression, Stereotype Accessibility, andIn-Group Favoritism. *Journal of Pesonality and Social Psychology*, *78*(4), 708–724. https://doi.org/10.1037//0022-3514.78.4.708

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels. *Journal of Consumer Research*, *35*(3), 472-482. https://doi.org/10.1086/586910

Han, X. (2018). Searching for an online space for feminism? The Chinese feminist group Gender Watch Women's Voice and its changing approaches to online misogyny. *Feminist Media Studies*, *18*(4), 734–749. https://doi.org/10.1080/14680777.2018.1447430

Hancock, L. C., & Henry, N. W. (2003). Perceptions, norms, and tobacco use of college residence hall freshmen: Evaluation of a social norms marketing intervention. In H. W. Perkins (Ed.), The social norms approach to preventing school and college age substance abuse: A Handbook for Educators, Counselors, and Clinicians (pp. 135–153). Jossey-Bass/Wiley.

Herring, S. (2003). Gender and power in online communication. In *The Handbook of Language and Gender* (pp. 202–228).

HM Crown Prosecution Service Ispectorate. (2019). *2019 Rape Inspection*. *December*.

Hoskin, R. A. (2019). Femmephobia: The Role of Anti-Femininity and Gender Policing in

LGBTQ+ People's Experiences of Discrimination. *Sex Roles*, *81*(11–12), 686–703. https://doi.org/10.1007/s11199-019-01021-3

House of Commons. (2017). *Online harassment and cyber bullying* (Issue 07967).

Jackson, S. (2018). Young feminists, feminism and digital media. *Feminism & Psychology*, *28*(1), 32–49. https://doi.org/10.1177/0959353517716952

Jane, E. (2014). Back to the kitchen, cunt: Speaking the unspeakable about online misogyny. In *Continuum* (Vol. 28, Issue 4, pp. 558–570). Taylor & Francis. https://doi.org/10.1080/10304312.2014.924479

Jane, E. (2018). Gendered cyberhate as workplace harassment and economic vandalism. *Feminist Media Studies*, *18*(4), 575–591. https://doi.org/10.1080/14680777.2018.1447344

Jane, T. (2019). *Creepy men slide into women's DMs all the time, but they can be shut down*. The Guardian. https://www.theguardian.com/commentisfree/2019/may/07/creepy-men-dm-online-harassment

Kilmartin, C., Smith, T., Green, A., Heinzen, H., Kuchler, M., & Kolar, D. (2008). A real time social norms intervention to reduce male sexism. *Sex Roles*, *59*(3–4), 264–273. https://doi.org/10.1007/s11199-008-9446-y

Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*, *28*(2), 434–443. https://doi.org/10.1016/j.chb.2011.10.014

Lindsay, M., Booth, J. M., Messing, J. T., & Thaller, J. (2016). Experiences of Online Harassment Among Emerging Adults: Emotional Reactions and the Mediating Role of Fear. *Journal of Interpersonal Violence*, *31*(19), 3174–3195. https://doi.org/10.1177/0886260515584344

Mallett, K. A., Bachrach, R. L., & Turrisi, R. (2008). Are all negative consequences truly negative? Assessing variations among college students' perceptions of alcohol related consequences. *Addictive Behaviors, 33*(10), 1375-1381. https://doi.org/10.1016/j.addbeh.2008.06.014

Mantilla, K. (2013). Gendertrolling: Misogyny adapts to new media. *Feminist Studies, 39*(2), 563-570. DOI.10.1353/fem.2013.0039

Megarry, J. (2014). Online incivility or sexual harassment? Conceptualising women's experiences in the digital age. *Women's Studies International Forum, 47*(PA), 46–55. https://doi.org/10.1016/j.wsif.2014.07.012

Met Police. (2021). *I'm being harassed by someone on social media. What can I do?* https://www.met.police.uk/advice/advice-and-information/har/harassment-on-social-media/#:~:text=You can report either harassment,by calling us on 101.

Munger, K. (2016). Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior*, 1–21. https://doi.org/10.1007/s11109-016-9373-5

Matias, N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(20), 9785–9789. https://doi.org/10.1073/pnas.1813486116

Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: what works? A review and assessment of research and practice. *Annual Review of Psychology*, *60*, 339–367. https://doi.org/10.1146/annurev.psych.60.110707.163607

Pei, X., Chib, A., & Ling, R. (2021). Covert resistance beyond# Metoo: mobile practices of marginalized migrant women to negotiate sexual harassment in the workplace. Information, *Communication & Society*, 1-18. https://doi.org/10.1080/1369118X.2021.1874036

Pennycook, G., Epstein, Z., Mosleh, M. Arechar, A., Eckles, D., & Rand, D. (2021) Shifting attention to accuracy can reduce misinformation online. *Nature 592*, 590–595. https://doi.org/10.1038/s41586-021-03344-2

Perkins, H. W., & Craig, D. W. (2006). A successful social norms campaign to reduce alcohol misuse among college student-athletes. *Journal of Studies on Alcohol, 67*(6), 880-889. https://doi.org/10.15288/jsa.2006.67.880

Regehr, K. (2020). In (cel) doctrination: How technologically facilitated misogyny moves violence off screens and on to streets. *New Media & Society*, https://doi.org/10.1177/1461444820959019.

Roberts, N., Donovan, C., & Durey, M. (2019). Agency, resistance and the non-'ideal'victim: how women deal with sexual violence. *Journal of Gender-Based Violence, 3*(3), 323-338. DOI: https://doi.org/10.1332/239868019X15633766459801

Rights of Women. (2021). *Rights of Women survey reveals online sexual harassment has increased, as women continue to suffer sexual harassment whilst working through the Covid-19 pandemic*. https://rightsofwomen.org.uk/news/rights-of-women-survey-reveals-online-sexual-harassment-has-increased-as-women-continue-to-suffer-sexual-harassment-whilst-working-through-the-covid-19-pandemic/

Ringrose, J., & Lawrence, E. (2018). Remixing misandry, manspreading, and dick pics: Networked feminist humour on Tumblr. *Feminist Media Studies, 18*(4), 686-704. https://doi.org/10.1080/14680777.2018.1450351

Schmidt, A., & Wiegand, M. (2017). *A Survey on Hate Speech Detection using Natural Language Processing*. *2012*, 1–10. https://doi.org/10.18653/v1/w17-1101

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2018). The Constructive, Destructive, and Reconstructive Power of Social Norms: Reprise. *Perspectives on Psychological Science*, *13*(2), 249–254. https://doi.org/10.1177/1745691617693325

Schwartz, H. A., & Ungar, L. H. (2015). Data-Driven Content Analysis of Social Media: A Systematic Overview of Automated Methods. *Annals of the American Academy of Political and Social Science, 659*(1), 78–94. https://doi.org/10.1177/0002716215569197

Stubbs-Richardson, M. S., Rader, N. E., & Cosby, A. G. (2018). Tweeting rape culture: Examining portrayals of victim blaming in discussions of sexual assault cases on Twitter. *Feminism & Psychology*, 28(1),90–108. https://doi.org/10.1177/0959353517715874

Thompson, L. (2018). ‘I can be your Tinder nightmare’: Harassment and mis- ogyny in the online sexual marketplace. *Feminism & Psychology*, *28*(1), 69–89. https://doi.org/10.1177/0959353517720226

Turley, E., & Fisher, J. (2018). Tweeting back while shouting back: Social media and feminist activism. *Feminism & Psychology*, *28*(1), 128–132. https://doi.org/10.1177/0959353517715875

Uhl, C., Rhyner, K., & Lugo, N. (2018). An examination of nonconsensual pornography websites. Feminism. *Feminism & Psychology*, *28*(1), 50–68. https://doi.org/10.1177/0959353517720225

UN Women UK. (2021). *Prevalence and reporting of sexual harassment in UK public spaces - A report by the APPG for UN Women*. *March*, 1–28.

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *ArXiv*, 1–22. https://doi.org/10.17605/OSF.IO/RGQW8.

Van Royen, K., Poels, K., Vandebosch, H., & Adam, P. (2017). “Thinking before posting?” Reducing cyber harassment on social networking sites through a reflective message. *Computers in Human Behavior, 66*, 345-352. https://doi.org/10.1016/j.chb.2016.09.040

Vera-Gray, F. (2017). Talk about a cunt with too much idle time’: Trolling feminist research.

Feminist Review, *115*(1), 61–78. https://doi.org/10.1057/s41305-017-0038-y

Vescio, T. K., Sechrist, G. B., & Paolucci, M. P. (2003). Perspective taking and prejudice reduction: The mediational role of empathy arousal and situational attributions. *European Journal of Social Psychology*, *33*(4), 455–472. https://doi.org/10.1002/ejsp.163

Vitis, L., & Gilmour, F. (2017). Dick pics on blast: A woman's resistance to online sexual harassment using humour, art and Instagram. *Crime, Media, Culture*, *13*(3), 335-355. https://doi.org/10.1177/1741659016652445

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2014). Cursing in English on twitter. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '14*, 415–425. https://doi.org/10.1145/2531602.2531734

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191-1207. https://doi.org/10.3758/s13428-012-0314-x

Woodzicka, J. A., Mallett, R. K., Hendricks, S., & Pruitt, A. V. (2015). It's just a (sexist) joke: Comparing reactions to sexist versus racist communications. *Humor*, *28*(2), 289-309. https://doi.org/10.1515/humor-2015-0025

Zimmerman, S., Kruschwitz, U., & Fox, C. (2018, May). Improving hate speech detection with deep learning ensembles. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). https://aclanthology.org/L18-1404.pdf

Biographical notes

Lilith A. Whiley is an inter-disciplinary psychologist whose work brings together Occupational Psychology, Human Resource Management, and Social Psychology. Her research centres around what makes people ‘different’, how ‘otherness’ is lived and embodied, and what organisations and society can do to improve inclusivity and redress inequalities.

Lukasz Walasek is a psychologist working in the areas of judgement and decision-making, subjective well-being, income inequality, valuation, natural language processing, and context effects.

Marie Juanchich is a behavioural scientist grounded in socio-cognitive psychology. Her work focuses on empowering people by helping them to understand uncertainty and probabilities, as well as reducing discrimination and increasing diversity.

Appendix I – List of commonly used sexist slurs

TERM	INCLUDE
arm candy	1
asking for it/asked for it	1
ball-breaker	1
ballbuster	1
battle axe	1
bimbo	1
bimbo	1
bint	1
bitch	0
bridezilla	1
bunny boiler	1
butch	1
butterface	1
catfight	1
chavette/girl chav?	1
cock tease	1
cougar	0
crank whore	1
crockadillapig	1
crone	1
cunt	0
daft bimbo	1
frigid bitch	1
frump	1
frumpy	1
fucking bimbo	1
fucking bitch	1
fucking cunt	1
gagging for it	1
ghetto bird	1
ghetto ho	1
gold digger	1
harridan	1
hoe	0
hooch	1
hoochie	1
hussy	1
huzzie	1
milf	0
MILF	0
minger	1
moll	1
moose	1
mousey	1

daft bitch	1
daft cow	1
daft cunt	1
damaged good	1
ditz	1
dizty	1
essex girl	1
fag hag	1
feminazi	1
flange	1
flipper	1
floozie	1
floozy	1
frigid	1
stupid bimbo	1
tart	1
town bike	1
tramp	1
troglodyte	1
trollop	1
vamp	1
village bicycle	1
what's-her-face	1
whatshername	1

old bag	1
pass around pussy	1
poon	1
poontang	1
prostitute	1
prude	1
pussy	0
sausage jockey	1
shrew	1
skank	1
skeezy ho	1
slag	1
slapper	1
sleaze	1

whore	0

1 = include; 0 = do not include.

List of figures

Tweet inclusion criteria.

Include in sample if Tweet meets the following criteria.

1. Tweets that use a sexist slur in an unambiguously derogatory way.
2. The slur is made against/about/in reference to women/a particular woman.

   The tweet should NOT be about a man.
3. The slur is NOT negated (e.g., “my mum is happy I am **not** a fucking bitch”).
4. The slur is NOT a report of someone else being derogatory (e.g., “someone screamed fucking bitch while I was driving. Okay cool”)
5. The slur is NOT used in an endearing/empowering way (e.g., “my fucking bitch”).
6. The slur is NOT self-deprecating (e.g., I am a fucking bitch I know…)
7. The slur is not associated with joking/laughing (e.g., “fucking bitch gave me a fright lol”)
8. The tweet does NOT come from porn companies (e.g., “slim east Asian, fucking bitch with her long dildo”).

Figure 1. Coding framework for manually identifying sexist slurs.

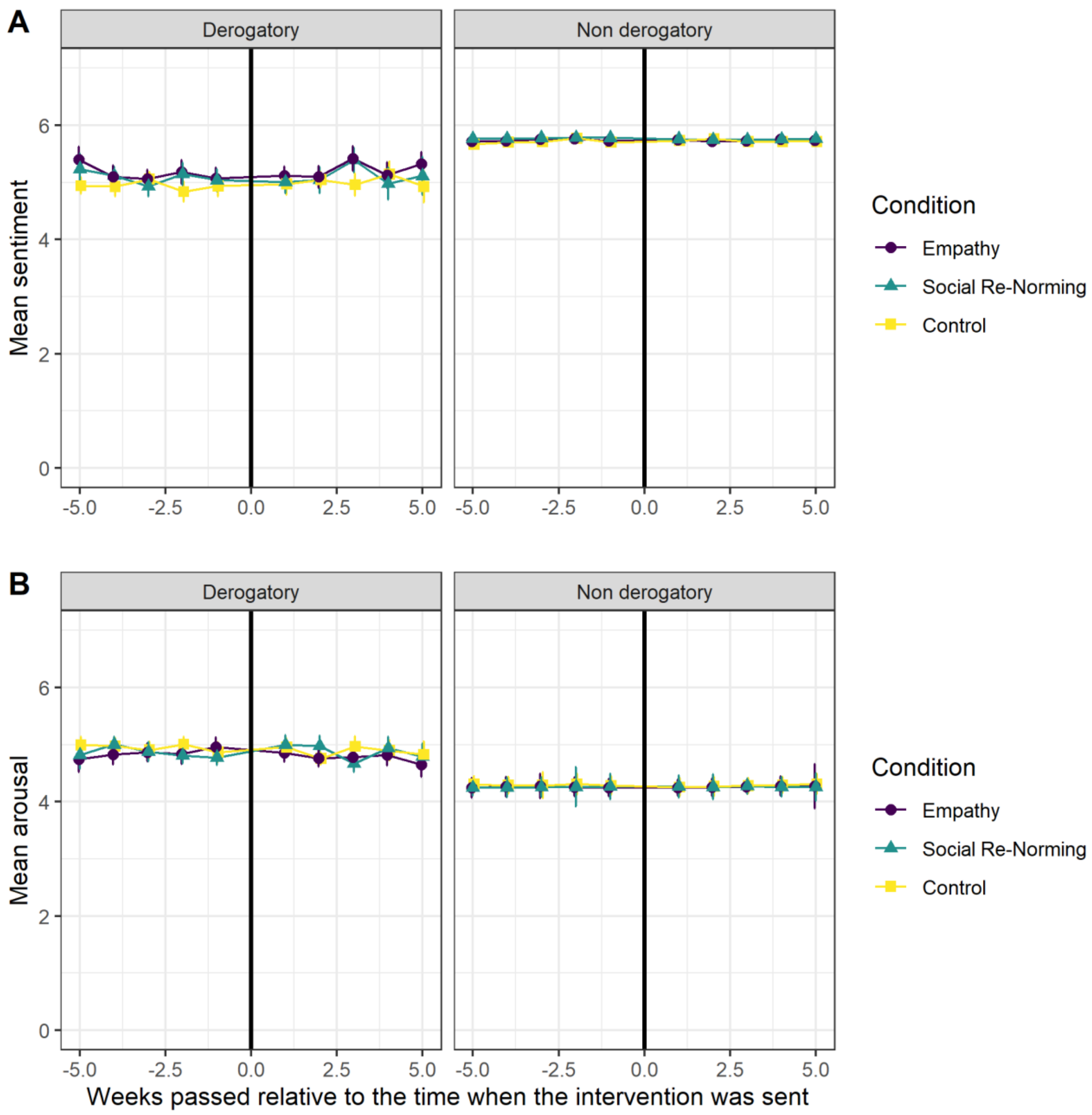

Figure 2. Panel A: Valence of tweets over the course of the study in weeks (ranging from 1: completely unhappy to 9: completely happy). Panel: B: Arousal of tweets over the course of the study in weeks (ranging from 1: completely calm to 9: completely aroused). In both Panel A and B, left panel shows tweets that include one of the sexist slurs; right panel: the remaining tweets. Error bars represent 2 standard errors of the means (they are too small to be clearly visible for non-derogatory tweets).

List of Tables

Table 1. Total and daily tweeting frequency and follower counts across experimental groups for Twitter users in our studies.

	Condition			
	Re-norming	Empathy	Control	Total
Number of users	218	214	234	666
Number of tweets over 62 days	164,621	145,335	177,703	487,659
Median number of tweets per day	7	8	9	8
Median followers count	775	529	573	619

Table 1. Percentage of sexist tweets and sexist users 7 days after our intervention

		7 days after our intervention tweets	
Condition of sample		**% of sexist tweets**	**% of sexist users**
Social re-norming	Before	0.42%	20%
	After	0.63%	21%
	Difference	**+0.24%**	**+1%**
Empathy	Before	0.73%	27%
	After	0.69%	24%
	Difference	**+0.03%**	**-3%**
Control	Before	0.57%	22%
	After	1.11%	26%
	Difference	**+0.54%**	**+4%**

Total	Before	0.57%	23%
	After	0.82%	24%

Table 3. Percentage of sexist tweets and users 31 days after our intervention tweets

		31 before/after	
Condition of sample		**% sexist tweets**	**% of sexist users**
Social re-norming	Before	0.57%	46%
	After	0.60%	46%
	Difference	**+0.11%**	**-/+0%**
Empathy	Before	0.48%	51%
	After	0.56%	53%
	Difference	**+0.06%**	**+2%**
Control	Before	0.68%	50%
	After	0.70%	54%
	Difference	**-0.03%**	**+4%**
Total	Before	0.58%	49%
	After	0.62%	51%