DOI: 10.1177/00222429231179942

JOURNAL of Marketing

For Shame! Socially Unacceptable Brand Mentions on Social Media Motivate Consumer Disengagement

Journal:	*Journal of Marketing*
Manuscript ID	JM.21.0334.R4
Manuscript Type:	Revised Submission
Research Topics:	Brand Communities, Consumer Identity, Digital Marketing - Customer, Social Media
Methods:	Hazard Models, Lab Experiments, Panel Data Models

For Shame! Socially Unacceptable Brand Mentions on Social Media Motivate Consumer Disengagement

Daniel Villanova*

Ted Matherly

* Daniel Villanova (dvillanova@walton.uark.edu) is Assistant Professor of Marketing, Walton College of Business, University of Arkansas, Business Building 302, Fayetteville, AR 72701. Ted Matherly (t.matherly@northeastern.edu) is Visiting Assistant Professor of Marketing, D'Amore-McKim School of Business, Northeastern University, Boston, MA 02115. Correspondence should be addressed to Daniel Villanova. Both authors contributed equally to this research.

Acknowledgements: The authors wish to thank the editor, associate editor, and three anonymous reviewers for their feedback during the review process. In addition, they wish to thank Rosellina Ferraro, Amna Kirmani, and Roland Rust for their comments on earlier versions of the manuscript, along with David Anderson, Zachary Arens, Brad N. Greenwood, Kelly Lee, Daniel Malter, Alejandra Rodriguez, Ashish Sharma, Steven Shepherd, and Kevin Voss for their comments on the studies and analysis.

For Shame! Socially Unacceptable Brand Mentions on Social Media Motivate Consumer Disengagement

Abstract

Brands invest tremendous resources into building engagement with their customers on social media. But considerably less focus is placed on addressing disengagement, when users actively choose to distance themselves from the brand, through reduced posting or even unfollowing. We find that the same self-brand connections that lead individuals to defensively protect the brand can also lead them to experience shame vicariously when others mention the brand in socially unacceptable ways. Experiencing vicarious shame motivates them to distance themselves from the brand, driving disengagement. In three mixed-method studies, we show that a socially unacceptable behavior - using profanity while mentioning the brand - leads highly connected consumers to experience vicarious shame, prompting disengagement motivations, and ultimately leading to real-world unfollowing behaviors on social media. We also show that proactive moderation behaviors by the brand can attenuate these responses. These results provide insight into the process by which self-brand connection interacts with socially unacceptable brand mentions and suggest a limitation to the insulating effects of strong self-brand connections.

Keywords: disengagement, social media, self-brand connection, shame

*If I see any more Bears fans talking s*** before a game, I will unfollow…You guys make us look terrible.*

--Anonymous Twitter user

Engaging with consumers on social media is an essential part of the modern marketing toolkit. Considerable research has studied how brands engage with customers (Hollebeek, Glynn, and Brodie 2014; Meire et al. 2019), and social media has greatly extended the ways in which brands can make and measure these interactions, such as likes, comments, or shares (Lee Hosanagar, and Nair 2018; Cucu 2021) or following a brand (Lamberton and Stephen 2016; Saboo, Kumar, and Ramani 2016). Moreover, social media engagement generates substantial value for companies (de Oliveira Santini et al. 2020).

By contrast, social media disengagement has received relatively little attention, which is surprising given the marketing aphorism that it is much easier to keep existing users than to acquire new ones. In this research, we seek to address this gap by developing a stronger understanding of one potential driver of social media disengagement. We define *social media disengagement* as the manifestation on social media of the psychological motivation to distance oneself from a brand. A striking case of this occurs when a consumer unfollows a brand. Just as customer retention is understood to have distinct drivers from customer acquisition (Reinartz, Thomas, and Kumar 2005), the reasons why consumers engage with brands on social media may not completely overlap with why those same consumers might want to disengage. Examining causes of disengagement is important to firms because disengagement involves the reduction in brand-related activities that have been linked to purchase intentions (Swani, Milne, and Miller 2021) and sales (Liadeli, Sotgiu, and Verlegh 2022; Saboo et al. 2016). Disengagement is also

important to understand because it can undermine the reach of future marketing content (Lee et al. 2018). For example, when someone unfollows a brand on Twitter, that consumer is no longer directly reachable by the brand's Twitter presence, and nor are that consumer's followers, who will not see the brand's content via that consumer's interactions.

The goal of the present research is to explore one potential reason why consumers who are highly connected to a brand may choose to disengage from it: their observation of socially unacceptable mentions of the brand on social media. We propose that these behaviors threaten the identity of highly connected consumers, leading them to experience vicarious shame, and in turn motivating them to disengage with the brand. In addition to its theoretical contributions to the domains of social media consumer behavior, self-brand connection, and social identity theory, this research also has practical implications for managers and marketing strategists. We demonstrate a previously unstudied risk accompanying individuals who are highly connected to brands and test the role that actively moderating posts with socially unacceptable brand mentions can play in mitigating this risk. We discuss these contributions in greater detail in the general discussion. In the following sections, we first develop the theoretical foundation for our work. After deriving our predictions, we present the results of three studies.

Conceptual Development

Social Media Disengagement, Motivation to Distance, and Vicarious Shame

Engagement is a well-trod concept in the marketing academy and in practice, and the wide adoption of social media across businesses and society has provided further opportunities for customers to engage directly with brands. Engagement mediated by social media has also attracted substantial research interest (see Hollebeek 2013 for a review). Across this research,

typical social media engagement behaviors such as liking or sharing content are conceptualized as behavioral outcomes of underlying engagement motivations (Brodie et al. 2011; Calder et a. 2016; Hollebeek et al. 2016; van Doorn et al. 2010). Furthermore, these conceptualizations also emphasize the role of attitudes as an antecedent to engagement (Barger et al. 2016; Kumar and Pansari 2016; van Doorn et al. 2010), suggesting that these should move together. Notably, prior research examines situations where attitudes toward the brand are expected to coincide with motivations to engage. By contrast, we study a situational motivation for social media *disengagement* that can arise despite holding a positive attitude toward the brand. Specifically, we consider the motivation to distance oneself (Lickel et al. 2005) as the driver of these behaviors.

In a social media context, distancing oneself from the brand can take the form of reduced intentions to post, but it can also rise to the level of choosing to unfollow the brand, where the potential impact can be far-reaching across an individual's social network. Thus, it is important to understand why consumers might feel motivated to distance themselves from and disengage with a brand. One source of these motivations is the feeling of shame. Shame is driven by negative self-evaluations (Niedenthal, Tangney, and Gavanski 1994), and involves negative attributions about the self (Johns, Schmader, and Lickel 2005; Tangney et al. 1998; Tracy and Robins 2004; Wicker, Payne, and Morgan 1983). Shame contrasts with guilt, which tends to occur due to specific actions an individual has taken. The difference in the generative process also leads to different responses to shame, as compared to guilt. Due to the personal sense of having done wrong that arises with guilt, its experience engenders a motivation to atone for the wrongdoing (Tangney 1999). However, as shame focuses on a flaw in a durable self, experiencing it leads to feelings of weakness and incompetence (De Hooge, Zeelenberg, and

Breugelmans 2007). Because the experience of shame leads individuals to feel unable to resolve the underlying flaw, they instead attempt to repair the negative feelings through emotion-focused coping (Duhachek et al. 2012). This motivates them to withdraw and distance themselves from the situation, alleviating the negative evaluations of the self (De Hooge et al. 2007).

Prior work has shown that individuals experience shame vicariously when observing the actions of another person who shares their social identity (Brown and Cehajic 2008; Brown et al. 2008). Individuals who identify with a group may experience vicarious shame because of another group member who behaves in a socially unacceptable way. Experiencing shame motivates them to distance themselves from the shared identity (Iyer, Schmader, and Lickel 2007; Lickel et al. 2005), and these effects are specific to negative information coming from in-group members (Doosje et al. 2006). For example, in the wake of 9/11, the extent to which an individual identified with the American identity predicted the degree of vicarious shame and motivation to distance they experienced after finding out a fellow American transgressed by being prejudicial (Johns et al. 2005). Additional research suggests that self-threat is key in the elicitation of vicarious shame (Piff, Martinez, and Ketner 2012; Welten, Zeelenberg, and Breugelmans 2012). Why, however, would an individual on social media experience shame when interacting with a brand?

Socially Unacceptable Brand Mentions

We argue that the widespread adoption of social media has heightened consumers' exposure to the behavior of others, and the worst of these behaviors are often made particularly salient. One key observable behavior is when others produce socially unacceptable brand mentions. We define socially unacceptable brand mentions as acts of violation of the implicit or

explicit rules guiding consumer interactions *by an individual mentioning the brand* (e.g., in Liu et al [2019]'s conceptualization of consumer-based offenses, an offending party is usually another consumer). These behaviors are distinct from "brand transgressions," which are "acts of violation of the implicit or explicit rules guiding consumer-brand relationship performance and evaluation," *by the brand itself* (Khamitov, Grégoire, and Suri 2020). Norms that guide consumer interactions are widely held on social media. For example, consumers hold a norm that excessively using profane language is considered socially unacceptable in public places generally (Bostrom, Baseheart and Roisster 1973) and on social media specifically (Feldman et al. 2017; Lafreniere, Moore and Fisher 2022; Sood, Antin, and Churchill 2012). These norms can also be more contextual within the communities surrounding a brand's platform (e.g., what constitutes an appropriate or inappropriate post may depend on additional factors other than profanity), so that evidence of a violation may more readily be observed in the community's response, such as through "down-votes" on social media. Socially unacceptable brand mentions are a violation of these implicit or explicit rules guiding consumer interactions by an individual mentioning the brand, and we focus on how consumers respond when witnessing these behaviors on social media. For example, a post politely expressing disagreement with someone who started a thread on social media might not be seen as socially unacceptable. However, a post lodging a profanity-laden ad hominem attack against the original poster is seen as socially unacceptable, as would be a similarly profane attack against a competing brand.

Since disengagement is driven by a motivation to distance oneself from a brand, and one reason this motivation arises is because of the shame-inducing behavior of other people, we suggest that social media disengagement may arise when an individual observes others mentioning the brand in a socially unacceptable way. Critically though, other people's behavior

generates vicarious shame when that behavior reflects poorly on their shared social identity. We note that not all individuals will feel a shared identity with other people on the internet, and we argue that the effects of socially unacceptable brand mentions will be experienced disproportionately by those who have integrated the brand into their identity to a greater degree.

Self-Brand Connections

Consumers often incorporate brands into their self-concepts, or their sense of self, resulting in rich relationships in which the consumer experiences emotional attachment (Escalas and Bettman 2003). Consumers who have integrated brands into their self-concepts are more likely to participate in brand communities (Algesheimer, Dholakia and Herrmann 2005) and to participate in brand-related activities on social media such as posting about the brand (Grewal, Stephen, and Coleman 2019). Further, if these connections are strong enough, consumers will readily defend brands they are connected to in the same way they would defend their loved ones or themselves. Research by Lisjak et al. (2012) demonstrates that after exposure to negative information about a brand in the form of a critical editorial, individuals higher in brand identification experience less negative attitude change than individuals whose identification with the brand is weaker. Similarly, Ahluwalia et al. (2000) suggest that brand commitment moderates the effect of negative publicity because committed consumers are resistant to counter-attitudinal information, and Swaminathan et al. (2007) identify counter-arguing as a mechanism by which individuals higher in self-concept connection resist the influence of negative brand information. Ultimately, consumers are motivated to discount incoming counter-attitudinal information regarding the brands to which they are highly connected.

In addition to defending against negative brand information, more highly connected consumers are also motivated to defend against other brand users who behave in socially unacceptable ways. Ferraro, Kirmani, and Matherly (2013) show that flaunting behavior negatively affects brand evaluations, but this effect is attenuated among individuals higher in self-brand connection, proposing an attitudinal transfer explanation where the attitude toward the socially unacceptable brand user transfers to the brand.

In contrast to this prior work finding weaker negative effects of socially unacceptable behaviors on brand attitudes for individuals higher in self-brand connection, we expect that social media disengagement will exhibit a different pattern of results owing to its distinct motivational driver. While prior research has shown that attitudes are predictive of engagement behaviors, attitudes are not the only antecedent, and motivational factors that can arise situationally also play a role (Kumar and Pansari 2016; van Doorn et al. 2010). We have argued that disengagement behaviors are driven by the motivation to distance oneself, which can arise due to situationally induced vicarious shame. While attitudes concern overall *evaluations* of the brand (Eagly and Chaiken 1993), social media disengagement is different in that it concerns the reduction in *expressions of association* with the brand. Thus, even if highly connected users may retain positive attitudes toward the brand, they may nonetheless choose to disengage from it on social media to attenuate the experience of identity threat.

We predict that socially unacceptable brand mentions will lead to vicarious shame for those higher in self-brand connection, which will in turn lead to social media disengagement. Our basis for this prediction stems from social identity theory, which suggests that in addition to a personal sense of identity, individuals also maintain identities related to the groups to which they belong (Tajfel and Turner 1979). Like personal identities, social identities can be expressed

through products and brands, and can be subject to identity threats in the same manner as the personal identity (White and Argo 2009). Consistent with this line of research, we argue that the brand platform community is an additional source of social identity. We propose that socially unacceptable brand mentions can threaten this social identity and generate vicarious shame (Doosje et al. 2006; Lickel, Schmader, and Spanovic 2007). Thus, we expect that the experience of vicarious shame when observing socially unacceptable brand mentions will be more likely for those who could view the brand platform community as an additional source of identity – in other words, for those higher in self-brand connection.

Taken together, we expect socially unacceptable brand mentions to constitute a social identity threat by introducing negative information about the brand-related social identity (Reed and Forehand 2016). In the context of more (as compared to less) socially unacceptable brand mentions, we expect a greater sense of vicarious shame for individuals higher in self-brand connection. The experience of vicarious shame motivates these individuals to distance themselves from the brand, and subsequently disengage from it on social media. In line with our definition of social media disengagement, we operationalize this construct in two general ways across our studies: 1) in Twitter data, as a behavioral outcome in unfollowing the brand, and 2) in experiments, measured as self-reported level of motivation to disengage and disengagement intentions. Formally, we predict:

H1: More (vs. less) socially unacceptable brand mentions will increase disengagement more for individuals higher (vs. lower) in self-brand connection.

H2: More (vs. less) socially unacceptable brand mentions will increase vicarious shame more for individuals higher (vs. lower) in self-brand connection.

H3: Vicarious shame will mediate the interactive effect of self-brand connection and socially unacceptable brand mentions on disengagement.

Research Overview

We present the results of three studies. In study 1, using real-world data from Twitter, we demonstrate that socially unacceptable brand mentions cause highly connected users to disengage by unfollowing brands (H1). In study 2, we test the psychological process, showing that vicarious shame drives consumers' level of motivation to disengage. We find that individuals with higher (vs. lower) levels of self-brand connection experience elevated levels of vicarious shame and disengagement motivation after observing socially unacceptable brand mentions (H1-3). We also examine alternative emotional responses as competing mechanisms. Finally, in study 3, we explore whether a potential managerial intervention, the moderation of posts with socially unacceptable media brand mentions, can mitigate the effect on disengagement intentions. We further show that our effects are separate from those on attitudes.

Study 1

The goal of our first study was to provide initial evidence that elevated levels of socially unacceptable brand mentions on social media would lead to disengagement by highly connected users. To do so, we examined sports brands on Twitter, collecting data on real-world user decisions to disengage by unfollowing brands.

Data Collection

The focal brands in this study were the 10 Major League Baseball (MLB) teams who competed in the 2018 postseason. During the postseason, each team played between one and 16 games against between one and three opponents, for a total of 33 games. Beginning the day before the first postseason game, we recorded all followers for each team via the Twitter API. The followers were recorded at the same time on each subsequent day until after the final game of the season, totaling 41 days, which permitted daily identification of those who had followed or unfollowed each team. In total, more than 15.5 million unique Twitter accounts were observed to have been following at least one team for at least one day during the observation period.

Among the observed accounts, 88% followed only one of the ten teams on Twitter. In our analysis we include accounts following multiple brands, with each individual's brand-specific variables calculated for each followed brand and incorporating individual and brand fixed effects. Overall, individuals' following status changed relatively infrequently, with only 1.4% of accounts having unfollowed a brand during the observation period. Although unfollowing was a rare event as a percentage of all followers, the absolute number of followers lost (210,761) combined with however many of *their* followers who would have seen retweets represents a practically important reduction in the potential reach of the brand on Twitter. Given the relative rarity of unfollowing events and rate limitations for data collection, we adopted a case-control design (Peng et al. 2018), which enabled us to make comparisons between the set of users who unfollowed a brand to a sample of users who did not. The focal set of 661,014 account-brand pairings (from 604,330 unique individuals) includes: 1) all 210,761 account-brand pairings who unfollowed a brand during the observation period, and 2) a simple random sample of 450,253 account-brand pairings who followed a team continuously throughout the observation period and had tweeted at least once during the regular season.

We selected our control sample to be approximately double the size of the case group, per recommendations from epidemiology research where case-control designs are common (Lewallen and Courtright 1998). In addition to the limited statistical benefit, the collection of additional observations for the control set of users who did not unfollow would have been prohibitive due to Twitter API rate limitations: at that time, gathering data on all 15.5 million individuals would have taken more than 10 years. To account for this oversampling of cases, we employed inverse probability weights in our model estimation (Borgan et al. 2000).

We then collected all tweets from the focal set of users for the period beginning at the start of the 2018 MLB season and running through the end of the observation period, totaling more than 49 million tweets. These tweets were divided into two periods: 1) tweets from the period prior to the post-season that we used to operationalize self-brand connection, and 2) tweets from the post-season period that we used to operationalize socially unacceptable brand mentions. The unit of analysis was the account-brand-day for the observation period, discretized based on the time of the follower checks which were begun each day at 2:00 A.M. U.S. Eastern standard time.

Measures

Disengagement. We operationalized our dependent measure as an indicator for whether the account unfollowed the team during the previous time period.

Self-brand connection. We operationalized self-brand connection at the account-brand level, based on the team they were identified as following. For those accounts that followed multiple teams, separate measures were created for each account-brand pairing. First, we examined each individual's tweets throughout the 2018 MLB regular season and counted the

number of tweets with body text that included the Twitter handle of the team they were following, or tweets that were retweets of the team's account. For a user who was following the Houston Astros at the beginning of the post season, their self-brand connection would be based on mentions of the word "astros" in their tweets, or retweets of the account @astros. We based this measure on the intuition that higher levels of activity involving the brand over a long period of time would reflect a higher level of connection to the brand (Decrop and Derbaix 2010). We log-transformed the resulting count of tweets, then divided by 100, to ease interpretation of the resulting regression coefficients, though the results are unaffected by this transformation.

To validate this measure of self-brand connection, we surveyed a different set of users on Twitter via direct message who followed sports brands that had recently won championships in major U.S. sports. Participants completed the self-brand connection scale (Escalas and Bettman 2003). We received 152 complete responses. For each respondent, we collected all of their tweets from the previous year and calculated the team mentions measure of self-brand connection using the same procedure as in the main study. The survey measure was reliable (α = .95), and the correlation between the team mentions measure and survey measure was acceptable (r = .30, p < .001). Further, other aggregations of Twitter activity provide additional support for the measure's validity. The number of tweets mentioning the sport or league and the overall number of tweets during the same time period were both uncorrelated with the survey measure (r = .01, p = .91; and r = .02, p = .86; respectively). However, the number of tweets mentioning the team's eventual championship opponent were negatively correlated with the survey measure (r = -.19, p = .02), supporting the validity of our approach to operationalizing self-brand connection based on tweeting behavior.

Socially unacceptable brand mentions. Socially unacceptable brand mentions were operationalized by profanity use and measured at the brand-day level. To capture the use of profanity in brand mentions, we counted tweets containing profanity that specifically mentioned the focal team through either a message directed "at" the team, or through a hashtag (e.g., "Wooooooooooooooo! LET'S FUCKIN GO. LET'S MOTHER FUCKING GO! #RedSox"). We employed a dictionary to identify profanity in the text derived from the Wiktionary list of English profanity (see Web Appendix for full list). We log-transformed the resulting count of tweets, then divided by 100, to ease interpretation of the resulting regression coefficients.

Total Volume. The overall volume of Twitter posts by brand followers and the volume of posts containing socially unacceptable brand mentions are correlated, but intuitively, we would expect these types of posts to have opposite effects on disengagement. Overall volume, indicating a thriving discussion, is likely to have a negative effect on disengagement, while our theory predicts that socially unacceptable brand mentions should increase disengagement. To cleanly identify the effect of socially unacceptable behavior, we include the log-transformed count (divided by 100) of the tweets of the focal followers of the brand as a control.

Empirical Estimation

As we are interested in the effects of exposure to socially unacceptable brand mentions on disengagement (the decision to unfollow a brand) over time, we adopted a discrete-time hazard approach (Singer and Willet 1993; Chandrasekaran and Tellis 2011), which allows us to capture the probability that, in each time period, an individual would unfollow the brand, conditioned on the fact that they had not already unfollowed it. Adopting this approach allows us to incorporate the time-varying effect of socially unacceptable brand mentions that individuals

would be exposed to in a particular period. However, the analysis is complicated by the fact that the effect of interest in our model is the interaction of self-brand connection and socially unacceptable brand mentions. Given known issues with the interpretation of interaction effects in non-linear models, such as the logit transformation typically used in modeling a discrete-time hazard process, we employ a linear probability model, an approach used in prior research (Wang et al. 2019). The linear probability model eases interpretation of the results and the tractability of estimation but can introduce issues with heteroskedasticity. To address this, we employ robust standard errors, clustered at the individual level. The base model to be estimated is:

$$Prob(y_{ijt} = 1 \mid y_{ijt-1} = 0) = \beta_0 + \beta_1 C_{jt} + \beta_2 S_{ij} + \beta_3 C_{jt} \times S_{ij} + \beta_4 V_{jt} + \boldsymbol{\Gamma}' \boldsymbol{G_{jt}} + \boldsymbol{\Theta}' \boldsymbol{A_{ijt}} + \mu_i + \tau_j + \epsilon_{ijt}$$

i 1,… 604,330 indexes the 604,330 individuals

j 1,…10 indexes the 10 MLB team brands

t 1,…41 indexes the 41 days of the observation period

y_{ijt} Indicator that is 1 if individual i unfollowed brand j in the period [t-1, t], 0 else

C_{jt} Socially unacceptable brand mentions by brand j followers in the period [t-1, t]

S_{ij} Self-brand connection for individual i for brand j

$C_{jt} \times S_{ij}$ Interaction between C_{jt} and S_{ij}

V_{jt} Total tweets by brand j followers in the period [t-1, t]

$\boldsymbol{G_{jt}}$ Vector of indicators for each brand j in each game matchup played at time t, defined as $\mathbf{G_{jt}} = (G_{jtm=1}, G_{jtm=2}, \ldots, G_{jtm=66})$ where G_{jtm} is 1 if brand j played in the brand-game matchup indexed by m at time t, 0 else

$\boldsymbol{A_{ijt}}$	Vector of indicators for time at risk for individual *i* for brand *j* at time *t,* defined as $\mathbf{A_{ijt}} = (A_{ijtn=2}, A_{ijtn=3}, \ldots, A_{ijtn=41})$ where A_{ijtn} is 1 if individual *i* had been following brand *j* for *n* days in the observation period at time *t*, 0 else
μ_i	Fixed effect for individual *i*
τ_j	Fixed effect for brand *j*
ϵ_{ijt}	Disturbance term
$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_4)$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Theta}$	Parameters to be estimated

Matchup Indicators. Regarding the matchup indicators included in our model, 33 games were played in the 2018 post-season, each featuring two teams. We estimate separate parameters for each team in each game because of the high likelihood the game would have a differential impact on followers of the competing teams. Using this approach captures the game outcome, as well as other idiosyncratic factors about the team and its performance in each game. A close loss, such as the New York Yankees' 4-5 loss in Game 1 against the Boston Red Sox in the American League Division Series, would likely have a different impact on the decision to unfollow the team for a Yankees fan than for a Red Sox fan. For the Yankees fan, that close loss might also feel quite different than a blowout, such as their 1-16 showing in Game 3 of the same series. Similarly, the Game 1 loss had the same margin as the 3-4 loss in Game 5, but that later game also eliminated the Yankees from the postseason, highlighting the possible differential effects of each game for both teams. Therefore, each matchup is represented by two indicators – one set to 1 for followers of team one on the date of the game and set to 0 otherwise, and a second set to 1 for followers of team two on the date of the game and set to 0 otherwise. Aside from the day the game was played for that team, they are set to 0, meaning they can also all take the value 0 if no

game was played on day *t*, for 66 total indicators, $\boldsymbol{G_{jt}}$. Although these matchup indicators share the same index as socially unacceptable brand mentions and total volume, note they are not collinear as the matchup indicators do not uniquely combine to represent every brand-day.

Time at Risk Indicators and Fixed Effects. To accommodate the possibility of nonlinear effects of time at risk (from 2 to 41 days at risk), time at risk is represented by 40 indicators – the first set to 1 when individual *i* following brand *j* on day *t* had been following for 2 days and set to 0 otherwise (the constant in our model subsumes the indicator for following for 1 day, since that is when the observation would enter our sample), the second set to 1 when they had been following for 3 days and set to 0 otherwise, and so on, for 40 total indicators, $\boldsymbol{A_{ijt}}$. Our approach allows us to accommodate individuals who began following a team after day 1. Although these time at risk indicators share the same index as our dependent variable and error term, note they are not collinear as the time at risk indicators do not uniquely combine to represent every individual-brand-day. We also included the 604,330 individual and 10 brand fixed effects, μ_i and τ_j, respectively. These capture individual- and brand-specific time-invariant heterogeneity.

Results

The full results are presented in table 1. In the base model (table 1 column 1), we observed the predicted, positive interaction effect of self-brand connection and socially unacceptable brand mentions ($b = .530$, $t(604{,}329) = 41.88$, $p < .001$), supporting H1. Although it is typical to inspect marginal effects at ±1 SD from the mean (Spiller et al. 2013), this is most appropriate when the moderator is normally distributed. However, as is common in social media data, the distribution of activity is highly skewed, with only 4.52% of accounts tweeting about the brand at least once during the study period, necessitating an alternative approach for our

marginal effects analysis. Therefore, we identified individuals as having lower self-brand connection when they had zero tweets mentioning the team, and higher self-brand connection was identified as the mean among those who tweeted about the team (3.53 tweets). Marginal effects analysis revealed that, for users with lower self-brand connection, the effect of socially unacceptable brand mentions on the likelihood of disengagement was positive and significant (b = .004, z = 28.99, p < .001). At higher self-brand connection, the effect on disengagement was positive as well (b = .010, z = 52.33, p < .001), and the increase between these two points was also significant. This implies that increases in socially unacceptable brand mentions were associated with increased disengagement to a higher degree for those with higher self-brand connection. Shown graphically in Figure 1, as expected, both lower and higher self-brand connection individuals were increasingly likely to disengage as socially unacceptable brand mentions increased, but the effect of more socially unacceptable brand mentions was stronger among those with higher self-brand connection.

Robustness Checks

Alternative sample. As mentioned earlier, a minority of individuals followed more than one of the ten brands. We accounted for this with our modeling approach, and we also note that our results are substantively unchanged if we: 1) treat each account-brand pairing independently, with fixed effects for each account-brand pairing (Web Appendix table W5 column 1), 2) exclude individuals who followed multiple brands (Web Appendix table W5 column 2), or 3) randomly select and assign the individual to only one of the teams that they followed (Web Appendix table W5 column 3).

Alternative self-brand connection measures. One potential concern is that results using our team mentions measure of self-brand connection may be driven by a long tail of heavy tweeters, since many of the individuals who followed the focal brands did not tweet about the brands prior to the start of the post-season. As we noted previously, 4.52% tweeted about the brand at least once during the study period, resulting in an average number of brand-mentioning tweets per account of .346, but the distribution was highly positively skewed (6.83). Due to the skew in the distribution, these heavy tweeters may exert sufficient leverage to bias our estimates. To test for this possibility, we created an indicator for having ever tweeted about the brand prior to the post-season (in essence, Winsorizing the self-brand connection measure), which was our first alternative measure. We also developed a second alternative measure using unsupervised machine learning, which is detailed in the Web Appendix. We show consistent results using the ever-tweeted self-brand connection measure (table 1 column 2) and the machine learning based self-brand connection measure (table 1 column 3).

Alternative socially unacceptable brand user mentions measure. We note that our results replicate using the profanity list developed by SurgeAI (SurgeAI 2022; Web Appendix table W5 column 4). In addition to these dictionary-based measures, we also created a platform feedback-based measure. This measure was derived from the ways in which other platform users interact with tweets, colloquially referred to as the "ratio" (Data for Progress 2019). On Twitter, users may respond to others' posts by favoriting/liking it, retweeting it (posting the tweet on their own timeline), or by replying to it, which is displayed with the original poster's tweet. While liking and retweeting a tweet implies a degree of endorsement and a desire to share the content, replies represent feedback directed to the original poster. The ratio of likes and retweets to replies is a proxy for how well the tweet has been received. A tweet that receives many likes and retweets

relative to replies may be seen by other platform users as a socially acceptable tweet, while a tweet with many replies and few likes and retweets has been "ratioed," reflecting its perceived unacceptability (Minot et al. 2021).

For example, a less socially unacceptable tweet was "@RedSox Another piece of hardware for the city of Boston | Congratulations to the 2018 #WorldSeries Champions @redsox!" ([likes+retweets]/replies: 20), and a more socially unacceptable tweet was "Is the #postseason now best of 5?" ([likes+retweets]/replies: .07). In this tweet, an individual did not understand that the different rounds of the MLB postseason have different rules, such that the second round Division Series are best-of-five, while the Championship and World Series are best-of-seven. Thus, the ratio reflects this highly contextual understanding of why this tweet was not socially acceptable. We calculated our alternative measure of socially unacceptable brand mentions as the number of tweets mentioning the target team that received more replies than retweets (Minot et al. 2021). We log-transformed the resulting count of ratioed tweets, then divided by 100. We obtain consistent results when using the platform user feedback-based Twitter "ratio" measure for socially unacceptable brand mentions (table 1 column 4).

Alternative treatment of matchups. As we discussed previously, games present considerable heterogeneity in their potential impact on unfollowing. Though we capture this as a fixed effect, another possibility is that some games may have induced more socially unacceptable brand mentions than others. This would in essence mean that socially unacceptable brand mentions would mediate the effect of the games, in which case controlling for all the $\boldsymbol{G_{jt}}$ *x* S_{ij} interactions could ensure a better estimate of the focal C_{jt} *x* S_{ij} interaction. We run a model where we add these 66 additional interaction terms, and our results remain consistent (table 1 column 5).

Discussion

This study demonstrates our hypothesized effect on disengagement as reflected in unfollowing a brand. This finding provides a strong test of the interactive effect of socially unacceptable brand mentions and self-brand connection on consumers' decisions to disengage and emphasizes the potential negative consequences for brands.

Figure 1: Predicted Probability of Unfollowing at Levels of Socially Unacceptable Brand Mentions for Individuals with Low and High Self-brand Connection (Study 1)

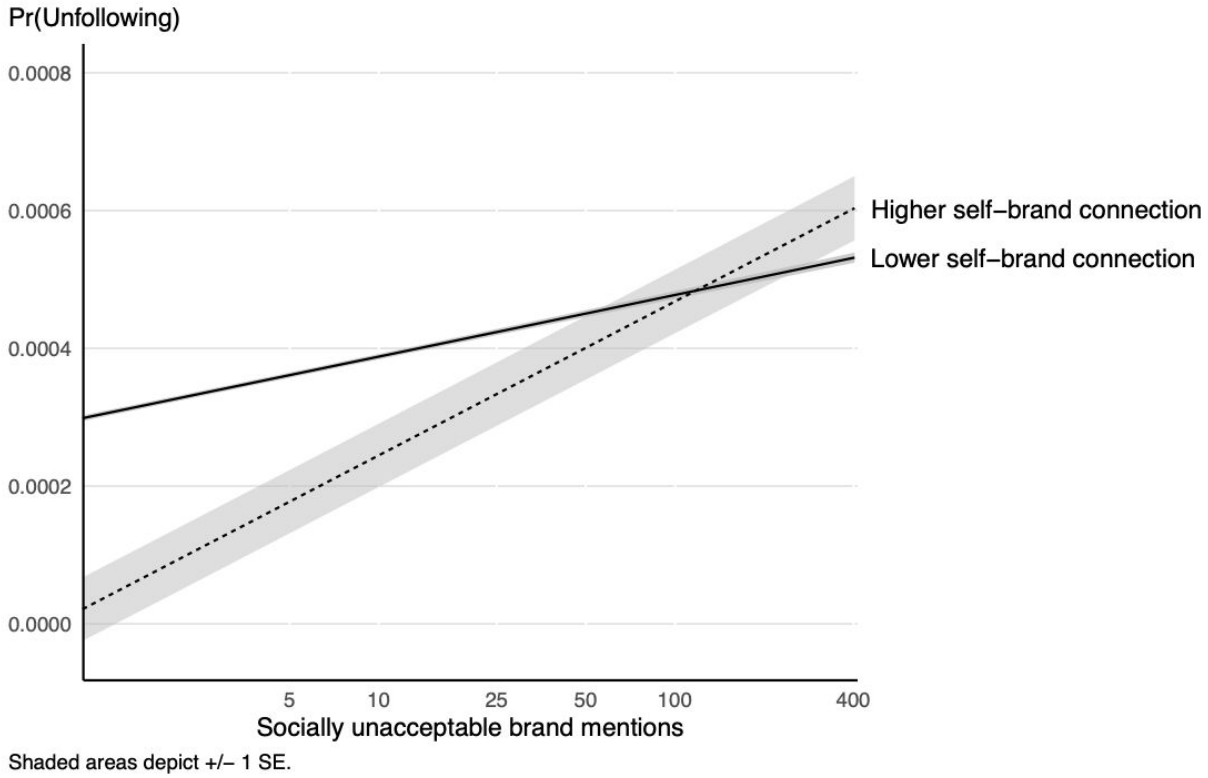


Table 1: Estimates of Relationship Between Self-Brand Connection and Socially Unacceptable Brand Mentions and Unfollowing (Study 1)

	(1) Base	(2) Ever Tweeted	(3) Machine Learning Indicator	(4) Ratioed Tweets	(5) Matchup Interactions
Self-brand connection Measure:	Team Mentions	Team Mentions Indicator	Cluster Indicator	Team Mentions	Team Mentions
Socially unacceptable brand mentions Measure:	Profane Tweets	Profane Tweets	Profane Tweets	Ratioed Tweets	Profane Tweets
Constant	0.0351***	0.0351***	0.0352***	0.0446***	0.0351***
	(0.000405)	(0.000405)	(0.000405)	(0.000447)	(0.000405)
Socially unacceptable brand mentions	0.00389***	0.00371***	0.00115***	0.0153***	0.00385***
	(0.000134)	(0.000134)	(0.000139)	(.000166)	(0.000134)
Self-brand connection	-0.0252***	-0.000403***		-0.0202***	-0.0269***
	(0.00453)	(0.0000759)		(0.00453)	(0.00454)
Socially unacceptable brand mentions X Self-brand connection	0.530***	0.0113***	0.0130***	0.241***	0.577***
	(0.0126)	(0.000193)	(0.000133)	(0.0124)	(0.0152)
Total Tweets	-0.335***	-0.336***	-0.336***	-0.427***	-0.335***
	(0.00392)	(0.00392)	(0.00392)	(0.00431)	(0.00392)
Marginal effects of Socially unacceptable brand mentions at…					
Lower Self-brand connection	0.00389***	0.00371***	0.00115***	0.0153***	0.00385***
	(0.000134)	(0.000134)	(0.000139)	(0.000166)	(0.000134)
Higher Self-brand connection	0.00970***	0.0150***	0.0141***	0.0189***	0.0102***
	(0.00185)	(0.000225)	(0.000165)	(0.000250)	(0.000204)
Matchup Indicators	Yes	Yes	Yes	Yes	Yes
Time at Risk Indicators	Yes	Yes	Yes	Yes	Yes
Individual Fixed Effects	Yes	Yes	Yes	Yes	Yes
Team brand Fixed Effects	Yes	Yes	Yes	Yes	Yes
Matchup X Self-brand connection Interactions	No	No	No	No	Yes
Observations	23,804,493	23,804,493	23,804,493	23,804,493	23,804,493
R-squared	0.07057	0.07057	0.07058	0.07059	0.07057

Notes: Robust standard errors in parentheses (clustered on the individual). Values are reported to three significant figures, except for R-squared values, which require four significant figures to show differences. *** p<0.001, ** p<0.01, * p<0.05. Column 1 contains the estimates for our base model, using the number of team mentions to measure of self-brand connection and the number of profane tweets to measure of socially unacceptable brand mentions. Column 2 reports the results from using an indicator for having ever tweeted to measure self-brand connection. Column 3 reports the results from using the machine learning based measure of self-brand connection. Because of how the machine learning approach clustered individuals based on observable characteristics, the simple effect of self-brand connection in this model was collinear with the fixed effects and was not estimable. Column 4 reports the results of using a community feedback-based "ratio" measure of socially unacceptable brand mentions. Column 5 reports the results from a model including interactions of matchups and self-brand connection, controlling for an additional source of heterogeneity.

Study 2

The goal of our second study was three-fold. First, we employed an experimental context to exogenously manipulate socially unacceptable brand mentions. Second, we sought to explore the process (vicarious shame) that led to the disengagement outcome observed in the first study. Finally, we also addressed several potential alternative accounts for our results, by measuring brand ownership and alternative emotions. Although both our theory and previous research (Ferraro et al. 2013) suggest brand ownership is not required for the brand to be incorporated into the social self-concept, we want to confirm this by measuring and controlling for brand ownership. We also measure and control for alternative emotions observers could experience that could operate as competing mechanisms to our proposed process of experiencing vicarious shame. We expected that highly connected users who were exposed to socially unacceptable brand mentions would become more motivated to disengage with the brand, and that this effect would be mediated by vicarious shame.

Method

The study employed a 2 (socially unacceptable brand mention: less, more) x 2 (brand: Patriots, Rams) between-subjects randomized factorial design, with participants randomly assigned to one of the four cells, and self-brand connection as a measured factor. Because we employed real brands, we sought to rule out potential idiosyncratic brand effects as the driver of our results, and thus we employed the two brands to serve as replicates. We conducted this study in two phases during the 2019 National Football League (NFL) playoffs, with self-brand connection measured one week prior to the main study. The NFL is composed of two conferences that have separate playoff brackets, and the conference champions compete in the

Super Bowl. Two days after the conference semi-final games, we recruited participants in the first phase. We targeted 400 responses in the main study and estimated a 50% follow-up response rate (Matherly 2019), so we recruited 800 U.S. participants using the TurkPrime platform (now CloudResearch; Litman, Robinson and Abberbock 2017). Participants completed self-brand connection scales ([Brand] reflects who I am, I can identify with [brand], I feel a personal connection to [brand], I (can) use [brand] to communicate who I am to other people, I think [brand] (could) help(s) me become the type of person I want to be, I consider [brand] to be "me" (it reflects who I consider myself to be or the way I want to present myself to others), [Brand] suits me well; 1 – Strongly Disagree / 7 – Strongly Agree; Escalas and Bettman 2003) for each of the four teams (NFC: Los Angeles Rams and New Orleans Saints, AFC: New England Patriots and Kansas City Chiefs) contending in the then-upcoming conference championship games. Participants also indicated whether they owned any products associated with each team, along with their age and gender. In the second phase, one week later and two days after the conference championships, we contacted panelists from the original sample to participate. Once we reached our target sample size of 400, we closed the study to new participants. Due to limitations in the MTurk platform, we received responses from 403 panelists from the original sample. Responses from eleven participants who took longer than three standard deviations above the mean time to complete the study were deleted, leaving 392 completed responses (mean age = 39.5, 43.8% female).

The main study proceeded as follows. First, participants in the Patriots (Rams) condition were reminded "On Sunday, the New England Patriots (Los Angeles Rams) won the AFC (NFC) conference championship, sending them to the Super Bowl." They were told that on the next page they would see a tweet by a Twitter user. They then saw a tweet by Tom, who was a fan of

the Patriots (Rams). We manipulated social acceptability by the inclusion of profanity in the fan's tweets (graphical stimuli are available in the Web Appendix). In the less socially unacceptable brand mention condition, Tom tweeted, "AWWW YEAH!!! CONFERENCE CHAMPS! [Patriots/Rams] are headed to the SB! LET'S GO!!!!!" In the more socially unacceptable brand mention condition, Tom tweeted, "FUCK YEAH!!! CONFERENCE CHAMPS! [Patriots/Rams] are headed to the SB! LET'S FUCKING GO!!!!!"

Measures

All participants then completed a measure of disengagement motivation (I want to be completely unassociated with [brand] products, I don't want to be associated in any way with [brand] products after seeing that post, 1 – Strongly Disagree / 7 – Strongly Agree, r = .93; Schmader and Lickel 2006). One could argue the six-item distancing measure from Schmader and Lickel (2006) is composed of multiple dimensions, despite its use as a unidimensional multi-item scale in previous research (the other four items: I want to be completely unassociated with [target], I don't want to be associated in any way with [target] after seeing that post, I wish [target] weren't a [brand] user, I feel like I want to disappear from the situation). Although a factor analysis suggests this distancing response is a unidimensional construct (all loadings > .77, eigenvalue = 4.65, 71% of variance explained) and our results replicate with the full scale (that is, individuals wanted to distance from the brand and the poster), we prefer the conceptual clarity of the two brand-specific items.

We measured the dependent variable first and then the mediators, in the following order: vicarious shame (I feel ashamed, I feel awkward, [Target]'s behavior reflected badly upon me, [Target]'s behavior made me look bad, 1 – Strongly Disagree / 7 – Strongly Agree, α = .92;

Welten et al. 2012), emotions (Please rate the extent to which you are experiencing each emotion right now after seeing [target]'s tweet; 1 – Not much at all / 7 – Very much; Johns et al. 2005) for guilt (guilty, regretful, remorseful, α = .88), anger (angry, offended, disgusted, α = .90), anxiety (nervous, anxious, r = .86), sadness (hurt, depressed, sad, upset, disappointed, α = .89), and positive affect (proud, good, happy, α = .94). They rated the social acceptability of the target's behavior using two semantic differential scales (To what extent do you think [target]'s behavior was… 1 – Inappropriate / 7 – Appropriate, 1 – Socially unacceptable / 7 – Socially acceptable, r = .90). The participant's self-brand connection from the first phase of the study for the team whose tweet they were exposed to served as the self-brand connection measure used in the main study (α = .90). That is, for participants who saw the tweet of a Patriots (Rams) fan, we used the participants' self-brand connection for the Patriots (Rams) in the subsequent analysis.

Results

A regression including indicators for socially unacceptable brand mention, brand, self-brand connection and all two and three-way interactions indicated non-significant three-way interactions on disengagement (t(383) = -1.01, p = .311) and vicarious shame (t(383) = -1.48, p = .139). Thus, the two-way interactions of theoretical interest were not significantly different across the two teams. Therefore, we pooled across teams for the analyses, while including a team indicator as a covariate. This covariate both reduces error variance and obtains a clean test of the focal effect of self-brand connection, controlling for any team-specific differences. Additionally, we include an indicator for product ownership to rule this out as a potential confound, but our results are unaffected by its inclusion as a covariate.

Manipulation Check. We regressed social acceptability on indicator-coded socially unacceptable brand mention, mean-centered self-brand connection, and their interaction as independent variables along with indicators for product ownership and brand replicate as controls. The analysis revealed a significant effect of the socially unacceptable brand mention manipulation ($t(386) = -13.71$, $p < .001$), with lower levels of social acceptability in the more socially unacceptable brand mention condition ($M = 3.61$) compared to the less socially unacceptable brand mention condition ($M = 5.80$), indicating a successful manipulation of social acceptability. The interaction effect of socially unacceptable brand mention and self-brand connection was not significant ($p = .837$).

Disengagement motivation. We regressed disengagement motivation on the same set of predictors. The analysis revealed significant simple effects of self-brand connection ($t(386) = -3.51$, $p < .001$) and socially unacceptable brand mention ($t(386) = 1.98$, $p = .049$), qualified by a significant interaction effect ($t(386) = 2.05$, $p = .042$), supporting H1. Although it is typical to inspect simple effects at ±1 SD from the mean (Spiller et al. 2013), -1 SD would have been below the scale minimum; thus, we use the scale minimum (-.74 SD) as our lower self-brand connection value. When self-brand connection was lower, the more (vs. less) socially unacceptable brand mention did not significantly increase disengagement motivation ($M_{less} = 3.48$, $M_{more} = 3.56$, $b = .08$, $t(386) = .33$, $p = .744$), but when self-brand connection was higher (+1SD), the more (vs. less) socially unacceptable brand mentions significantly increased disengagement motivation ($M_{less} = 2.53$, $M_{more} = 3.32$, $b = .79$, $t(386) = 2.82$, $p = .005$).

Vicarious shame. We regressed vicarious shame on the same predictors. The analysis revealed significant simple effects of self-brand connection ($t(386) = 2.20$, $p = .028$) and socially unacceptable brand mention ($t(386) = 5.18$, $p < .001$), qualified by a significant interaction effect

($t(386) = 3.43$, $p < .001$), supporting H2. When self-brand connection was lower, the more (vs. less) socially unacceptable brand mention significantly increased vicarious shame ($M_{less} = 1.33$, $M_{more} = 1.62$, $b = .29$, $t(386) = 2.04$, $p = .042$), but when self-brand connection was higher (+1 SD), the more (vs. less) socially unacceptable brand mention significantly increased vicarious shame even more strongly ($M_{less} = 1.68$, $M_{more} = 2.66$, $b = .98$, $t(386) = 6.03$, $p < .001$).

Mediation. We tested the overall model relating vicarious shame and disengagement to the effects of self-brand connection and the socially unacceptable brand mentions using bootstrapping. We employed the PROCESS macro with 5,000 resamples to generate confidence intervals for the indirect effect, as recommended by Hayes (2013). The index of moderated mediation for vicarious shame was significant ($b_{index} = .19$, SE = .07, 95% CI: [.06, .34]), supporting H3. We observed a significant conditional indirect effect of the more (vs. less) socially unacceptable brand mention on disengagement through vicarious shame when self-brand connection was lower ($b = .19$, SE = .08, 95% CI: [.03, .36]). The indirect effect was even stronger when self-brand connection was higher ($b = .66$, SE = .15, 95% CI: [.38, .98]).

These results were robust to the inclusion of the five other emotion measures as competing (i.e., parallel) mediators ($b_{index,\ vicarious\ shame} = .07$, SE = .04, 95% CI: [.00, .18]; $b_{lower,\ vicarious\ shame} = .07$, SE = .04, 95% CI: [.00, .16]; $b_{higher,\ vicarious\ shame} = .25$, SE = .12, 95% CI: [.03, .51]). We note that, although these confidence intervals are in the "Goldilocks Zone" in proximity to zero, we repeated the bootstrapping procedure 1000 times and only .2% of the confidence intervals contained zero. Based on the recommendations of Götz et al. (2021), we further triangulated our estimate of the indirect effect by constructing Monte Carlo confidence intervals, which did not contain zero (95% MC CI: [.018, .135]). Taken together, these results suggest that our estimate of the indirect effect is not a statistical artifact.

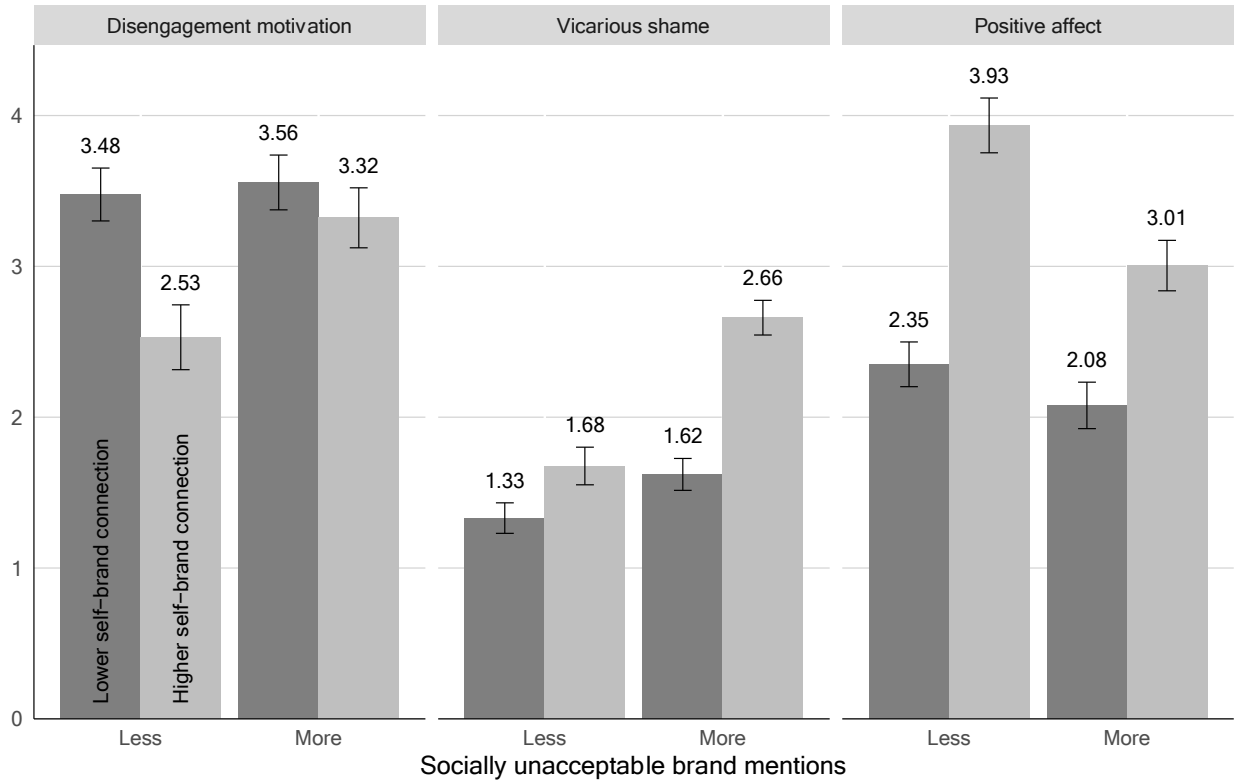
The only other emotion that operated similarly (i.e., had a significant index of moderated mediation in the parallel mediation model) was positive affect (b_{index} = .06, SE = .03, 95% CI: [.01, .12]). Inspecting the first-step regression with positive affect as the dependent variable revealed significant simple effects of self-brand connection ($t(386) = 6.94$, $p < .001$) and socially unacceptable brand mention ($t(386) = -3.36$, $p < .001$), qualified by a significant interaction effect ($t(386) = -2.23$, $p = .026$). When self-brand connection was lower, the more (vs. less) socially unacceptable brand mention non-significantly reduced positive affect (M_{less} = 2.35, M_{more} = 2.08, $b = -.27$, $t(386) = -1.32$, $p = .19$), but when self-brand connection was higher (+1 SD), the more (vs. less) socially unacceptable brand mention more strongly and significantly reduced positive affect (M_{less} = 3.93, M_{more} = 3.01, $b = -.93$, $t(386) = -3.92$, $p < .001$). Accordingly, the conditional indirect effect of the more (vs. less) socially unacceptable brand mention on disengagement through positive affect was non-significant when self-brand connection was lower (b = .06, SE = .05, 90% CI: [-.02, .15]), and the conditional indirect effect was significant when self-brand connection was higher (b = .21, SE = .07, 95% CI: [.10, .36]).

Discussion

The results of our second study conceptually replicate the findings of study 1, showing that users who are higher in self-brand connection are motivated to disengage from the brand in the face of socially unacceptable brand mentions. We also find that this is mediated by the experience of vicarious shame, and that, except for the relationship with positive affect, the results cannot be explained by other emotions. Most notably, we do not find support for guilt as a mediator, suggesting that the process is distinctly related to the vicarious experience of shame, as our theory predicts. However, one limitation of this study was that, in the interest of using an

established scale from the social identity literature, we measured general disengagement motivation as opposed to social media-specific disengagement intentions. Thus, it is possible that, by measuring general disengagement motivation, the role for positive affect may have been stronger than if we had measured social media-specific disengagement intentions. We address this limitation in our next study.

Figure 2: Disengagement Motivation, Vicarious Shame, and Positive Affect by Socially Unacceptable Brand Mention Condition for Individuals with Lower and Higher Self-brand Connection (Study 2)



Error bars depict +/− 1 SE.

Study 3

In our final study, we sought to demonstrate the full process again experimentally, from socially unacceptable brand mentions through vicarious shame to social media disengagement intentions. We also sought to separate our effects from any relationship with brand attitudes,

which have shown in prior work (Cheng, White, and Chaplin 2012; Ferraro et al. 2013). Finally, we explored whether a brand's response to socially unacceptable brand mentions can potentially mitigate the risks of disengagement for highly connected individuals. We focused on those higher in self-brand connection and expected that a brand actively responding to the bad behavior could potentially blunt the effect of socially unacceptable brand mentions on disengagement.

Method

We recruited 301 CloudResearch-approved (Litman et al. 2017) U.S. MTurk panelists. Responses from six participants who took longer than three standard deviations above mean time to complete the study were deleted, leaving 295 completed responses (mean age = 39.3, 41% female).

The study employed a 3 cell (socially unacceptable brand mention: less, more, more with brand reaction) randomized between-subjects design. First, to ensure participants had higher self-brand connection, we adopted a procedure from Dagogo-Jack and Forehand (2018). Participants ranked five sportswear brands (Nike, Adidas, Under Armour, Puma, Reebok) based on which ones they felt they had the most personal connection with. We used the top-ranked brand as the focal brand, customized to each participant. They were told that on the next page they would see a post from the social media website Reddit and to assume they were generally active in the brand's subreddit/discussion forum. They then saw a Reddit post in the focal brand's subreddit forum by a user, Alex. As in the prior study, we manipulated social acceptability by the inclusion of profanity in Alex's post (graphical stimuli are available in the Web Appendix). In the less socially unacceptable brand mention condition, Alex posted (for a participant who top-ranked

Nike), "I don't agree with the original poster. Anyone who doesn't like Nike is really missing out." In the more socially unacceptable brand mention condition, Alex posted, "The original poster is such a fucking dumbass! Anyone who doesn't like Nike deserves a flaming bag of dog shit on their porch." In the reaction condition, participants saw the same post as in the more socially unacceptable condition, but on the next page were told that Alex's post had been removed by the brand's moderator team and that he had been banned from future posting.

Measures

All participants then completed measures of disengagement intentions (I am likely to decrease my engagement with the [brand] subreddit, I am likely to decrease my posting in the [brand] subreddit, I am likely to decrease my reading in the [brand] subreddit, I am likely to unsubscribe from the [brand] subreddit, 1 – Strongly Disagree / 7 – Strongly Agree, α = .96; based on Hollebeek, Glynn, and Brodie 2014), vicarious shame (α = .91, same as in Study 2), brand attitudes (Please rate your attitudes towards the [brand] brand. 1 – Dislike / 7 – Like, 1 – Bad / 7 – Good, 1 – Unfavorable, 7 – Favorable, α = .96; Ferraro et al. 2013), social acceptability (r = .94, same as in Study 2), brand response (How strong do you think the [brand] brand's moderator team's response was, 1 – No response at all / 7 – Very strong response), and self-brand connection (α = .95, same as in Study 2), in that order.

Results

Discriminant validity. Although we expected disengagement intentions and attitudes to behave differently, to further distinguish these concepts, we also conducted a discriminant validity analysis of these measures. The correlation between disengagement intentions and

attitudes was r = -.28, providing a strong indication of discriminant validity in these measures. To further underscore this, the average variance extracted (AVE) for each scale (disengagement intentions = .87, attitudes = .88) was higher than their squared correlation (.08), indicating satisfactory discriminant validity (Fornell and Larcker 1981).

Self-brand connection. An ANOVA showed the effect of socially unacceptable brand mention condition on self-brand connection was non-significant (F(2, 292) = .17, p = .844), indicating that our approach ensured a higher degree of self-brand connection (M = 4.02, SD = 1.56) but did not exhibit differences across the conditions.

Manipulation checks. An ANOVA with planned contrasts revealed a significant effect of the socially unacceptable brand mention manipulation on social acceptability (F(2, 292) = 113.38, p < .001). Participants perceived lower levels of social acceptability in the more socially unacceptable brand mention condition (M = 1.77, t(292) = 13.44, p < .001) and more socially unacceptable brand mention with reaction condition (M = 1.90, t(292) = 12.62, p < .001) compared to the less socially unacceptable brand mention condition (M = 4.74). The more socially unacceptable brand mention conditions did not significantly differ from each other (t(292) = .58, p = .56), confirming a successful manipulation of social acceptability.

An ANOVA with planned contrasts revealed a significant effect of the socially unacceptable brand mention manipulation on brand response (F(2, 292) = 103.37, p < .001). Participants assumed an accompanying brand response would be marginally significantly stronger for the more socially unacceptable brand mention (M = 3.05) compared to the less socially unacceptable brand mention (M = 2.59, t(292) = 1.94, p = .054) even when we did not specify any reaction. Importantly, participants saw the brand's reaction as significantly stronger when the brand reacted by deleting the post and banning Alex, compared to the other conditions

(M = 5.79; $t(292)_{\text{vs. more}}$ = 11.51, p < .001; $t(292)_{\text{vs. less}}$ = 13.33, p < .001), confirming a successful manipulation of the brand reaction.

Disengagement Intentions. An ANOVA with planned contrasts revealed a significant effect of the socially unacceptable brand mention manipulation on disengagement intentions (F(2, 292) = 10.85, p < .001). Participants reported greater disengagement intentions in the more socially unacceptable brand mention condition (M = 3.70) compared to the less socially unacceptable brand mention condition (M = 2.51, t(292) = 4.58, p < .001), again supporting H1. The brand's reaction reduced disengagement intentions (M = 2.93, $t(292)_{\text{vs. more}}$ = 2.95, p = .003) to a level similar to that in the less socially unacceptable brand mention condition ($t(292)_{\text{vs. less}}$ = 1.58, p = .12).

Vicarious shame. An ANOVA with planned contrasts revealed a significant effect of the socially unacceptable brand mention manipulation on vicarious shame (F(2, 292) = 4.06, p = .018). Participants experienced higher levels of vicarious shame in the more socially unacceptable brand mention condition (M = 2.29) compared to the less socially unacceptable brand mention condition (M = 1.73, t(292) = 2.82, p = .005), again supporting H2. The brand's reaction did not significantly reduce this feeling (M = 2.08, $t(292)_{\text{vs. more}}$ = 1.04, p = .30), though in comparison to the less socially unacceptable brand mention condition, the difference was marginally significant ($t(292)_{\text{vs. less}}$ = 1.73, p = .083). While the brand's reaction did not completely address the emotional response resulting from the more socially unacceptable brand mention, it did stop it from translating into increased disengagement intentions.

Mediation. We tested the overall model relating vicarious shame and disengagement intentions to the effects of the socially unacceptable brand mention and the brand's reaction using PROCESS. We used the less socially unacceptable condition as the reference category and

included indicators for more socially unacceptable (more = 1, else = 0) and the more socially unacceptable with reaction (more with reaction = 1, else = 0) conditions. We observed a significant indirect effect of the more (vs. less) socially unacceptable brand mention on disengagement intentions through vicarious shame when participants did not know the brand's reaction (b = .30, SE = .12, 95% CI: [.09, .54]), again supporting H3. The indirect effect was weaker and only marginally significant when participants were told the brand had banned Alex (b = .19, SE = .11, 95% CI: [-.02, .42], 90% CI: [.02, .37]).

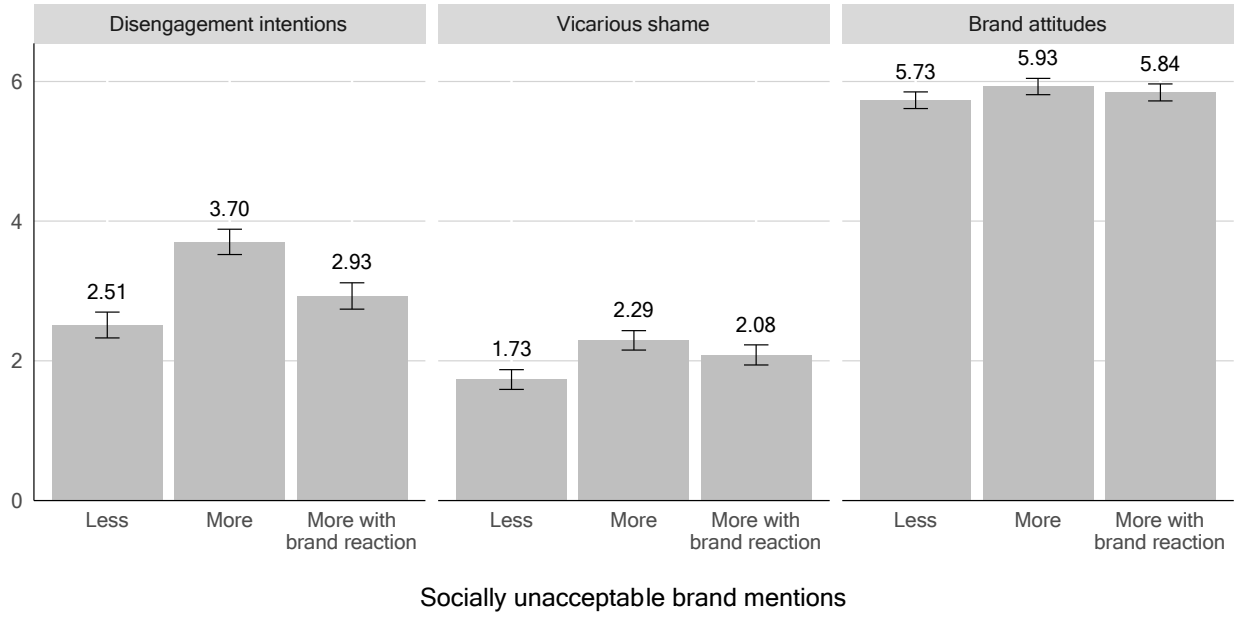
Attitudes. An ANOVA indicated the effect of the socially unacceptable brand mention condition on brand attitudes was non-significant ($F(2, 292) = .70$, $p = .50$). Attitudes toward the brand remained positive regardless of condition ($M_{less} = 5.73$; $M_{more} = 5.93$; $M_{more+reaction} = 5.84$). This pattern of results is quite different from those for the disengagement intentions and vicarious shame measures, verifying that our effects are separate from those on brand attitudes shown in prior work. As would be expected, using the previous mediation model while replacing vicarious shame with brand attitudes showed that the indirect effect on disengagement intentions through attitudes was non-significant (b = .25, SE = .23, 95% CI: [-.18, .71]), further highlighting the distinction of the path towards disengagement through vicarious shame.

Discussion

The results of our last study replicate and extend the findings of our earlier studies, showing that users who are higher in self-brand connection intend to disengage from the brand on social media in the face of socially unacceptable brand mentions, and that this relationship is mediated by the experience of vicarious shame. We also show that when the brand takes steps to

remove the socially unacceptable post, doing so reduces the likelihood of disengagement, and while it mitigates the experience of vicarious shame, it does not completely eliminate it.

Figure 3: Disengagement Intentions, Vicarious Shame, and Brand Attitudes by Socially Unacceptable Brand Mention Condition for Individuals with Higher Self-brand Connection (Study 3)



Error bars depict +/− 1 SE.

General Discussion

The goal of the present research was to explore a potential reason why consumers who are highly connected to a brand may choose to disengage from it when observing socially unacceptable brand mentions. We proposed that these behaviors threaten the identity of the consumers, that this threat leads them to experience vicarious shame, and this motivates them to disengage with the brand. Across three studies, we showed that observing socially unacceptable brand mentions leads to disengagement by individuals with stronger self-brand connections.

In study 1, we provide real-world evidence for the deleterious effect of socially unacceptable brand mentions among individuals higher in self-brand connection, showing that

these highly connected users will take steps to disengage with the brand. We provide evidence for the role of social identity threat as the underlying process driving this response, as shown in the mediation by vicarious shame in study 2. In study 3, we show that moderating a post and banning the offending poster can help mitigate the effect on social media disengagement. We further observe that these effects are robust across different brands and socially unacceptable brand mentions, and they arise even when brand attitudes remain positive.

Empirically, we show that socially unacceptable brand mentions constitute a social identity threat, expanding the scope of the process found in work by Cheng et al. (2012) and others by identifying brand mentions by social media users as a potential source of negative information. Further, our results demonstrate that, despite the importance placed on developing strong relationships by marketing managers (Connors et al. 2021), the resilience these relationships provide in the face of negative information can be limited. While higher self-brand connections help consumers maintain positive views of the brand (Ferraro et al. 2013) and may encourage adoption of higher status products from the same brand (Wang and John 2019), we find that the vicarious shame induced when observing socially unacceptable brand mentions can undermine the benefits of self-brand connection by leading to social media disengagement. Our findings also are an interesting complement to the broader recent work on brand relationships within social media (Grewal et al. 2019; Zhang and Patrick 2021).

Our results offer several important theoretical insights. First, we provide initial evidence that social media disengagement can occur even while brand attitudes are maintained. We show that disengagement motivation can be situationally induced by socially unacceptable brand mentions, which generate vicarious shame. We extend existing work on how self-brand connection affects consumers' responses to negative information. Confirming that socially

unacceptable brand mentions reflect on the brand, we show that observing these behaviors interacts with self-brand connection to increase vicarious shame and motivate disengagement. We also contribute to work on social identity threats in consumer behavior. While prior research focuses on larger, more general social groups (e.g., nationality, gender), our findings suggest that posters who mention the brand are a social group that can generate identity threats. We find that the identity threat prompted by socially unacceptable brand mentions induces vicarious shame and manifests in heightened disengagement motivation, intentions, and behaviors. We also provide corroborating evidence of in-group transgression as an antecedent to vicarious shame.

These findings further contribute to the emerging literature on consumption-based offenses. Recently, Liu et al. (2019) proposed a framework for characterizing consumption-based offenses based on whether they violate one's values, relationships, or self-views. Within this framework, we empirically identify socially unacceptable brand mentions as a (social) self-view consumption-based offense, in that it generates a social identity threat and vicariously elicits the self-conscious emotion of shame, ultimately leading to disengagement.

Our work also has practical implications for managers and marketing strategists, as our results imply that the insulating effects of strong brand relationships may not be as unequivocal as has been suggested. While prior work has shown that more highly connected consumers are able to maintain positive attitudes towards the brand when they are exposed to negative brand information, our results suggest that socially unacceptable brand mentions by others may lead these users to experience vicarious shame, which inclines them to step away from the brand. Brand managers and marketers may therefore want to avoid relying too much on the protective properties of strong self-brand connections, and to be proactive in mitigating the potential damage of in-group transgressions. We also highlight the importance of managing social media

platforms and communities to reduce these effects (as in study 3). Our results suggest that actively moderating posts with socially unacceptable brand mentions can stop these behaviors from translating into heightened disengagement intentions among higher self-brand connection consumers, complementing recent work by Liu, Yildirim, and Zhang (2022) on platform incentives for moderation in attracting and retaining users. Brands may be able to use moderation activities as an opportunity to educate consumers and help them become productive members of the community by providing explanations for why they removed content (Jhaver, Bruckman and Gilber 2019). They may also consider using platform features, such as peer awards (Burtch et al. 2021), to produce high quality content to crowd out socially unacceptable brand mentions.

Our results also present several important opportunities for future research. First, distancing is a general motivation and other cognitions and emotions that may prompt this motivation would be worthy of study. Another potential avenue involves expanding the scope of socially unacceptable behaviors. Our studies focused on socially unacceptable social media behaviors. While we generally treat the use of profanity as socially unacceptable in our contexts and find supporting evidence for this throughout our studies, we note that there can be considerable nuance in the interpretation of the use of profanity. Some recent work in marketing highlights potentially positive aspects of using profanity in conveying complex meanings and ideas (Lafreniere et al. 2022), suggesting that boundary conditions to our effects may exist. Many other types of socially unacceptable behaviors that brand mentioners or users can engage in, such as publicly disparaging the brand, either on social media or offline (Wilson, Giebelhausen, and Brady 2017) may be more ambiguous to interpret and have opaque motivations. Even more extreme behavior also may be worthy of study in the future. Likewise, given many possible approaches brands can take to respond to socially unacceptable brand

mentions, exploring the impact of these could provide useful insights to both academics and practitioners.

Another area for further study involves the scope of the concept of disengagement. We measured disengagement through unfollowing behaviors on social media and motivation/intention to disengage from the brand. Furthermore, we focused on disengagement from the brand itself as a function of the behavior of other users on the brand's platform. However, the decision to unfollow a brand on social media necessarily means also choosing not to participate in the brand community that exists around the platform. It is possible there may be scenarios where individuals may choose to distance from the community while still maintaining contact with the brand; for instance, if the community becomes "toxic." Identifying contexts that allow exploring the distinction between disengagement from the brand versus its community may provide fruitful avenues for further research.

Conclusion

Disengagement from a brand on social media is a strong response that is costly to brands, and it is important for both researchers and practitioners to understand its drivers. When consumers observe socially unacceptable brand mentions, greater self-brand connection motivates them to disengage from the brand on social media due to the experience of vicarious shame. Vicarious shame arises because socially unacceptable brand mentions introduce a social identity threat. Thus, it is important to recognize the limits of the buffering effects of self-brand connection and that the potential negative impact of socially unacceptable brand mentions may be broader than previously thought.

References

Ahluwalia, Rohini, Robert E. Burnkrant, and H. Rao Unnava (2000), "Consumer Response to Negative Publicity: The Moderating Role of Commitment," *Journal of Marketing Research*, 37 (May), 203-214.

Algesheimer, René, Utpal M. Dholakia, and Andreas Herrmann (2005), "The Social Influence of Brand Community: Evidence from European Car Clubs," *Journal of Marketing* 69 (3), 19-34.

Barger, Victor, James W. Peltier, and Don E. Schultz (2016), "Social Media and Consumer Engagement: A Review and Research Agenda," *Journal of Research in Interactive Marketing,* 10 (4), 268-287.

Borgan, Ornulf, Bryan Langholz, Sven Ove Samuelsen, Larry Goldstein, and Janice Pogoda (2000), "Exposure Stratified Case-cohort Designs," *Lifetime Data Analysis*, 6 (1), 39-58.

Bostrom, Robert N., John R. Baseheart, and Charles M. Rossiter Jr. (1973), "The Effects of Three Types of Profane Language in Persuasive Messages," *Journal of Communication*, 23 (4), 461-475.

Brodie, Roderick J., Linda D. Hollebeek, Biljana Jurić, and Ana Ilić (2011), "Customer Engagement: Conceptual Domain, Fundamental Propositions, and Implications for Research," *Journal of Service Research* 14 (3), 252-271.

Brown, Rupert, and Sabina Cehajic (2008), "Dealing with the Past and Facing the Future: Mediators of the Effects of Collective Guilt and Shame in Bosnia and Herzegovina." *European Journal of Social Psychology*, 38 (4), 669-684.

Brown, Rupert, Roberto Gonzalez, Hanna Zagefka, Jorge Manzi, and Sabina Cehajic (2008), "Nuestra Culpa: Collective Guilt and Shame as Predictors of Reparation for Historical Wrongdoing," *Journal of Personality and Social Psychology*, 94 (1), 75-90.

Burtch, Gordon, Qinglai He, Yili Hong and Dokyun Lee (2021), "How Do Peer Awards Motivate Creative Content? Experimental Evidence from Reddit," *Management Science,* articles in advance.

Calder, Bobby J., Edward C. Malthouse, and Ewa Maslowska (2016), "Brand Marketing, Big Data and Social Innovation as Future Research Directions for Engagement," *Journal of Marketing Management,* 32 (5-6), 579-585.

Chandrasekaran, Deepa, and Gerard J. Tellis (2011), "Getting a Grip on the Saddle: Chasms or Cycles?" *Journal of Marketing* 75 (4), 21-34.

Cheng, Shirley Y.Y., Tiffany B. White, and Lan N. Chaplin (2012), "The Effects of Self-Brand Connections on Responses to Brand Failure: A New Look at the Consumer-Brand Relationship," *Journal of Consumer Psychology*, 22 (2), 280-288.

Connors, Scott, Mansur Khamitov, Matthew Thomson, and Andrew Perkins (2021), "They're Just Not that into You: How to Leverage Existing Consumer-Brand Relationships through social Psychological Distance," *Journal of Marketing*, forthcoming, DOI: 10.1177/0022242920984492.

Cucu, Elena (2021), "Instagram Engagement Stats in 2021," *Socialinsider.io* (accessed February 26, 2022), www.socialinsider.io/blog/instagram-engagement

Dagogo-Jack, Sokiente W., and Mark R. Forehand (2018), "Egocentric Improvement Evaluations: Change in the Self as an Anchor for Brand Improvement Judgments," *Journal of Marketing Research*, 55 (6), 934-950.

Data for Progress (2019), "The Ratio Richter Scale," dataforprogress.org, (accessed October 7, 2022), www.dataforprogress.org/the-ratio-richter-scale

Decrop, Alain, and Christian Derbaix (2010), "Pride in Contemporary Sport Consumption: A Marketing Perspective," *Journal of the Academy of Marketing Science*, 38 (5), 586-603.

De Hooge, Ilona E., Marcel Zeelenberg, and Seger M. Breugelmans (2007), "Moral Sentiments and Cooperation: Differential Influences of Shame and Guilt," *Cognition and Emotion*, 21 (5), 1025-1042.

Doosje, Bertjan E.J., Nyla R. Branscombe, Russell Spears, and Antony S.R. Manstead (2006), "Antecedents and Consequences of Group-based Guilt: The Effects of Ingroup Identification," *Group Processes & Intergroup Relations*, 9 (3), 325-338.

Duhachek, Adam, Nidhi Agrawal, and DaHee Han (2012), "Guilt Versus Shame: Coping, Fluency, and Framing in the Effectiveness of Responsible Drinking Messages," *Journal of Marketing Research*, 49 (6), 928-941.

Eagly, Alice H., and Shelly Chaiken (1993), *The Psychology of Attitudes,* Fort Worth, TX: Harcourt Brace Jovanovich.

Escalas, Jennifer E., and James R. Bettman (2003), "You Are What They Eat: The Influence of Reference Groups on Consumers' Connections to Brands," *Journal of Consumer Psychology*, 13 (3), 339-348.

Feldman, Gilad, Huiwen Lian, Michal Kosinski, and David Stillwell (2017), "Frankly, We Do Give a Damn: The Relationship between Profanity and Honesty," *Social Psychological and Personality Science*, 8 (7), 816-826.

Ferraro, Rosellina, Amna Kirmani, and Ted Matherly (2013), "Look at Me! Look at Me! Conspicuous Brand Usage, Self-Brand Connection, and Dilution," *Journal of Marketing Research*, 50 (August), 477-488.

Götz, Martin, Ernest H. O'Boyle, Erik Gonzalez-Mulé, George C. Banks, and Stella C. Bollmann (2021), "The 'Goldilocks Zone': (Too) Many Confidence Intervals in Tests of Mediation Just Exclude Zero," *Psychological Bulletin*, 147 (1), 95-114.

Grewal, Lauren, Andrew T. Stephen, and Nicole Verrochi Coleman (2019), "When Posting about Products on Social Media Backfires: The Negative Effects of Consumer Identity Signaling on Product Interest," *Journal of Marketing Research*, 56 (2): 197-210.

Hayes, Andrew F. (2013), *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*, New York, NY: Guilford Press.

Hollebeek, Linda D. (2013), "The Customer Engagement/Value Interface: An Exploratory Investigation," *Australasian Marketing Journal,* 21 (1), 17-24.

Hollebeek, Linda D., Mark S. Glynn, and Roderick J. Brodie (2014), "Consumer Brand Engagement in Social Media: Conceptualization, Scale Development, and Validation," *Journal of Interactive Marketing*, 28, 149-165.

Hollebeek, Linda D., Jodie Conduit, and Roderick J. Brodie (2016), "Strategic Drivers, Anticipated and Unanticipated Outcomes of Customer Engagement," *Journal of Marketing Management*, 32 (5-6), 393-398.

Iyer, Aarti, Toni Schmader, and Brian Lickel (2007), "Why Individuals Protest the Perceived Transgressions of Their Country: The Role of Anger, Shame, and Guilt," *Personality and Social Psychology Bulletin*, 33 (4), 572-587.

Jhaver, Shagun, Amy Bruckman, and Eric Gilbert, (2019), "Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit," *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW: 1-27.

Johns, Michael, Toni Schmader, and Brian Lickel (2005), "Ashamed to be an American? The Role of Identification in Predicting Vicarious Shame for Anti-Arab Prejudice After 9-11," *Self and Identity*, 4, 331-348.

Khamitov, Mansur, Yany Grégoire, and Anshu Suri (2020), "A Systematic Review of Brand Transgression, Service Failure Recovery and Product-Harm Crisis: Integration and Guiding Insights," *Journal of the Academy of Marketing Science*, 48 (3), 519-542.

Kumar, V., and Anita Pansari (2016), "Competitive Advantage through Engagement," *Journal of Marketing Research,* 53 (4), 497-514.

Lafreniere, Katherine C., Sarah G. Moore, Robert J. Fisher (2022), "The Power of Profanity: The Meaning and Impact of Swearwords in Word-of-Mouth," *Journal of Marketing Research*, https://doi.org/10.1177/00222437221078606

Lamberton, Cait, and Andrew T. Stephen (2016), "A Thematic Exploration of Digital, Social Media, and Mobile Marketing: Research Evolution from 2000 to 2015 and an Agenda for Future Inquiry," *Journal of Marketing*, 80, 146-172.

Lee, Dokyun, Kartik Hosanagar, and Harikesh S. Nair (2018), "Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook," *Marketing Science*, 64 (11), 5105-5131.

Lewallen, Susan, and Paul Courtright (1998), "Epidemiology in Practice: Case-Control Studies," *Community Eye Health*, 11 (28), 57-58.

Liadeli, Georgia, Francesca Sotgiu, and Peeter W.J. Verlegh (2022), "A Meta-Analysis of the Effects of Brand Owned Social Media on Social Media Engagement and Sales," *Journal of Marketing*, https://doi.org/10.1177/00222429221123250

Lickel, Brian, Toni Schmader, Mathew Curtis, Marchelle Scarnier, and Daniel R. Ames (2005), "Vicarious Shame and Guilt," *Group Processes and Intergroup Relations*, 8 (2), 145-157.

Lickel, Brian, Toni Schmader, and Marija Spanovic (2007), "Group-conscious Emotions," in Tracy, J.L., R.W. Robins, and J.P. Tangney (Eds.), *The Self-conscious Emotions: Theory and Research*, 351-366. New York: The Guilford Press.

Lisjak, Monika, Angela Y. Lee, and Wendi L. Gardner (2012), "When a Threat to the Brand is a Threat to the Self: The Importance of Brand Identification and Implicit Self-Esteem in Predicting Defensiveness," *Personality and Social Psychology Bulletin*, 38 (9), 1120-1132.

Litman, Leib, Jonathan Robinson, and Tzvi Abberbock (2017), "TurkPrime.com: A Versatile Crowdsourcing Data Acquisition Platform for the Behavioral Sciences," *Behavior Research Methods*, 49 (2), 433-442.

Liu, Peggy J., Cait Lamberton, James R. Bettman, and Gavan J. Fitzsimons (2019), "Delicate Snowflakes and Broken Bonds: A Conceptualization of Consumption-Based Offense," *Journal of Consumer Research*, 45 (6), 1164-1193.

Liu, Yi, Pinar Yildirim, and Z. John Zhang (2022), "Implications of Revenue Models and Technology for Content Moderation Strategies," *Marketing Science*, 41 (4), 831-847.

Matherly, Ted (2019), "A Panel for Lemons? Positivity Bias, Reputation Systems and Data Quality on MTurk," *European Journal of Marketing*, 53 (2), 195-223.

Meire, Matthijs, Kelly Hewett, Michel Ballings, V. Kumar, and Dirk Van den Poel (2019), "The Role of Marketer-Generated Content in Customer Engagement Marketing," *Journal of Marketing*, 83 (6), 21-42.

Minot, Joshua R., Michael V. Arnold, Thayer Alshaabi, Christopher M. Danforth, and Peter Sheridan Dodds (2021), "Ratioing the President: An Exploration of Public Engagement with Obama and Trump on Twitter," *PLoS ONE*, 16 (4), e0248880.

Niedenthal, Paula M., June Price Tangney, and Igor Gavanski (1994), "'If Only I Weren't' Versus 'If Only I Hadn't': Distinguishing Shame and Guilt in Counterfactual Thinking," *Journal of Personality and Social Psychology*, 67 (4), 585-595.

de Oliveira Santini, Fernando, Wagner Junior Ladeira, Diego Costa Pinto, Márcia Maurer Herter, Claudio Hoffmann Sampaio, and Barry J. Babin (2020), "Customer Engagement in Social Media: A Framework and Meta-analysis," *Journal of the Academy of Marketing Science*, 48 (6), 1211-1228.

Peng, Jing, Ashish Agarwal, Kartik Hosanagar, and Raghuram Iyengar (2018), "Network Overlap and Content Sharing on Social Media Platforms," *Journal of Marketing Research*, 55 (4), 571-585.

Piff, Paul K., Andres G. Martinez, and Dacher Keltner (2012), "Me Against We: In-group Transgression, Collective Shame, and In-group-directed Hostility," *Cognition and Emotion*, 26 (4), 634-649.

Reed II, Americus, and Mark R. Forehand (2016), "The Ebb and Flow of Consumer Identities: The Role of Memory, Emotions and Threats," *Current Opinion in Psychology*, 10, 94-100.

Reinartz, Werner, Jacquelyn S. Thomas, and Viswanathan Kumar (2005), "Balancing Acquisition and Retention Resources to Maximize Customer Profitability," *Journal of Marketing*, 69 (1), 63-79.

Saboo, Alok, R., V. Kumar, and Girish Ramani (2016), "Evaluating the Impact of Social Media Activities on Human Brand Sales," *International Journal of Research in Marketing*, 33 (3), 524-541.

Schmader, Toni, and Brian Lickel (2006), "The Approach and Avoidance Function of Guilt and Shame Emotions: Comparing Reactions to Self-caused and Other-caused Wrongdoing," *Motivation and Emotion*, 30 (1), 42-55.

Singer, Judith D., and John B. Willett (1993) "It's about Time: Using Discrete-time Survival Analysis to Study Duration and the Timing of Events," *Journal of Educational Statistics*, 18 (2), 155-195.

Sood, Sara, Judd Antin, and Elizabeth Churchill (2012), "Profanity Use in Online Communities" In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1481-1490.

Spiller, Stephen A., Gavan J. Fitzsimons, John G. Lynch Jr., and Gary H. McClelland (2013), "Spotlights, Floodlights, and the Magic Number Zero: Simple Effects Tests in Moderated Regression," *Journal of Marketing Research*, 50 (2), 277-288.

Swaminathan, Vanitha, Karen L. Page, and Zeynep Gürhan-Canli (2007), "'My' Brand or 'Our' Brand: The Effects of Brand Relationship Dimensions and Self-Construal on Brand Evaluations," *Journal of Consumer Research*, 34 (2), 248-259.

Swani, Kunal, George R. Milne, and Elizabeth G. Miller (2021), "Social Media Service Branding: The Use of Corporate Brand Names," *Journal of Business Research*, 125, 785-797.

SurgeAI (2022), "The Obscenity List," (accessed September 19, 2022), https://github.com/surge-ai/profanity

Tajfel, Henri, and John C. Turner (1979), "An Integrative Theory of Group Conflict," in *The Social Psychology of Intergroup Relations*, W.G. Austin and S. Worchel, eds. Monterey, CA: Brooks-Cole, 33-48.

Tangney, June Price (1999), "The Self-Conscious Emotions: Shame, Guilt, Embarrassment and Pride," in *Handbook of Cognition and Emotion*, ed. Tim Dalgleish and Mick J. Power, New York: Wiley, 541-568.

Tangney, June Price, Paula M. Niedenthal, Michelle Vowell Covert, and Deborah Hill Barlow (1998), "Are Shame and Guilt Related to Distinct Self-Discrepancies? A Test of Higgins's (1987) Hypotheses." *Journal of Personality and Social Psychology*, 75 (1), 256-268.

Tracy, Jessica L., and Richard W. Robins (2004), "Putting the Self into Self-conscious Emotions: A Theoretical Model," *Psychological Inquiry*, 15 (2), 103-125.

Wang, Helen Si, Charles H. Noble, Darren W. Dahl, and Sungho Park (2019), "Successfully Communicating a Cocreated Innovation," *Journal of Marketing*, 83 (4): 38-57.

Wang, Yajin, and Deborah Roedder John (2019), "Up, Up, and Away: Upgrading as a Response to Dissimilar Brand Users," *Journal of Marketing Research*, 56 (1), 142-157.

Welten, Stephanie C. M., Marcel Zeelenberg, and Seger M. Breugelmans (2012), "Vicarious Shame," *Cognition and Emotion*, 26 (5), 836-846.

White, Katherine, and Jennifer J. Argo (2009), "Social Identity Threat and Consumer Preferences," *Journal of Consumer Psychology*, 19 (3), 313-325.

Wicker, Frank W., Glen C. Payne, and Randall D. Morgan (1983), "Participant Descriptions of Guilt and Shame," *Motivation and Emotion*, 7 (1), 25-39.

Wilson, Andrew E., Michael D. Giebelhausen, and Michael K. Brady (2017), "Negative Word of Mouth can be a Positive for Consumers Connected to the Brand," *Journal of the Academy of Marketing Science*, 45 (4), 534-547.

van Doorn, Jenny, Katherine N. Lemon, Vikas Mittal, Stephan Nass, Doreén Pick, Peter Pirner, and Peter C. Verhoef (2010), "Customer Engagement Behavior: Theoretical Foundations and Research Directions," *Journal of Service Research*, 13 (3), 253-266.

Zhang, Zhe, and Vanessa M. Patrick (2021), "Mickey D's Has More Street Cred Than McDonald's: Consumer Brand Nickname Use Signals Information Authenticity," *Journal of Marketing*, 85 (5), 58-73.

Web Appendix

For Shame! Socially Unacceptable Brand Mentions on Social Media Motivate Consumer Disengagement

Daniel Villanova (dvillanova@walton.uark.edu)

Ted Matherly (t.matherly@northeastern.edu)

Table of Contents

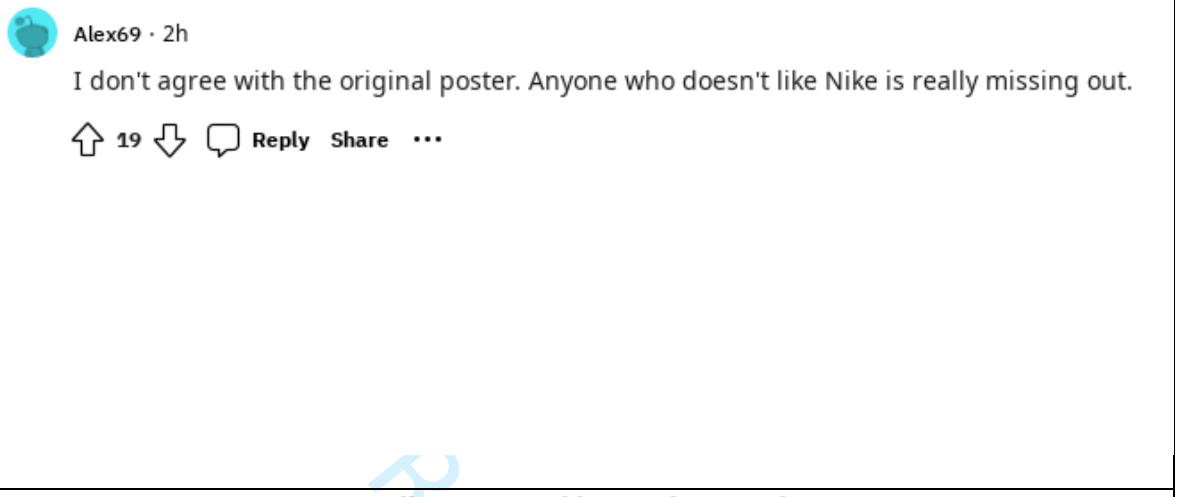
Disclosure: These materials have been supplied by the authors to aid in the understanding of their paper. The AMA is sharing these materials at the request of the authors.

A. Manipulations from Studies 2 and 3

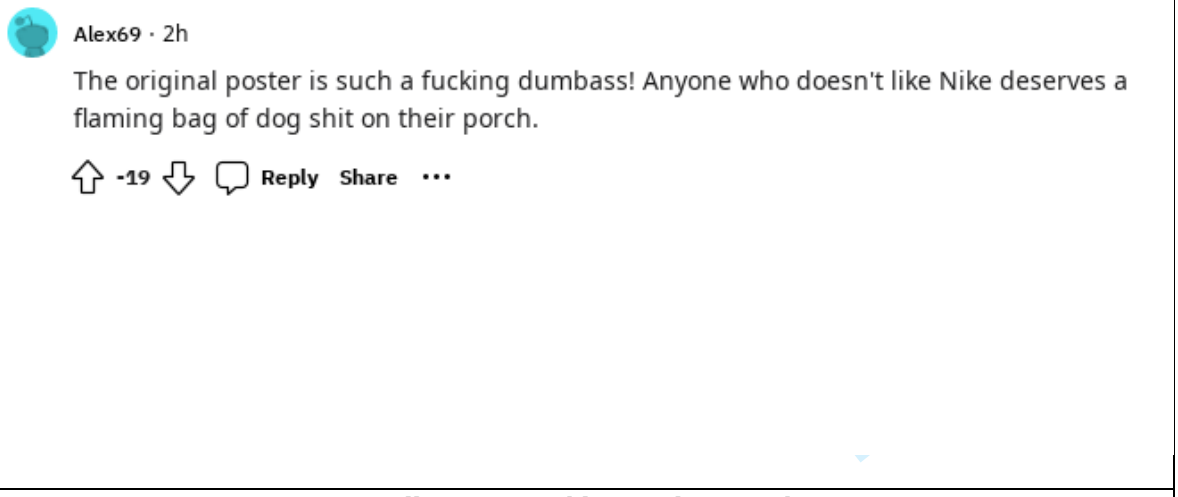
MANIPULATIONS FROM STUDY 2

	
Less Socially Unacceptable Brand User Behavior	More Socially Unacceptable Brand User Behavior

MANIPULATIONS FROM STUDY 3



Less Socially Unacceptable Brand User Behavior



More Socially Unacceptable Brand User Behavior

B. Machine Learning Clustering Analysis

We explored an alternative approach to capturing differences in the users' relationships with brands through an unsupervised machine learning approach. We extracted 24 features from the data, grouped into four meta-categories (see table W4): user account level features (four features), brand specific features (five features), overall text features (four features), and brand mention window features (eleven features). User features were derived from account-level data reported from the Twitter API, while team-specific features were based on counts of the users' tweets that mentioned the brand. For the text of all of the users' contributions, we coded the sentiment of the words in the complete text (positive, neutral, negative, and compound), and applied this analysis to those tweets that mentioned the brand, using the VADER sentiment analysis library (Hutto and Gilbert 2014). Finally, for each of the brand mentions, we extracted windows of five words on either side of the mention and coded the sentiment using VADER), and we also employed the LIWC library (Pennebaker, et al. 2001; Tausczik and Pennebaker 2010) to code for emotional words (affect, positive emotion, negative emotion, swearing, anxiety, and anger). To conduct our analysis, the values of the features were scaled and normalized, and then a k-means clustering algorithm was applied to the matrix to classify users. To simplify the interpretation of these groupings, we set the number of clusters to two, with the presumption that the features would help to separate users into those lower and higher in self-brand connection, and that additional groupings would not have provided a clear mapping on to self-brand connection conceptually. Thus, our machine learning based measure was a binary indicator, with the non-zero value indicating membership in the group presumed to have higher self-brand connections.

References

Hutto, Clayton, and Eric Gilbert (2014), "Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," In *Proceedings of the International AAAI Conference on Web and Social Media*, 8 (1), 216-225.

Pennebaker, James W., Martha E. Francis, and Roger J. Booth (2001), "Linguistic Inquiry and Word Count: LIWC 2001," *Mahway: Lawrence Erlbaum Associates*, 71 (2001).

Tausczik, Yla R., and James W. Pennebaker (2010), "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," *Journal of Language and Social Psychology*, 29 (1), 24-54.

C. Supplementary Tables

Table W1: Profanity dictionary for Study 1

asshole
bitch
cunt
damn
[ ]fag?(s|got)[ ]
fuck
nigger
puss(y|ies)
whore

Dictionary words are written in regular expression form.

Table W2: Summary of data collected for target brand Twitter accounts in Study 1)

Team name	League	Handle	Season start	Season end	Observation end	Maximum followers[1]	Total sampled followers[2]	Active followers[3]	Observed tweets
Boston Red Sox	American	redsox	3/29/18	10/28/18 23:15	10/29/18 2:00	2,102,022	88,096	19,328	1,898,440
Houston Astros	American	astros	3/29/18	10/18/18 23:41	10/19/18 2:06	1,445,937	60,708	13,537	1,365,990
New York Yankees	American	yankees	3/29/18	10/9/18 23:35	10/10/18 2:09	3,393,165	146,175	27,606	2,553,800
Cleveland Indians	American	indians	3/29/18	10/8/18 17:32	10/9/18 2:03	1,037,690	43,093	10,708	1,096,161
Oakland Athletics	American	athletics	3/29/18	10/3/18 23:33	10/4/18 2:03	548,992	23,190	5192	587,372
Los Angeles Dodgers	National	dodgers	3/29/18	10/28/18 23:15	10/29/18 2:03	2,123,537	91,790	18,762	2,048,904
Milwaukee Brewers	National	brewers	3/29/18	10/20/18 23:24	10/21/18 2:00	584,016	26,298	5805	579,301
Atlanta Braves	National	braves	3/29/18	10/8/18 20:12	10/9/18 2:00	1,257,863	53,104	11,657	1,117,841
Colorado Rockies	National	rockies	3/29/18	10/7/18 19:51	10/8/18 2:09	552,525	24,207	5964	639,288
Chicago Cubs	National	cubs	3/29/18	10/3/18 1:03	10/3/18 2:06	2,543,954	104,353	23,477	2,142,454
Total						15,589,701	661,014	142,036	14,029,551

[1] During follower observation period, 10/1/19 - 10/29/18.
[2] Individuals could follow more than one team.
[3] Individuals with at least one tweet during overall observation period, 3/29/18-10/29/18.

Table W3: Summary statistics and correlations for measures in Study 1

	Variable	Mean[1]	SD	(1)	(2)	(3)	(4)	(5)	(6)
1	Disengagement	0.009	0.094						
	Self-brand connection								
2	Regular season team mentions	0.357	4.547	−0.0086					
3	Regular season team mentions indicator	0.047	0.211	−0.0107	0.8449				
4	Machine learning cluster indicator	0.213	0.409	−0.0180	0.3598	0.4258			
	Socially unacceptable brand mentions								
5	Profane tweets mentioning team	14.833	40.512	−0.0324	−0.0099	−0.0113	−0.0059		
6	Ratioed tweets	10.718	43.009	−0.0211	−0.0029	−0.0036	0.0014	0.7947	
	Controls								
7	Total tweets	37,855.607	21,421.937	0.0092	−0.0236	−0.0246	−0.0141	0.4812	0.7087

[1] Untransformed values reported here.

Table W4: Features extracted for clustering analysis in Study 1

Category	Study 1
Account	• Overall tweets • Overall likes • Overall retweets • Overall replies
Brand	• Brand mentions • Tweets with brand mentions • Likes for tweets with brand mentions • Retweets for tweets with brand mentions • Replies for tweets with brand mentions
Text	• Sentiment-Positive • Sentiment-Neutral • Sentiment-Negative • Sentiment-Compound
Brand mentions [1]	• LIWC-Affect • LIWC-Positive emotions • LIWC-Negative emotions • LIWC-Swearing • LIWC-Anxiety • LIWC-Anger • Sentiment-Positive • Sentiment-Neutral • Sentiment-Negative • Sentiment-Compound • Total tokens in window

[1] Around every mention of the brand in the individual's tweets, windows of up to five words on either side were extracted, and these tokens were coded using the LIWC and VADER text analysis programs, which were then averaged across all mentions.

For example, in the tweet:

> Good morning world wow what an awesome time to be alive tomorrow *it's my bday and my* **redsox** *just won the World Series* #DoDamage

the words in italics would be considered within this window.

Table W5: Estimates of Relationship Between Self-Brand Connection and Socially Unacceptable Brand Mentions and Unfollowing from Additional Robustness Checks (Study 1)

	(1) Account-brand Fixed Effects	(2) Exclude Multiple Brand Followers	(3) Account-brand Random Assignment	(4) SurgeAI Dictionary
Self-brand connection Measure:	Team Mentions	Team Mentions	Team Mentions	Team Mentions
Socially unacceptable brand mentions Measure:	Profane Tweets	Profane Tweets	Profane Tweets	SurgeAI Profane Tweets
Constant	0.0351***	0.0397***	0.0318***	0.0446***
	(0.000405)	(0.00911)	(0.000298)	(0.000447)
Socially unacceptable brand mentions	0.00406***	.00335***	0.00336***	0.0153***
	(0.00124)	(0.000106)	(0.000112)	(.000166)
Self-brand connection				-0.0202***
				(0.00453)
Socially unacceptable brand mentions X Self-brand connection	0.532***	0.397***	0.440***	0.241***
	(0.0111)	(0.00911)	(0.00968)	(0.0124)
Total Tweets	-0.340***	-0.297***	-0.304***	-0.427***
	(0.00312)	(0.00280)	(0.00288)	(0.00431)
Marginal effects of Socially unacceptable brand mentions at…				
Lower Self-brand connection	0.00406***	0.00335***	0.00336***	0.0153***
	(0.000124)	(0.000106)	(0.000112)	(0.000166)
Higher Self-brand connection	0.00991***	0.00771***	0.00820***	0.0189***
	(0.000162)	(0.000137)	(0.000144)	(0.000250)
Matchup Indicators	Yes	Yes	Yes	Yes
Time at Risk Indicators	Yes	Yes	Yes	Yes
Individual Fixed Effects	No	Yes	Yes	Yes
Team brand Fixed Effects	No	Yes	Yes	Yes
Individual-team brand Fixed Effects	Yes	No	No	No
Observations	23,803,320	21,366,611	22,211,260	23,804,493
R-squared	0.06748	0.06687	0.06722	0.07056

Notes: Robust standard errors in parentheses (clustered on the individual). Values are reported to three significant figures, except for R-squared values, which require four significant figures to show differences. *** p<0.001, ** p<0.01, * p<0.05. Column 1 reports the results from a model using individual-team brand (i.e., account-brand) fixed effects. In this model, the simple effect of self-brand connection was collinear with the fixed effects and was not estimable. Column 2 reports the results from a model excluding individuals who followed multiple brands. Column 3 reports the results from a model where the individuals who followed multiple brands are randomly assigned to one of the teams they followed. In both models, there is only one brand associated with each individual, and because self-brand connection is time invariant, its simple effect is collinear with the fixed effects and was not estimable. Column 4 reports the results from a model using the SurgeAI dictionary to measure socially unacceptable brand mentions.