Computers in Human Behavior Reports 3 (2021) 100056

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-reports

Tweet valence, volume of abuse, and observers' dark tetrad personality factors influence victim-blaming and the perceived severity of twitter cyberabuse

Christopher J. Hand [a,*], Graham G. Scott [b], Zara P. Brodie [b], Xilei Ye [c], Sara C. Sereno [c,d]

[a] Department of Psychology, Glasgow Caledonian University, Glasgow, UK
[b] Applied Psychology Research Group, School of Education and Social Sciences, University of the West of Scotland, Paisley, UK
[c] School of Psychology, University of Glasgow, Glasgow, UK
[d] Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK

ARTICLE INFO

Keywords:
Cyberbullying
Dark triad
Sadism
Twitter
Victim blame
Warranting theory

ABSTRACT

Previous research into Twitter cyberabuse has yielded several findings: victim-blaming (VB) was influenced by victims' initial tweet-valence; perceived severity (PS) was influenced independently by tweet valence and abuse volume; VB and PS were predicted by observer narcissism and psychopathy. However, this previous research was limited by its narrow focus on celebrity victims, and lack of consideration of observer sadism. The current study investigated 125 observers' VB and PS perceptions of lay-user cyberabuse, and influence of observers' Dark Tetrad scores (psychopathy, narcissism, Machiavellianism, sadism). We manipulated initial-tweet valence (negative, neutral, positive) and received abuse volume (low, high). Our results indicated that VB was highest following negative initial tweets; VB was higher following high-volume abuse. PS did not differ across initial-tweet valences; PS was greater following a high abuse volume. Regression analyses revealed that observer sadism predicted VB across initial-tweet valences; psychopathy predicted PS when initial tweets were 'emotive' (negative, positive), whereas Machiavellianism predicted PS when they were neutral. Our results show that perceptions of lay-user abuse are influenced interactively by victim-generated content and received abuse volume. Our current results contrast with perceptions of celebrity-abuse, which is mostly determined by victim-generated content. Findings are contextualised within the Warranting Theory of impression formation.

1. Introduction

Online abuse is an increasing problem on social media and can have serious negative impact on victims (Allcott & Gentzkow, 2017; John et al., 2018). Such abuse can take the form of private direct messages, but also involves abusive posts in public forums. Nevertheless, observers often attribute blame to victims for the abuse perpetrated against them (Scott et al., 2019; Weber et al., 2013). To better understand the blame and lack of sympathy directed at online abuse victims, we manipulated the valence of the initial tweet and the volume of abuse received in artificially-constructed Twitter interactions, measuring participants' victim-blame (VB) and perceived incident severity (PS). Additionally, we explored the role played by the *Dark Tetrad* of observer personality factors (psychopathy, narcissism, Machiavellianism, and sadism; Jones & Paulhus, 2013) on these perceptions.

1.1. Online abuse

Social media is growing in importance, especially among the younger population, and in recent years has also become increasingly diverse (Allcott & Gentzkow, 2017; Mohsin, 2020; Villlanti et al., 2017). Facebook is the largest social media platform with 2.23 billion active monthly users. Other popular sites include the primarily photo-based site Instagram with 1 billion active monthly users, Twitter, which allow users to broadcast 'tweets' using a limited number of characters, with 335 million active monthly users, and Snapchat, which allows the sharing of temporary text and picture messages which disappear after viewing, with 291 million active monthly users (Chaffey, 2019). Social media is used to satisfy users' need to belong by sustaining friendships and relationships, and to organise and document activities, and to allow users to manage their self-presentation (Garcia & Sikström, 2014; Nadkarni & Hofmann,

* Corresponding author. Department of Psychology, George Moore Building, Glasgow Caledonian University, Cowcaddens Road, Glasgow, G4 0BA, UK.
E-mail address: Christopher.Hand@gcu.ac.uk (C.J. Hand).

https://doi.org/10.1016/j.chbr.2021.100056
Received 7 October 2020; Received in revised form 4 January 2021; Accepted 6 January 2021
Available online 25 January 2021

2012; Tosun, 2012). Its use is increasingly becoming embedded into the daily life of users (Chaffey, 2019).

With elevated usage of online social networks, cases of online abuse are also increasing (Hearn & Hall, 2019; Mendez, Jorquera, Ruiz-Esteban, Martinez-Ramon, & Fernandez-Sogorb, 2019; Vakhitova et al., 2019). Online abuse manifests in different ways, with distinct patterns and combinations of abusive behaviours often classified as cyberbullying, cyberaggression, cyberharassment, and cyberstalking (Maple et al., 2012; Menesini & Nocentini, 2009). Specific categories of abuse often overlap and are difficult to define (Jurgens et al., 2019; Menesini et al., 2012). Abuse sometimes includes private communications between individuals (i.e., private or direct messaging) but it also commonly takes place in the public domain (e.g., as posts or comments on a Facebook timeline or tweets or comments on a Twitter page). The Warranting Theory of impression formation (Walther & Parks, 2002) states that we form impressions of others online based on claims individuals make about themselves (*identity claims*) and by evidence left unintentionally (*behavioural residue*), and that the latter carry more weight. In the context of online communication, particularly on social media, messages and statements from third parties, including online abuse, constitute behavioural residue (Scott et al., 2019). The Hyperpersonal Theory of communication (Walther, 1996, 1997) suggests that, in contrast to offline contexts, impressions formed in online domains will be exaggerated and stereotyped based on the limited information available. Taken together, these suggest that publically visible abuse on social networks will significantly impact the impressions formed of online abuse victims by observers.

The impact on victims of online abuse can be extremely serious and damaging (e.g., John et al., 2018). Potential negative outcomes for victims include depression, anxiety, self-harm, loneliness, enforced changes to personal and work lifestyle, and even suicide (Gini & Espelage, 2014; Mechanic et al., 2000; Short & McMurray, 2009; van Geel et al., 2014). Because of the nature of the online domain, abuse suffered there may be more damaging, and longer lasting, than offline abuse. Because of the importance of social media in different aspects of daily life, and the constant accessibility enabled by mobile technologies, it can be difficult for victims to find a 'safe space' away from online abuse without also cutting themselves off from the benefits afforded by such technologies. Additionally, the permanence of public posts means that, unlike offline abuse, online abuse can linger and continue to impact the victim long past the time of the initial incident (e.g., Aoyama et al., 2011).

Despite these negative effects, those who suffer are typically given little support or sympathy from either friends or authorities (e.g., Chen et al., 2015; Dredge et al., 2014; Gahagan et al., 2016). This lack of regard may be due to observers attributing blame to victims for abuse (e.g., Russell & Hand, 2017; Scott et al., 2019; Scott et al., 2020). There are two theoretical explanations for why this occurs: Just World Theory (Lerner & Simmons, 1966) and Defensive Attribution (Shaver, 1970). Just World Theory states that people '*get what they deserve*', as the world is a just place. The Defensive Attribution hypothesis states that observers often blame the cause of an unpleasant event on the disposition of the individual involved in order to increase their own sense of control.

The attribution of blame to victims, and thus the lack of support offered, can be influenced both by aspects of the online interaction that is presented, and by the personality of the observer (e.g., Scott et al., 2020). The warrants associated with public online abuse which shape victim blame will be examined, and how individuals' Dark Tetrad personality factors impact their processing of abusive acts will be assessed.

1.2. Online victim blame

Scott et al. (2019) demonstrated that Facebook users' perception of cyberbullying victims (both VB and perceptions of victim attractiveness) was influenced by the volume of abuse directed towards victims, and whether the abuse was generated by a single source or multiple ones. Timeline owners who received a lower volume of abuse from a single source were perceived as more blameworthy. This may have been due to desensitization. However, on Facebook the majority of friends are not online-only acquaintances, but individuals with whom the profile owner has established offline relationships, and so in this context abuse was interpreted by observers as friendly 'banter' (Scott et al., 2019). As both volume and source of abuse constitute aspects of third party-generated content, they can be classified as behavioural residue warrants. Weber et al. (2013) demonstrated that identity claims can also affect attributed victim blame. They manipulated the amount of content generated by social media users (identity claims) and found that victims who generated increased content were at greater risk of being blamed.

The only study to have manipulated both identity claims and behavioural residue experimentally focused on celebrity users of Twitter. Scott et al. (2020) manipulated the valence of original tweets of celebrity Twitter users (identity claims), and the volume of abusive replies by non-celebrities (behavioural residue) in manufactured Twitter exchanges. Celebrities were attributed more blame following negative tweets, and when a negative tweet was followed by a high volume of abuse it was perceived as being least severe, demonstrating that identity claims influenced blame perceptions, while both categories of warrant were indicators of severity. However, given that the vast majority of social media users (not to mention the majority of human beings) are lay-users rather than celebrities, the work of Scott et al. (2020) has limited generalisability.

It is likely that lay-users (non-celebrities) on Twitter may be perceived slightly differently from celebrity Twitter users and lay Facebook users. Whereas celebrities often use Twitter and other social networks for self-publicity (e.g., Gayle & Lawson, 2013; Lee & Lim, 2016; Lim, 2017), the majority of users are lay-users. Non-celebrities have a wider variety of less overtly self-promoting motivations for use of these sites (Hargittai & Litt, 2011; Yoo et al., 2012). The relationships among and interactions between users on Twitter also differ from those on Facebook. Facebook users' posts typically can only be seen and commented on by their Facebook friends (individuals who have mutually agreed to be friends on the site). Twitter profiles are typically public, and can be viewed and responded to by any other user (unless that user has been actively blocked). Thus, interactions on Twitter are more likely to occur between, and be viewed by, parties who have no pre-existing offline relationship.

Although observers may be cognisant of varying motivations driving distinct categories of social media users, they will still form impressions from the available online warrants. As behavioural residue usually carries more weight than identity claims, and negative warrants carry more weight than positive ones (Walther et al., 2009), negative online abuse will likely contribute negatively to any impression formed. However, as identity claims have been shown to drive conceptions of certain characteristics and attributions in online settings (Scott et al., 2020; Scott & Ravenscroft, 2017), it is likely that tweets will also contribute to impressions formed and blame attributed. On Twitter individuals who interact are not guaranteed to be acquainted offline, or to be considered legitimate friends as they would on Facebook (Phua et al., 2017); thus, it is also possible that abuse in the current study will not be perceived as jovial or "*banterous*", as may have been the case in previous studies (Scott et al., 2019).

Another aspect of perceived abuse which has received limited attention is the role played by individual differences between observers. While all viewers will base the impressions they form on the identity claims and behavioural residue of a target's social media profile, how these are interpreted may differ between viewers.

1.3. Dark tetrad personality factors

Although the population in general underestimates the severity of online abuse and its impact on victims, individuals differ in terms of how abusive incidents are interpreted. Specifically, individuals scoring high in the *Dark Triad* of personality traits – psychopathy, narcissism, Machiavellianism (Jones & Paulhus, 2013) – may be likely to underplay the

severity of online abuse and to attribute more blame to victims. In addition to the three traits that encompass the Dark Triad, there has been a recent theoretical shift towards the inclusion of subclinical Sadism as a distinct but interrelated construct (Johnson et al., 2019), leading to increased consideration of what is now known as a *Dark Tetrad* (DT) of personality. Machiavellianism is reflected by a manipulative and deceptive nature, a lack of concern with conventional morality, and a lack of interpersonal affect (Deluga, 2001). Narcissism, while primarily reflected by high levels of vanity and self-enhancement tendencies not commonly associated with Machiavellianism (Paulhus & Williams, 2002), is similarly characterised by an exploitative interpersonal style, a sense of superiority and entitlement, and selfishness (Millon & Davis, 1996). Psychopathy reflects several aversive interpersonal (e.g., callousness, remorselessness) and behavioural (e.g., anti-social behaviour, impulsivity) characteristics (Douglas et al., 2012). While there is a level of conceptual overlap between the original dark triad traits and subclinical sadism, such as empathy deficits and callous behaviour (Međedović & Petrović, 2015), a body of literature has established the incremental validity of this trait particularly in the context of externalising behaviours such as cyber-aggression and trolling (Buckels et al., 2014). Subclinical sadism, or 'everyday sadism', is characterised by deriving pleasure from witnessing the distress or pain experienced by others, as well as diminished disgust sensitivity and a predatory interpersonal style (Meere & Egan, 2017).

Recent research has identified that those high in psychopathy, Machiavellianism, and sadism are more likely to engage in trolling behaviours (Buckels et al., 2014), cyberaggression and cyberbullying (Brown et al., 2019; Pabian et al., 2015), and are more inclined to use profane and aggressive language online (Sumner et al., 2012). Some findings indicate that, while all four traits predict online disinhibition, psychopathy and Sadism alone are independent predictors of cyberaggression (Kurek et al., 2019), while other research argues that psychopathy is a stand-alone, independent predictor of cyberbullying behaviour (Gibb & Devereux, 2014; Goodboy & Martin, 2015) and Facebook trolling (Craker & March 2016). However, van Geel and colleagues found that while psychopathy, Machiavellianism, and sadism were related to traditional bullying, Sadism alone was a significant predictor of cyberbullying (van Geel et al., 2017), suggesting some inconsistencies in the current literature. Further, while these studies have explored dark tetrad (DT) personality predictors of online abuse *perpetration*, only one study, to the authors' knowledge, has considered their relation to factors relevant to outsider *observation* of abuse. Scott et al. (2020) indicated that psychopathy and narcissism were predictive of reduced perceived severity of abuse received by celebrities online, while narcissism predicted increased victim blame. While providing initial insight into the links between dark traits and reactions to online trolling, as mentioned previously, this study focused solely on celebrities and is thus limited in its generalisability. Furthermore, this study failed to account for subclinical sadism – the trait that is arguably the most conceptually-relevant when examining responses to the victimization of others (Buckels et al., 2014).

1.4. The current study

As stated previously, the work of Scott et al. (2020) was limited to the perception of celebrity Twitter abuse, and additionally did not account for observer subclinical sadism. The current study addressed these gaps by presenting carefully controlled stimuli which demonstrated cyberabuse of Twitter lay-users (unknown members of the public), accounting for observer subclinical sadism (in addition to psychoticism, narcissism, and Machiavellianism). We present a novel investigation how different types of tweets by lay-users, as well as observers' DT personality scores, influence attributed VB and perceived abuse severity (PS). We investigated how the Valence of tweets written by 'victims' (identity claims: negative, neutral, or positive) and the Volume of abusive responses by followers (behavioural residue: low or high) affected participants' attribution of VB, and perceptions of incident severity (PS). We also examined whether participants' DT personality traits influenced their victim-blaming and severity perceptions. The findings of the current study provide informative insights into observer perceptions of the victims of online abuse and illuminate the complex relationship between 'victim' behaviour, abuser behaviour, and observers' internalised characteristics.

Based on the theoretical frameworks and arguments provided by Warranting Theory (Walther & Parks, 2002), Just World Theory (Lerner & Simmons, 1966) and Defensive Attribution (Shaver, 1970) – as well as the empirical findings of Scott et al. (2020) – we predicted that:

H1a. Negative initial tweets will result in greater VB.

H1b. Negative initial tweets will result in lower perceptions of abuse severity.

H2. Increased volume of abuse (i.e., number of abusive responses) would be associated with higher perceived severity of abuse.

H3a. Attributed VB will be higher among participants scoring higher in DT personality factors.

H3b. Perceived severity will be lower among participants scoring higher in DT personality factors.

2. Method

2.1. Design and participants

The current study employed a 3 (Initial Tweet Valence: negative, neutral, positive) × 2 (Abuse Volume: low, high) within-participants design. Following the presentation of each tweet and associated replies, we measured participants' judgements of VB and PS. After presentation of all tweets and replies, we measured participants' Dark Tetrad personality traits (Machiavellianism, narcissism, psychopathy, and sadism). The study was carried out online and hosted by a UK university. Participants were 125 volunteers who indicated on a questionnaire that they used Twitter regularly (84 female, 39 male, 2 non-binary; M_{age} = 25.06 years, SD_{age} = 4.94; range 18–53 years; median = 24 years; mode = 24 years). All participants were native or highly proficient readers/speakers of English. Participants represented a diverse set of nations – 32 from Great Britain and Northern Ireland, 28 from China, 20 from continental Europe, 13 from North America, 12 from South East Asia, 5 from India, 5 from Australia, 4 from African nations, 2 from Middle Eastern countries, and 1 from South America. The remaining three participants did not disclose their national identity. An a priori power analysis aligned against typical parameters (e.g., Cohen, 1988) was conducted using G*Power 3.1.9.4. This analysis estimated that a sample of 84 would yield power of .95 given a repeated-measures design and an anticipated effect size of $f = 0.20$ with α=.05 (Cohen, 1988); our final sample exceeded this estimate. Participants were recruited through convenience/opportunity sampling. Online recruitment took place via advertisements on the researchers' social media networks; on-campus advertisements (physical posters) were used to publicise the study and the link to the online survey.

2.2. Materials

Participants were presented with a series of screenshots of six Twitter pages featuring an initial tweet by a fictitious lay-person (i.e., a non-celebrity male profile owner). A male-only victim-set was chosen for two main reasons: first, so that results could be directly compared to previous work by Scott et al. (2020) who used the same manipulation, and second, because twitter is used predominantly by males (Statistia, 2018a, 2018b). Further, there is some evidence which suggests that males are more likely to be abused on this platform (Demos, 2014). The profile pictures of the six 'victims' were prototype-based male faces

assessed for masculinity preference (see, e.g., Penton-Voak et al., 1999; Tiddeman et al., 2001; Welling et al., 2007). The 'names' of the artificial lay-person 'victims' were generated using a list of popular names in the United Kingdom (Embury-Dennis, 2016). Artificial account-owner names were: *Noah Anderson, Charles Hughes, Oliver Jones, Jacob Smith, James Williams,* and *George Wilson*.

Participants were presented with six target stimuli which were composed of an initial tweet by a male profile owner, followed by six comments from Twitter users who were unknown to the participants. Stimuli were manufactured using GIMP (https://www.gimp.org/). Each stimulus contained the following, in order from top to bottom: the 'victim' name and profile picture; the tweet itself; the number of 'retweets' and 'favourites' (the numbers for each of these were counterbalanced); and the six comments. The initial tweet was either Negative, Neutral, or Positive; of the six replies, either two (Low volume) or four (High volume) were abusive, with the rest being neutral. An example stimulus is presented in Appendix A.

The content of initial tweets and associated comments were identical to those used by Scott et al. (2020), and full details of stimulus norming procedures are provided within their article, and summarised here. Scott et al.'s naïve norming participants ($N = 28$) were presented with a list of 90 tweets and comments (taken from Twitter) and rated each on 7-point Likert-type scales of valence (*1 = negative – 7 = positive*), arousal (*1 = not arousing – 7 = very arousing*) and politeness (*1 = abusive – 7 = polite*). The mean valence, arousal, and politeness ratings for the negative, neutral, and positive tweets and comments obtained by Scott et al. and selected for the current study are presented in Table 1 (please note that these are identical to those in Scott et al., 2020).

Negative content was low in both valence and politeness, but high in arousal; neutral content was neither high nor low in either valence or arousal, but high in politeness; positive content was high in all three dimensions. Examples of tweets used in the current study included: Positive – "*Be disciplined about doin' the little things for your goals – daily. Consistency adds up to success. #ChaseYourGreatness*"; Neutral – "*Weathers getting chilly. I think summer is over*"; Negative – "*Isn't it annoying that the really illiterate & rude people on Twitter are so fucking stupid that they forgot to kill themselves today.*". For a complete list of the Negative and Neutral comments, please see Scott et al. (2020) Appendix B.

2.3. Measures

Measures of VB and PS were identical to those used by Scott et al. (2020), which in turn were derived from the study by Weber et al. (2013). VB and PS were measured on four- and two-item scales respectively, using 5-point Likert-type scales [VB: Cronbach's $\alpha = 0.90$, mean inter-item correlation = .692, $F(3,749) = 22.351$, $p < .001$; PS: Cronbach's $\alpha = 0.65$, inter-item correlation = .479, $F(1,749) = 148.523$, $p < .001$]. An example item from the VB measure was: "*Did the victim provoke the abuse?*" (*1 = strongly disagree – 5 = strongly agree*) and from the PS measure was: "*How severe was the abuse?*" (*1 = not severe at all – 7 = very severe*). VB and PS measures were based on participants' mean item responses. Participant responses to these six items (4 VB and 2 PS) were examined via Principal Components Analysis (Direct Oblimin rotation) with an eigenvalue threshold of 1. Assumptions were met (KMO = .783; Bartlett's Test < 0.001). As anticipated, two components were returned, explaining approximately 77% of the variance in item scores. Investigation of the component matrix revealed that Component 1 reflected the 4 DVB items (component scores between 0.782 and 0.928) and Component 2 reflected the 2 PS items (component scores .848 and .856). There were no issues with cross-loading (max. cross-load score of 0.253).

Table 1
Mean Ratings (plus standard deviations) of Valence, Arousal, and Politeness for Tweets (Negative, Neutral, Positive) and Comments (Negative, Neutral) from Scott et al. (2020).

		Mean Rating (Standard Deviation)		
Stimulus	Valence	Valence	Arousal	Politeness
Tweet	Negative	1.46 (0.10)	5.52 (0.03)	1.38 (0.03)
	Neutral	4.23 (0.58)	3.41 (0.18)	5.05 (0.43)
	Positive	5.88 (0.28)	4.96 (0.05)	5.82 (0.51)
Comment	Negative	1.60 (0.36)	5.19 (0.69)	1.47 (0.36)
	Neutral	4.06 (0.28)	3.60 (0.38)	4.50 (0.42)

Note: Participant judgments were measured on 7-point scales with endpoints 1 and 7 labelled, respectively, as follows: Valence (*very negative – very positive*); Arousal (*not arousing – very arousing*); and Politeness (*abusive – polite*).

Dark Tetrad personality factors were measured using a 36-item questionnaire, with each item having a five-point Likert-type response scale (*1 = Strongly Disagree – 5 = Strongly Agree*; 27 dark triad items of the SD3, Jones & Paulhus, 2013; 9 sadism items from Johnson et al., 2019). Three items of Psychopathy, two items of Narcissism, and one item of Sadism were reverse-scored. Example statements from each of the Dark Tetrad dimensions are: Psychopathy – "*Payback needs to be quick and nasty.*"; Narcissism – "*People see me as a natural leader.*"; Machiavellianism – "*You should wait for the right time to get back at people.*"; Sadism – "*Being mean to others can be exciting.*". Cronbach's alphas ($n_{items} = 9$) for Psychopathy, Narcissism, Machiavellianism, and Sadism were 0.714, 0.730, 0.700, and 0.855, respectively; all mean inter-item correlations were >0.195 [all $Fs > 6.594$, all $ps < .001$]. Each Dark tetrad dimensional score was based on participants' mean item responses. Participant responses to the 27 items measuring psychoticism, narcissism, and Machiavellianism (Jones & Paulhus, 2013) were examined via Principal Components Analysis (Direct Oblimin rotation) with an eigenvalue threshold of 1. Assumptions were met (KMO = .734; Bartlett's Test < 0.001). As anticipated, three components were returned, explaining approximately 39% of the variance in item scores. Investigation of the component matrix revealed that Component 1 reflected the psychoticism items (component scores between 0.406 and 710), Component 2 reflected the narcissism items (component scores between 0.340 and 654), and Component 3 reflected the Machiavellianism items (component scores between 0.309 and 0.600). There were no issues with cross-loading (max. cross-load score of 0.294; no cross-loading scores were within 0.100 of their dominant loading). The nine sadism items (Johnson et al., 2019) were examined in a similar manner. Assumptions were met (KMO = .858; Bartlett's Test < 0.001). As anticipated, one component was returned, explaining approximately 49% of the variance in item scores (component scores between 0.384 and 0.837).

2.4. Procedure

The study was designed in accordance with British Psychological Society (2014) principles; participants were ensured that their responses would be anonymous, that they could withdraw from the study at any time without reason and without penalty, that their data would be stored securely, etc. Ethical approval was granted by the host university's Ethics Committee. Participants were tested online using the QuestionPro platform (https://www.questionpro.com/). After providing informed consent, participants were given access to one of six questionnaires which presented the profiles in one of six pseudo-random orders. After reading task instructions, participants completed a short demographic questionnaire. For each tweet (and associated replies), participants were asked to form an impression of the tweeter and could view each target stimulus for as long as they wanted. After processing each target stimulus, participants made VB and PS judgements via the scales described in section 2.2. After responding to all stimuli, participants then completed the Dark Tetrad items – as described in section 2.2 – before receiving full debriefing information. Survey participation lasted approximately 20 min.

2.5. Data analysis

Participant data ($N = 125$) was first checked for completeness, and it was found that there were no missing values across profile ratings or Dark

Tetrad measures. Three distinct sets of inferential analyses were conducted. Two 3 (Initial Tweet Valence: negative, neutral, positive) × 2 (Abuse Volume: low, high) repeated measures analyses of variance (ANOVAs) were performed on ratings of VB and PS; these analyses tested hypotheses H_{1a} (predicted effect of initial tweet valence on VB), H_{1b} (predicted effect of initial tweet valence on PS), and H_2 (predicted effect of abuse volume on PS). In order to test H_{3a} (predicted positive relationship between participants' attributed VB and DT scores) and H_{3b} (predicted positive relationship between participants' PS and DT scores), Pearson's correlations were used to investigate the relationships between participant-observers' Dark Tetrad ratings and perceptions of blame and severity (NB, these were conducted on global ratings of VB and PS collapsed across conditions, and performed individually across each level of Initial Tweet Valence). Finally, a series of stepwise multiple regressions investigated the predictive value of participant-observers' Dark Tetrad attributes on ratings of VB and PS (as with the correlational analyses, there were performed globally and independently across Initial Tweet Valence conditions).

3. Results

3.1. ANOVAs

Mean ratings (and associated 95% confidence intervals) of VB and PS across levels of Initial Tweet Valence and Abuse Volume are presented in Table 2.

Ratings of VB and PS were first tested for normality, and both Kolmogorov-Smirnov (both $ps > .200$) and Shapiro-Wilk tests (both $ps > .05$) indicated that the VB and PS data could be assumed to approximate normal distributions.

Analysis of VB ratings revealed a significant, large main effect of Valence [$F(2,248) = 114.335$, $p < .001$, $\eta_p^2 = .480$]. Planned follow-up comparisons (Bonferroni) illustrated that attributed VB was significantly higher following lay-persons' negative initial tweets (3.45) than either their neutral tweets (2.51; $p < .001$) or positive tweets (2.43; $p < .001$). There was no significant difference between the attributed VB following neutral (2.51) versus positive tweets (2.43; $p = .153$). The main effect of Abuse Volume on VB was significant, but smaller than that of Valence [$F(1,124) = 20.684$, $p < .001$, $\eta_p^2 = .143$]. VB was greater when Abuse Volume was high (2.89) versus low (2.71). Finally, the ANOVA indicated no evidence of a significant interaction between Tweet Valence and Abuse Volume on attributed VB [$F < 1$; see Fig. 1].

A second ANOVA considered PS ratings. There was no evidence of a main effect of Valence [$F < 1$]. The main effect of Abuse Volume on PS was highly significant and large [$F(1,124) = 91.493$, $p < .001$, $\eta_p^2 = .425$]. Observers' judgements of PS were greater when Abuse Volume was high (3.51) versus low (2.90). Finally, the ANOVA yielded a significant interaction between Valence and Volume on PS [$F(2,248) = 3.276$, $p = .039$, $\eta_p^2 = .026$; see Fig. 2].

Analyses revealed that the simpe main effect of Volume was significant across all tweet valences (all $ps < .001$). When initial lay-person tweets were negative, PS was greater following high volumes of abuse (3.52) than lower volumes (2.78). When initial lay-person tweets were neutral, PS was again greater following higher volumes of abuse (3.44) than lower volumes (3.00). When initial tweets were positive, PS was greater when Abuse Volume was high (3.57) vs. low (2.92). Although all of these comparisons are significant, it is clear that the interaction is being driven by larger differences between high and low volumes of abuse following negative tweets (0.74) and positive tweets (0.64) than following neutral tweets (0.44). When Abuse Volume was high, there was no difference in PS between negative initial-tweets and neutral tweets ($p > .99$), nor between negative and positive tweets ($p > .99$); there was no difference between PS between neutral and positive tweets ($p = .580$). When Abuse Volume was low, there was no difference in PS between negative initial-tweets and neutral tweets ($p = .177$), nor between negative and positive tweets ($p = .597$); there was no difference between PS between neutral and positive tweets when Abuse Volume was high ($p > .99$).

Table 2
Mean Ratings (plus standard deviations) of Victim Blame (VB) and Perceived Severity (PS) with 95% CIs across Experimental Conditions.

Tweet Valence	Abuse Volume	VB (SD)	VB 95% CI	PS (SD)	PS 95%CI
Negative	Low	3.35 (0.69)	[3.23–3.47]	2.78 (1.07)	[2.60–2.97]
	High	3.54 (0.66)	[3.43–3.66]	3.52 (0.88)	[3.37–3.68]
Neutral	Low	2.41 (0.64)	[2.30–2.52]	3.00 (1.08)	[2.81–3.20]
	High	2.62 (0.71)	[2.49–2.74]	3.44 (1.12)	[3.25–3.64]
Positive	Low	2.36 (0.75)	[2.23–2.49]	2.92 (1.01)	[2.75–3.10]
	High	2.50 (0.74)	[2.37–2.63]	3.57 (1.13)	[3.37–3.77]

Note: Participant judgments were measured on 5-point scales with endpoints 1 (least blame/severity) and 5 (greatest blame/severity).

3.2. Correlations

Pearson's correlations (one-tailed) were conducted to identify relationships between Dark Tetrad traits, VB, and PS both globally and independently across Initial Tweet Valence conditions. Furthermore, these analyses were crucial in identifying potential relationships for further regression analyses (see section 3.3). Pearson's correlation coefficients (r) and significance values are presented in Tables 3 and 4. Several significant relationships were identified, with strengths ranging from small to large, based on Cohen's (1988) standards: small, r=.1; medium, r=.3; large, r=.5.

Considering the global total ratings of VB, collapsed across Initial Tweet Valence and Abuse Volume, significant (but small) positive relationships were observed with Total VB and all four Dark Tetrad dimensions (see Table 3). Psychopathy ratings explained 4.45% of the variance in Total VB, narcissism 2.69%, Machiavellianism 6.35%, and sadism 7.02%.

In the negative Initial Tweet condition, small-to-medium significant negative correlations were observed between VB and psychopathy, narcissism, and sadism scores (see Table 4; all $ps < .01$). There was no significant relationship between Machiavellianism and VB in the negative Initial Tweet Valence condition. When considering PS of abuse received following an initial negative tweet, significant but small negative correlations were found between perceptions of abuse severity and psychopathy, Machiavellianism, and sadism scores (see Table 4; all $ps < .05$). There was no significant relationship between narcissism and PS in the negative Initial Tweet Valence condition.

When initial tweets were neutral, significant small-to-medium strength positive correlations were observed between VB and all four Dark Tetrad dimension scores (see Table 4; all $ps < .01$). In this condition, significant small-to-medium negative correlations were found between PS and psychopathy, Machiavellianism, and sadism scores (see Table 4; all $ps < .05$). There was no significant relationship between narcissism and PS in the neutral Initial Tweet Valence condition.

Finally, we considered positive initial tweets. Significant medium-strength positive correlations were observed between VB and all four Dark Tetrad dimension scores (see Table 4; all $ps < .001$). In terms of ratings of PS following a positive initial tweet, the only significant correlation observed was between PS and psychopathy (see Table 4).

3.3. Regressions

As a result of the correlational analyses in section 3.2., and following the protocol of Scott et al. (2020), correlations between VB, PS and Dark Tetrad variables at $p < .10$ at the univariate level were considered as candidates for multivariable models, as typical significance limits (e.g., $p \leq .05$) frequently fail to establish significance in variables known to be predictive (Bursac et al., 2008). Initial data screening indicated that assumptions related to multicollinearity, independence of error terms,

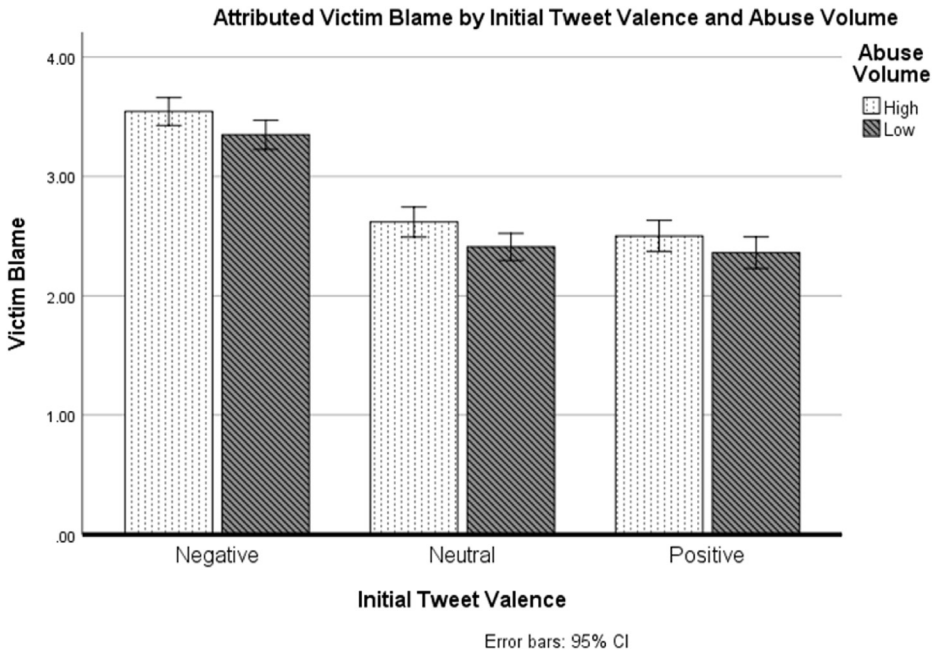


Fig. 1. Attributed victim blame of lay-persons by initial tweet valence and abuse volume.

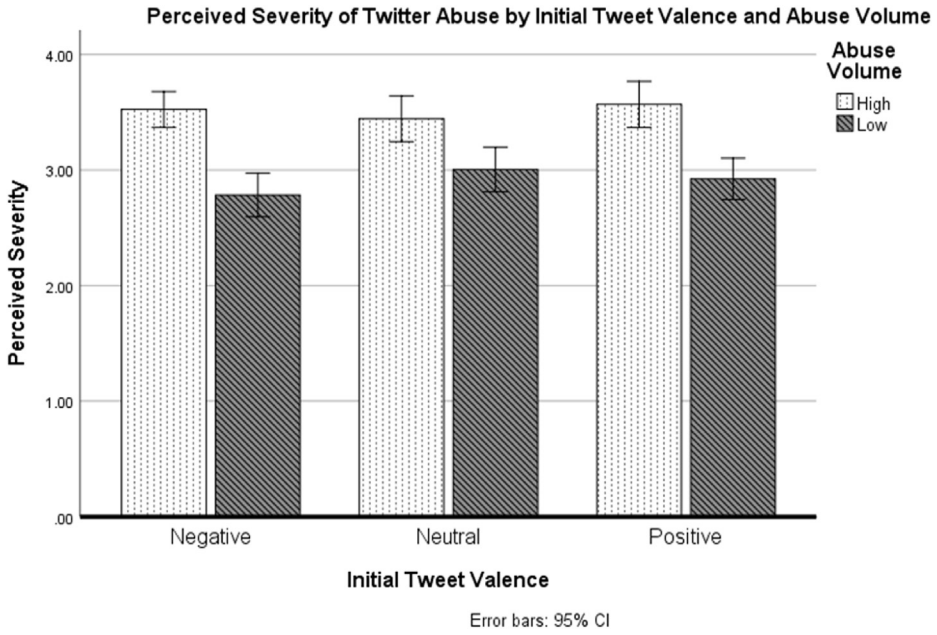


Fig. 2. Perceived severity of twitter abuse by initial tweet valence and abuse volume.

non-zero variances, normality, homoscedasticity and linearity were not violated.

A series of stepwise regressions were conducted to identify the predictive value of the four Dark Tetrad (DT) traits for VB and PS across all three initial tweet conditions (negative, neutral, and positive). These results are presented in Table 5.

First, three models were constructed to determine whether the four DT traits predicted VB across tweet conditions. These models indicate that sadism was a significant independent predictor of VB in the negative, neutral, and positive tweet conditions (small-to-medium effects), explaining 12%, 9%, and 24% of variance, respectively (see Table 5). Psychopathy, narcissism, and Machiavellianism were not significant predictors of VB in any tweet condition. This suggests that as sadism increases, VB following negative, neutral, and positive tweets also increases (see Table 5).

Three further models were conducted to determine whether the four DT traits predicted PS of abuse in each tweet condition (see Table 5). In the negative and positive tweet conditions, psychopathy was found to be the only significant independent predictor of PS (small-to-medium effect), explaining 9% and 4% of variance, respectively. Narcissism, Machiavellianism, and Sadism were not significant predictors in these two initial tweet conditions. This suggests that, as psychopathy increases, PS of abuse following positive and negative tweets decreases (see Table 5). In contrast, in the neutral tweet condition, Machiavellianism was the only significant predictor (small-to-medium effect), explaining 5% of variance in PS of abuse, while psychopathy, narcissism, and sadism

Table 3
Pearson's Correlations of Dark Tetrad Components VB, and PS Ratings collapsed across Valence and Volume.

	1	2	3	4	5	6
1. Psychopathy	1	.457***	.367***	.702***	.211**	-.276**
2. Narcissism		1	.420***	.392***	.164*	-.108 .115
3. Machiavellianism			1	.381***	.252**	-.227**
4. Sadism				1	.265**	-.199*
5. Total Victim Blame					1	.043 .316
6. Total Perceived Severity						1

Note: *$p < .05$; **$p < .01$.; ***$p < .001$; $n = 125$.

Table 4
Pearson's correlations of dark tetrad components, VB, and PS ratings by initial tweet valence.

	Negative		Neutral		Positive	
	VB	PS	VB	PS	VB	PS
Psychopathy	-.312***	-.187*	.277**	-.179*	.393***	-.207**
Narcissism	-.278** .204	-.075	.231**	-.103 .127	.324***	-.086 .170
Machiavellianism	-.090 .159	-.210**	.231**	-.246**	.312***	-.104 .123
Sadism	-.352***	-.187*	.311***	-.170*	.491***	-.136 .065
Negative – VB	1	.104 .124	–	–	–	–
Negative – PS		1	–	–	–	–
Neutral – VB			1	.009 .460	–	–
Neutral – PS				1	–	–
Positive – VB					1	.002 .490
Positive – PS						1

Note: Upper row = Correlation coefficients, lower row = significance value (*$p <$ .05; **$p < .01$.; ***$p < .001$); $N = 125$.

failed to predict significant variance. This suggests that as Machiavellianism increases, PS of abuse following neutral tweets decreases (see Table 5).

4. Discussion

We examined the independent and combined effects of lay-user initial tweet valence and abuse volume generated by other lay-users, on participant-observer perceptions of victim blame (VB) and perceived incident severity (PS). This follows on from the work of Scott et al. (2020) who examined VB and PS of twitter abuse, but only involving celebrity 'victims'. In the current study, we examined the role of participant-observers' Dark Tetrad personality factors on their perceptions – another extension from Scott et al. (2020) who had only examined only Dark *Triad* dimensions. Our first hypothesis (H_{1a}) was upheld – greater VB was observed following negative initial tweets than either neutral or positive tweets. Our second hypothesis (H_{1b}) was rejected – there was no significant main effect of initial tweet valence on PS ratings. As predicted (H_2), an increased volume of abuse resulted in higher PS ratings. Finally, as hypothesised, participants who scored higher on Dark Tetrad dimensions attributed greater VB (H_{3a}) and lesser PS (H_{3b}) than those who scored lower on these personality dimensions.

Similar to the findings of Scott et al. (2020), our results reveal a difference between identity claims (lay-user initial tweets) and behavioural residue (volume of abuse received) on observers' interpretation of online abuse incidents. We found that the valence of a lay-user's initial tweet influenced VB, but not PS, with higher VB following negative initial tweets than either neutral or positive initial tweets (which did not differ from each other). We found that the volume of abuse received influenced both VB and PS – both measures were higher following high versus low volumes of abuse. A significant interaction between valence and volume was observed on ratings of PS – across all tweet valences, PS was higher following high versus low abuse; however, the simple main effects of valence were non-significant at each level of abuse volume. Regression analyses revealed that Sadism was a significant predictor of VB attribution, regardless of initial tweet valence; PS of abuse was predicted by Psychopathy following emotionally-salient initial tweets (negative or positive valence), whereas Machiavellianism was the sole predictor of PS following neutral initial tweets.

4.1. The impact of initial tweet valence and abuse volume on attributed victim blame

The current results showed that when lay-tweeters are abused, VB judgements were shaped by both identity claims (initial tweet valence) and behavioural residue (volume of abuse). This is in contrast to Scott et al.'s (2020) celebrity tweeters, for whom only identity claims influenced VB attributions. The finding that more blame was attributed following negative (vs. neutral or positive) initial tweets supports H_{1a}. It also aligns with previous findings that, in online environments such as social media, both celebrity- and lay-users are blamed for abuse they received if there is evidence that they did something to provoke such a response (DeSmet et al., 2012; Scott et al., 2020; Shultz et al., 2014; Weber et al., 2013).

The current findings contrast with those of Scott et al. (2020) in how observers view lay-versus celebrity-users of Twitter, and which warrants that observers use in order to make judgements about profile owners/victims. In the current study, VB of lay-users was only inflated following a negative initial tweet, whereas for celebrity-users more VB was attributed following a neutral tweet than following a positive tweet (Scott et al., 2020). This indicates that lay-users are conceptualized somewhat differently than celebrities who might be assumed to be using the site primarily for personal gain and self-promotion, rather than less overtly self-serving objectives (Lim, 2017; Yoo et al., 2012). As such lay-compared to celebrity-users are given more 'benefit of the doubt', and only receive increased blame if they were overtly negative, rather than being merely non-positive.Whereas blame attributed to celebrity Twitter users was dependent exclusively on identity claims (initial tweet), the current study suggests that both identity claims and behavioural residue

Table 5
Summary of stepwise regressions for victim blame and perceived severity across initial tweet valence.

Outcome	Valence	Predictor	R	R^2	R^2_{adj}	F	p	β
Victim Blame	Negative	Sadism	-.352	.124	.117	17.392	<.001	−0.352
	Neutral	Sadism	.311	.097	.089	13.174	<.001	0.311
	Positive	Sadism	.491	.241	.235	39.049	<.001	0.491
Perceived Severity	Negative	Psychopathy	-.306	.094	.087	12.744	.001	−0.306
	Neutral	Machiavellianism	-.246	.060	.053	7.916	.006	−0.246
	Positive	Psychopathy	-.207	.043	.035	5.524	.020	−0.207

Note: Valence = Initial Tweet Valence; ANOVA degrees of freedom = 1,123; β = Standardised Coefficient.

(volume of abuse) were taken into account when attributing VB against lay-users. More blame was attributed to victims following a high volume of abuse from other lay-users. This pattern is different from reported VB against lay-victims of online abuse on Facebook (Scott et al., 2019) and could highlight observers' understanding of the distinct relationships between users of these different social networking sites.

Scott et al. (2019) found no difference in VB when abuse came from multiple sources, and the opposite pattern (more blame attributed following a lower volume of abuse) when abuse came from a single source. Explanations for this seemingly counterintuitive finding focused on the close personal nature of friendships on Facebook, with the majority of friends on the site also having offline interactions (e.g., Ellison et al., 2007). It was proposed by Scott et al. (2019) that comments which might superficially appear abusive were 'banter' and evidence of a dark but playful relationship between online interlocutors, because good-natured teasing is a common facilitator of strong friendships (Shultz et al., 2014; Tragesser & Lippman, 2005).

Because interactions on Twitter are not limited to those who are friends, or even acquaintances, in real life (Phua et al., 2017), it would seem that observers judge these interactions differently, and do not necessarily assume that negative content is intended in a friendly manner. Fewer assumptions can be made about the relationships between users interacting on Twitter compared to such interactions on Facebook. This may explain why Twitter observers utilise behavioural residue as well as identity claims, rather than only the latter (as is the case when judging Facebook interactions), as sources of information when attributing blame. This finding has potential implications going forward as the diversity of social media continues to grow. Depending on the nature of individual platforms, and the interaction allowed between users having little or no social connection, then assumptions about the nature of existing relationships may be used as a lens through which to filter observed information. In an already relatively impoverished online environment, the hyperpersonal model predicts that such assumptions would likely lead to a magnified effect (Walther, 1996, 1997), and in the case of online abuse, have an increased impact on online victims.

4.2. The impact of initial tweet valence on perceived severity

Our findings illustrate that PS of lay-person abuse was influenced only by behavioural residue (i.e., volume of abuse), supporting H_2 but not H_{1b}. This is in contrast to Scott et al. (2020) whose celebrity victim PS data was found to be influenced by both identity claims and behavioural residue. The present findings of volume of abuse effects on both VB and PS are consistent with classifications of online abuse, including cyberbullying and cyberstalking, that place emphasis on the importance of frequency of abuse as a driver of negative victim experience (e.g., Garett et al., 2016; Menesini & Nocentini, 2009). To draw a parallel with our study, frequency of abuse is indexed by a high volume of abusive content within a chronological record of interaction – in this case, abusive replies to an initial tweet.

In contrast to celebrity-users of social media, lay-users are likely perceived as being more genuine (e.g., Lim, 2017). This could explain why PS was not influenced by the nature of the initial tweet, and H_{1b} was not supported. Scott et al. (2020) speculated that some celebrity-users may inhabit online characters or personas, posting negative or controversial content to garner attention and elicit reactions. As such, tweets would be tools to attract publicity, and backlash in the form of abuse may not negatively impact the celebrity who posted them, but instead be welcomed as the desired reaction. As most lay-users do not use Twitter in this way, it is likely that the current results reflect an understanding of this of participant observers. Lay-users, whether posting positive, negative, or neutral tweets, do so genuinely, and thus any negative reaction they receive as a consequence will likely impact them negatively.

It is likely that PS of any abusive incident is judged in much the same way across most social networking sites which conform to the same basic layout as both Twitter and Facebook: segments of content (either a tweet or a timeline status update) which can be reacted to or commented on by third parties, appearing in a chronological order from top to bottom. Both form a record of warrants documenting interactions from which observers can draw conclusions about the nature of any abuse. While it is probable that the nature of the relationship between interactants on any social networking platform (e.g., close friends vs. potential strangers) may colour any interpretation of this digital record, volume of abuse, as manifested here, will comprise a key component on any judgements of online abuse (Garett et al., 2016).

4.3. Outcomes regarding for lay-person social media use

Taken together, the findings for both VB and PS have implications for the application of Warranting Theory to online impression formation (Walther & Parks, 2002). Our findings reveal that not only are different categories of warrants (identity claims and behavioural residue) considered differently when forming impressions of lay-versus celebrity-tweeters, but that the content of identity claims contribute differentially to impressions. Negative content has previously been shown to carry more weight than positive content (Walther et al., 2009). For lay-users, negative tweets resulted in increased VB compared to both neutral and positive tweets. For celebrities, the impact of identity statements (initial celebrity tweets) were less categorical and more continuous, with more negative tweets resulting in increased blame attribution (Scott et al., 2019). This extends previous findings that identity claims can carry at least as much weight in impression formation as behavioural residue (Fullwood et al., 2015; Scott & Ravenscroft, 2017).

Lay-user identity claims did not affect PS; PS was determined solely by the actions of others, another finding distinct from celebrity abuse judgments (Scott et al., 2019). This finding adds to growing evidence indicating that aspects of Warranting Theory need to be reconsidered. Due to the evolution of technology and the consequent expansion of social media, new affordances have become available and the interaction between different types of users has been facilitated (e.g., Dare-Edwards, 2014; Scott et al., 2019; Weber et al., 2013).

Current findings outline that judgments of AB and PS on lay-victims of online abuse reflect a tacit acknowledgment on the part of observers that such users are different from celebrity users in their use of social media. Whereas celebrities may use the platform for self-promotion (Marwick & Boyd, 2010), lay-users are more genuine (Tosun, 2012). As such, lay-users are only attributed blame if they receive abuse after posting explicitly negative content (vs. celebrities for whom blame attribution is more continuous). Also, it is acknowledged that incident severity is independent of a lay-users' own actions.

Compared to celebrities, lay-users seem to be perceived more 'fairly' when subjected to online abuse, but they are still attributed *some* blame. This accords with previous findings which highlight a lack of sympathy and support for victims (Chen et al., 2015; Dredge et al., 2014; Gahagan et al., 2016), and is troubling given the negative outcomes associated with online abuse (Gini & Espelage, 2014; van Geel et al., 2014). Due to the serious nature of potential consequences (e.g., Hinduja & Patchin, 2010), this topic should continue to be investigated to better understand the cognitive processes underlying online victim blame.

4.4. The role of dark tetrad factors

Following on from the work of Scott et al. (2020), who examined the influence of participants Dark Triad personality dimensions on judgements of VB and PS, we considered the additional dimension of *Sadism*. A large body of research demonstrates consistent links between the Dark Tetrad (DT) personality traits and deficits in both affective and cognitive empathy (e.g. Buckels et al., 2013; Pajevic et al., 2018). A reduced capacity for perspective-taking and physiological responses to others' distress may influence how individuals high in these traits interpret and respond to observed instances of online trolling.

In the present study, results indicated that sadism, specifically, predicted victim blame in instances where the abuse bore no relation to the

valence of the initial tweet. While previous literature indicates that sadism is a significant predictor of engagement in cyberbullying and trolling behaviours (Brown et al., 2019; Buckels et al., 2014), ours is the first study to consider sadistic attitudes towards abuse perpetrated by others. As those high in this trait exude pleasure in witnessing the pain or distress of others (Meere & Egan, 2017), it is unsurprising that this trait also predicts higher attribution of blame to a victim of cyberabuse. Our findings suggest that the pleasure sadists experience at others' misfortune comes not from the perpetration of the abuse itself, but from its presumed negative effect on the unobserved reaction of the victim – the outcome of which remains the same regardless of the abuse perpetrator.

This study also found that psychopathy was a significant predictor of perceived severity of abuse that followed positive tweets by the victim. The association between psychopathy and perceived severity was also found by Scott et al. (2020), whose results indicated that psychopathy was a significant predictor of PS in the positive tweet condition. This suggests that positive initial tweets by both celebrities and lay-users result in reduced perceived severity of abuse received by the victim in those high in this trait. As highlighted by Scott et al., it's possible that the positive tweets were deemed by participants high in psychopathy as 'showing off' which, given their superiority complex and tendency towards envy (Jonason et al., 2015; Walker & Jackson, 2017), may make these individuals view the abuse as deserved. However, our study also found psychopathy to be predictive of PS following negative tweets. In this case, it may be that those high in psychopathy view the abuse as less severe because the initial negative tweets insinuate that the individual can handle the abuse they receive. Again, due to their feelings of superiority, they may view a lay-user target as even more inferior than the celebrity targets used by Scott et al., and so may be more likely to attribute minimised impact on the victim.

Machiavellianism was found to predict perceived severity of abuse following neutral victim tweets only. The fact that Machiavellianism was not predictive of VB or PS in the negative or positive tweet condition is in line with the findings of Scott et al. (2020). However, the current study indicates that Machiavellianism may be relevant in instances where the victim's initial tweet is neutral in valence. Given the goal-oriented nature of Machiavellianism (Deluga, 2001), it may be that these individuals view any neutral social media posts as futile and non-goal-directed. For this reason, they may demonstrate less sympathy as to the impact of the abuse on that individual.

Interestingly, narcissism was not a significant predictor of victim blame or perceived severity of abuse. This is in contrast to the findings of Scott et al. (2020), who found narcissism was the sole predictor of VB and PS following negative tweets by celebrities. A fundamental aspect of narcissism is heightened ego-threat monitoring (Horvath & Morf, 2009), making those high in this trait quick to respond negatively and aggressively to potential ego threats (whether real or imagined). It is possible that the lay-users portrayed in this experiment represented a lesser threat to those high in narcissism than did the successful and wealthy celebrities portrayed in Scott et al.'s study. Those high in narcissism may find it harder to relate to the 'everyday person' given their feelings of eminence, and therefore may be somewhat disinterested and unresponsive to observed instances of abuse. Indeed, the literature supports the notion that narcissism is associated with enhanced interest in celebrities (Greenwood et al., 2018) and celebrity worship (Ashe et al., 2005).

4.5. Limitations and future research

The current research does have some limitations. The most notable of these is the fact that all stimuli involved male 'victims'. This was done for two reasons: first, so that results could be directly compared to previous work by Scott et al. (2020) who used the same manipulation, and second, because twitter is used predominantly by males (Statistia, 2018a, 2018b). Further, there is some tentative evidence to suggest that well-known males are the most-abused users on this platform (Demos, 2014). We think it likely that female twitter users would be perceived in a similar fashion, but further research is required to confirm the generalisability of results. Another gender-sex imbalance can be found in the sample of participants, the majority of whom were female ($n = 84$). Although we believe that the number of male respondents was adequate ($n = 39$), future research should aim to recruit a more balanced set of participants by gender-sex (or equal proportions of 'men' and 'women') – or, indeed, proportionate, representative groups of male, female, non-binary/trans respondents. Previous research into victim blame has found that men are more likely to both attribute blame to victims (e.g., Gerber et al., 2004; Grubb & Turner, 2012) and downplay the severity of abusive incidents (Ben-David & Schneider, 2005; Davies et al., 2008). We suggest that, if anything, the inclusion of more males in the sample would strengthen the pattern of effects found here. The four- and two-item scales which measure VB and PS, respectively, could be considered to be problematic, as scales with few items might be vulnerable to 'extreme' responses to particular items. The Cronbach's alpha associated with the VB scale was very good ($\alpha = 0.90$), and although the Cronbach's alpha associated with the two-item PS scale was not particularly good ($\alpha = 0.65$), the associated inter-item correlations were more-than-adequate ($p < .001$).

Two extensions of the current research relate to the generalisability of results across social media platforms, and the effects of other individual differences not measured here. The differences between abuse on Twitter and Facebook have been discussed above, but other social media platforms are available (e.g., Whatsapp, Snapchat, Instagram), each of which not only offer distinct modes of communication, but afford different ways of posting content which may be commented upon by others, and assume different relationships between online communicators. Further systematic investigations of the factors contributing to perceptions of VB and severity of abuse across platforms will advance our understanding of the cognitions behind such attributions. Finally, other personality factors such as the Big Five (Costa & McCrae, 1985) and self-esteem (e.g., Baumeister et al., 2003) have been shown to mediate online behaviour, and their role in how observers attribute VB and PS in instances of online abuse could be investigated.

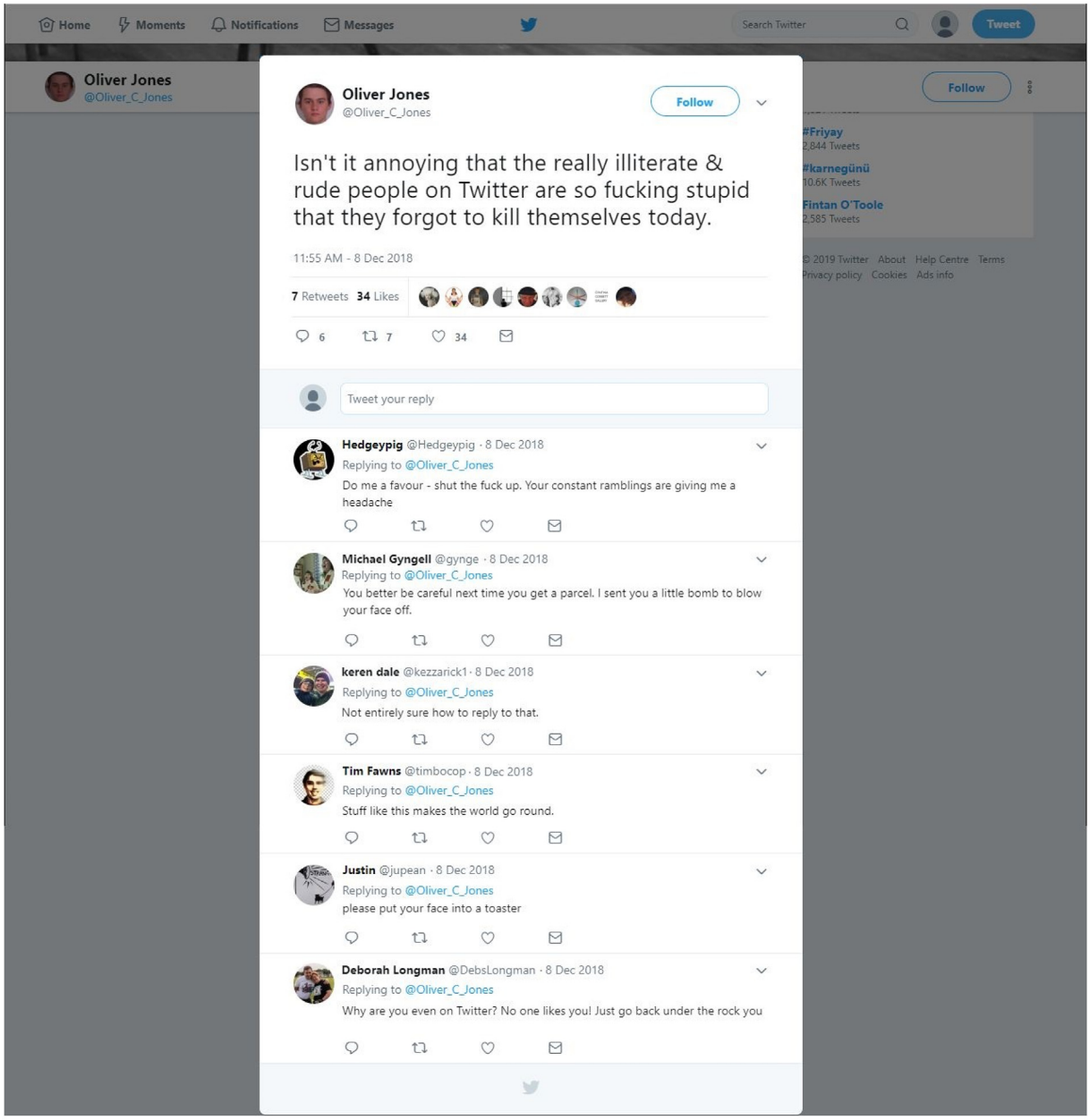
5. Conclusions

We have shown that online impression formation of lay-user victims of Twitter abuse is influenced by both the actions of the lay-user victim (initial tweet valence; identity claims) and the actions of the abusers (volume of abuse; behavioural residue). These findings are in partial contrast to those of Scott et al. (2020) who investigated perceptions of abused celebrities and found that initial tweet valence/identity claims were the principal driver of attributed victim blame (VB) and perceived incident severity (PS). In the current study, we considered the role of observer *Dark Tetrad* personality characteristics, and found that observer sadism scores predicted VB regardless of the valence of victims' initial tweets; PS judgements were predicted by observer psychopathy scores when initial tweets were highly emotional (positive, negative). Taken together, our findings provide novel and informative insights into observer perceptions of the victims of online abuse, and demonstrate a complex but rational interplay between victim-generated content, the responses of online 'abusers', and individual differences within the cohort of observers.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Example Experimental Stimulus



Note: Negative initial tweet Valence, High abuse Volume condition

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *The Journal of Economic Perspectives, 31*(2), 211–236. https://doi.org/10.1257/jep31.2.211

Aoyama, I., Barnard-Brak, L., & Talbert, T. (2011). Cyberbullying among high school students: Cluster analysis of sex and age differences and the level of parental monitoring. *International Journal of Cyber Behavior, Psychology and Learning, 1*(1), 25–35. https://doi.org/10.4018/ijcbpl.2011010103

Ashe, D., Maltby, J., & McCutcheon, L. E. (2005). Are celebrity-worshipers more prone to narcissism? A brief report. *North American Journal of Psychology, 7*, 239–246.

Baumeister, R. F., Campbell, J. D., Krueger, J. I., & Vohs, K. D. (2003). Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles? *Psychological Science in the Public Interest, 4*(1), 1–44.

Ben-David, S., & Schneider, O. (2005). Rape perceptions, gender role attitudes, and victim-perpetrator acquaintance. *Sex Roles, 53*(5/6), 385–399. https://doi.org/10.1007/s11199-005-6761–6764

British Psychological Society. (2014). *Code of human research Ethics*. Leicester: BPS.

Brown, W. M., Hazraty, S., & Palasinski, M. (2019). Examining the dark tetrad and its links to cyberbullying. *Cyberpsychology, Behavior, and Social Networking, 22*(8), 552–557.

Buckels, E. E., Jones, D. N., & Paulhus, D. L. (2013). Behavioral confirmation of everyday sadism. *Psychological Science, 24*(11), 2201–2209.

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences, 67*, 97–102.

Bursac, Z., Gauss, C. H., Williams, D. K., & Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine, 3*(17). https://doi.org/10.1186/1751-0473-3-17

Chaffey, D. (2019). *Global social media research summary 2020*. Smartinsights.com. https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/. (Accessed 31 March 2020).

Chen, L. M., Cheng, W., & Ho, H. (2015). Perceived severity of school bullying in elementary schools based on participants' roles. *Educational Psychology, 35*, 484–496. https://doi.org/10.1080/01443410.2013.860220

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.

Costa, P. T., & McCrae, R. R. (1985). *The NEO personality inventory manual*. Odessa, FL: Psychological Assessment Resources.

Craker, N., & March, E. (2016). The dark side of Facebook®: The Dark Tetrad, negative social potency, and trolling behaviours. *Personality and Individual Differences, 102*, 79–84.

Dare-Edwards, H. L. (2014). 'Shipping bullshit': Twitter rumours, fan/celebrity interaction and questions of authenticity. *Celebrity Studies, 5*(4), 521–524. https://doi.org/10.1080/19392397.2014.981370

Davies, M., Rogers, P., & Bates, J. (2008). Blame towards male rape victims in a hypothetical sexual assault as a function of victim sexuality and degree of resistance. *Journal of Homosexuality, 55*(3), 533–544. https://doi.org/10.1080/00918360802345339

Deluga, R. J. (2001). American presidential Machiavellianism: Implications for charismatic leadership and rated performance. *The Leadership Quarterly, 12*, 334–363.

Demos. (2014). Male celebrities receive more abuse on Twitter than women. https://www.demos.co.uk/press-release/demos-male-celebrities-receive-more-abuse-on-twitter-than-women-2/.

DeSmet, A., Bastiaensens, S., Van Cleemput, K., Poels, K., Vandebosch, H., & De Bourdeaudhuij, I. (2012). Mobilizing bystanders of cyber-bullying: An exploratory study into behavioural determinants of defending the victim. *Annual Review of Cybertherapy and Telemedicine, 181*, 58–63.

Douglas, H., Bore, M., & Munro, D. (2012). Construct validity of a two-factor model of psychopathy. *Psychology, 3*(3), 243–248.

Dredge, R., Gleeson, J. F. M., & de la Piedad Garcia, X. (2014). Risk factors associated with impact severity of cyberbullying victimization: A qualitative study of adolescent online social networking. *Cyberpsychology, Behavior, and Social Networking, 17*, 287–291. https://doi.org/10.1089/cyber.2013.0541

Ellison, N., Steinfeld, C., & Lampe, C. (2007). The benefits of Facebook "friends": Exploring the relationship between college students' use of online social friends' networks and social capital. *Journal of Computer-Mediated Communication, 12*(4), 1143–1168.

Embury-Dennis, T. (2016). The 25 most common surnames in Britain – and what they say about your family history. https://www.independent.co.uk/news/uk/home-news/the-25-common-surnames-britain-family-history-university-west-england-bristol-uk-a7423196.html.

Fullwood, C., Quinn, S., Chen-Wilson, J., Chadwick, D., & Reynolds, K. (2015). Put on a smiley face: Textspeak and personality perceptions. *Cyberpsychology, Behavior, and Social Networking, 18*(3), 147–151.

Gahagan, K., Vaterlaus, J. M., & Frost, L. R. (2016). College student cyberbullying on social networking sites: Conceptualization, prevalence, and perceived bystander responsibility. *Computers in Human Behavior, 55*(B), 1097–1105. https://doi.org/10.1016/j.chb.2015.11.019

Garcia, D., & Sikström, S. (2014). The dark side of Facebook: Semantic representations of status updates predict the Dark Triad of personality. *Personality and Individual Differences, 67*, 92–96. https://doi.org/10.1016/j.paid.2013.10.001

Garett, R., Lord, L. R., & Young, S. D. (2016). Associations between social media and cyberbullying: A review of the literature. *mHealth, 2*(46). https://doi.org/10.21037/mhealth.2016.12.01

Gayle, S. S., & Lawson, K. (2013). Twitter as a way for celebrities to communicate with fans: Implications for the study of parasocial interaction. *North American Journal of Psychology Winter Garden, 15*(2), 339–354.

van Geel, M., Goemans, A., Toprak, F., & Vedder, P. (2017). Which personality traits are related to traditional bullying and cyberbullying? A study with the Big five, dark triad and sadism. *Personality and Individual Differences, 106*(1), 231–235.

van Geel, M., Vedder, P., & Tanilon, J. (2014). Relationship between peer victimization, cyberbullying, and suicide in children and adolescents: A meta-analysis. *JAMA Pediatrics, 168*(5), 435–442. https://doi.org/10.1001/jamapediatrics.2013.4143

Gerber, G. L., Cronin, J. M., & Steigman, H. J. (2004). Attributions of blame in sexual assault to predators and victims of both genders. *Journal of Applied Social Psychology, 34*(10), 2149–2165.

Gibb, Z. G., & Devereux, P. G. (2014). Who does that anyway? Predictors and personality correlates of cyberbullying in college. *Computers in Human Behavior, 38*, 8–16.

Gini, G., & Espelage, D. L. (2014). Peer victimization, cyberbullying, and suicide risk in children and adolescents. *Journal of the American Medical Association, 312*, 545. https://doi.org/10.1001/jama.2014.3212

Goodboy, A. K., & Martin, M. M. (2015). The personality profile of a cyberbully: Examining the Dark Triad. *Computers in Human Behavior, 49*, 1–4.

Greenwood, D., McCutcheon, L. E., Collisson, B., & Wong, M. (2018). What's fame got to do with it? Clarifying links among celebrity attitudes, fame appeal, and narcissistic subtypes. *Personality and Individual Differences, 131*, 238–243.

Grubb, A., & Turner, E. (2012). Attribution of blame in rape cases: A review of the impact of rape myth acceptance, gender role conformity and substance use on victim blaming. *Aggression and Violent Behavior, 17*, 443–452. https://doi.org/10.1016/j.avb.2012.06.002

Hargittai, E., & Litt, E. (2011). The tweet smell of celebrity success: Explaining variation in Twitter adoption among a diverse group of young adults. *New Media & Society, 13*(5), 824–842. https://doi.org/10.1177/1461444811405805

Hearn, J., & Hall, M. (2019). New technologies, image distribution and cyberabuse. *NOTA News, 88*, 10–13.

Hinduja, S., & Patchin, J. (2010). Bullying, cyberbullying, and suicide. *Archives of Suicide Research, 14*, 206–221.

Horvath, S., & Morf, C. C. (2009). Narcissistic defensiveness: Hypervigilance and avoidance of worthlessness. *Journal of Experimental Social Psychology, 45*(6), 1252–1258.

John, A., Glendenning, A. C., Marchant, A., Montgomery, P., Stewart, A., Wood, S., Lloyd, K., & Hawton, K. (2018). Self-harm, suicidal behaviours, and cyberbullying in children and young people: Systematic review. *Journal of Medical Internet Research, 20*(4), e129. https://doi.org/10.2196/jmir.9044

Johnson, L. K., Plouffe, R. A., & Saklofske, D. H. (2019). Subclinical sadism and the dark triad: Should there be a dark tetrad? *Journal of Individual Differences, 40*(3), 127–133. https://doi.org/10.1027/1614-0001/a000284

Jonason, P. K., Wee, S., & Li, N. P. (2015). Competition, autonomy, and prestige: Mechanisms through which the Dark Triad predict job satisfaction. *Personality and Individual Differences, 72*, 112–116.

Jones, D. N., & Paulhus, D. L. (2013). Introducing the short dark triad (SD3): A brief measure of dark personality traits. *Assessment, 21*(1), 28–41. https://doi.org/10.1177/1073191113514105

Jurgens, D., Hemphill, L., & Chandrasekharan, E. (2019). A just and comprehensive strategy for using NLP to address online abuse. *Proceedings of ACL*. https://arxiv.org/abs/1906.01738.

Kurek, A., Jose, P. E., & Stuart, J. (2019). 'I did it for the LULZ': How the dark personality predicts online disinhibition and aggressive online behavior in adolescence. *Computers in Human Behavior, 98*, 31–40. https://doi.org/10.1016/j.chb.2019.03.027

Lee, J., & Lim, Y.-S. (2016). Generated campaign tweets: The cases of hillary Clinton and Donald Trump. *Public Relations Review, 42*(5), 849–855. https://doi.org/10.1016/j.pubrev.2016.07.004

Lerner, M., & Simmons, C. H. (1966). Observer's reaction to the 'innocent victim': Compassion or rejection? *Journal of Personality and Social Psychology, 4*(2), 203–210.

Lim, Y. J. (2017). Decision to use either Snapchat or Instagram for most powerful celebrities. *Research Journal of the Institute for Public Relations, 3*(2), 1016.

Maple, C., Short, E., Brown, A., Bryden, C., & Salter, M. (2012). Cyberstalking in the UK: Analysis and recommendations. *International Journal of Distributed Systems and Technologies, 3*(4), 34–51.

Marwick, A. E., & Boyd, D. (2010). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society, 13*(1), 114–133. https://doi.org/10.1177/1461444810365313

Mechanic, M. B., Uhlmansiek, M. H., Weaver, T. L., & Resick, P. A. (2000). The impact of severe stalking experienced by acutely battered women: An examination of violence, psychological symptoms, and strategic responding. *Violence & Victims, 15*, 443–458.

Meere, M., & Egan, V. (2017). Everyday sadism, the Dark Triad, personality, and disgust sensitivity. *Personality and Individual Differences, 112*, 157–161. https://doi.org/10.1016/j.paid.2017.02.056

Mendez, I., Jorquera, A. B., Ruiz-Esteban, C., Martinez-Ramon, J. P., & Fernandez-Sogorb, A. (2019). Emotional intelligence, bullying, and cyberbullying in adolescents. *International Journal of Environmental Research and Public Health, 16*(23), 4837. https://doi.org/10.3390/ijerph16234837

Menesini, E., & Nocentini, A. (2009). Cyberbullying definition and measurement: Some critical considerations. *Journal of Psychology, 217*(4), 230–232.

Menesini, E., Nocentini, A., Palladino, B. E., Frisén, A., Berne, S., Ortega-Ruiz, R., & Smith, P. K. (2012). Cyberbullying definitions among adolescents: A comparison across six European countries. *Cyberpsychology, Behavior, and Social Networking, 15*(9), 455–463.

Međedović, J., & Petrović, B. (2015). The Dark Tetrad: Structural properties and location in the personality space. *Journal of Individual Differences, 36*(4), 228–236. https://doi.org/10.1027/1614-0001/a000179

Millon, T., & Davis, R. D. (1996). *Disorders of personality: DSM-IV and beyond*. New York: John Wiley & Sons.

Mohsin, M. (2020). *10 social media statistics you need to know in 2020*. Oberlo.co.uk. https://www.oberlo.co.uk/blog/social-media-marketing-statistics. (Accessed 31 March 2020).

Nadkarni, A., & Hofmann, S. G. (2012). Why do people use Facebook? *Personality and Individual Differences, 52*(3), 243–249. https://doi.org/10.1016/j.paid.2011.11.007

Pabian, S., De Backer, C. J. S., & Vandebosch, H. (2015). Dark Triad personality traits and adolescent cyber-aggression. *Personality and Individual Differences, 75*, 41–46.

Pajevic, M., Vukosavljevic-Gvozden, T., Stevanovic, N., & Neumann, C. S. (2018). The relationship between the Dark Tetrad and a two-dimensional view of empathy. *Personality and Individual Differences, 123*(1), 125–130.

Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: Narcissism, machiavellianism, and psychopathy. *Journal of Research in Psychology, 36*, 556–563.

Penton-Voak, I., Perrett, D., Castles, D., et al. (1999). Menstrual cycle alters face preference. *Nature, 399*, 741–742. https://doi.org/10.1038/21557

Phua, J., Jin, S. V., & Kim, J. (2017). Uses and gratifications of social networking sites for bridging and bonding social capital: A comparison of Facebook, twitter, Instagram, and Snapchat. *Computers in Human Behavior, 72*, 115–122. https://doi.org/10.1016/j.chb.2017.02.041

Russell, K. J., & Hand, C. J. (2017). Rape myth acceptance, victim blame attribution and just world beliefs: A rapid evidence assessment. *Aggression and Violent Behavior, 37*, 153–160. https://doi.org/10.1016/j.avb.2017.10.008

Scott, G. G., Brodie, Z. P., Wilson, M. J., Ivory, L., Hand, C. J., & Sereno, S. C. (2020). Celebrity abuse on Twitter: The impact of tweet valence, volume of abuse, and dark triad personality factors on victim blaming and perceptions of severity. *Computers in Human Behavior, 103*, 109–119. https://doi.org/10.1016/j.chb.2019.09.020

Scott, G. G., & Ravenscroft, K. (2017). Bragging on Facebook: The interaction of content source and focus on online impression formation. *Cyberpsychology, Behavior, and Social Networking, 20*(1), 58–63. https://doi.org/10.1089/cyber.2016.0311

Scott, G. G., Wienercz, S., & Hand, C. J. (2019). The frequency and source of online abuse impacts attribution of victim blame and perceptions of victim attractiveness. *Computers in Human Behavior, 92*, 119–127. https://doi.org/10.1016/j.chb.2018.10.037

Shaver, K. G. (1970). Defensive attribution: Effects of severity and relevance on the responsibility assigned for an accident. *Journal of Personality and Social Psychology, 14*, 101–113.

Short, E., & McMurray, I. (2009). Mobile phone harassment: An exploration of students' perceptions of intrusive texting behaviour. *Human Technology, 5*(2), 163–180.

Shultz, E., Heilman, R., & Hart, K. J. (2014). Cyber-bullying: An exploration of bystander behavior and motivation. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 8*(4). https://doi.org/10.5817/CP2014-4-3. article 3.

Statistia. (2018a). Distribution of Twitter users worldwide as of July 2018, by gender Accessed 17th August 2018 from https://www.statista.com/statistics/828092/distribution-of-users-on-twitter-worldwide-gender/.

Statistia. (2018b). *Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2018 (in millions)*. Statistia Accessed 19th June 2018 from https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.

Sumner, C., Byers, A., Boochever, R., & Park, G. J. (2012). Predicting Dark Triad personality traits from Twitter usage and a linguist analysis of tweets. *Paper presented at the international conference of machine learning and applications (IMCLA)*.

Tiddeman, B. P., Burt, D. M., & Perrett, D. I. (2001). Prototyping and transforming facial textures for perception research. *IEEE Computer Graphics and Applications, 21*, 42–50.

Tosun, L. P. (2012). Motives for Facebook use and expressing "true self" on the internet. *Computers in Human Behavior, 28*, 1510–1517.

Tragesser, S. L., & Lippman, L. G. (2005). Teasing: For superiority or solidarity? *The Journal of General Psychology, 132*(3), 255–266.

Vakhitova, Z., Alston-Knox, C. L., Reynald, D. M., Townsley, M. K., & Webster, J. L. (2019). Lifestyles and routine activities: Do they enable different types of cyber abuse? *Computers in Human Behavior, 101*, 225–237. https://doi.org/10.1016/j.chb.2019.07.012

Villlanti, A. C., Johnson, A. L., Ilakkuvan, V., Jacobs, M. A., Graham, A. L., & Rath, J. M. (2017). Social media use and access to digital technology in US young adults in 2016. *Journal of Medical Internet Research, 19*(6), e196. https://doi.org/10.2196/jmir.7303

Walker, B. R., & Jackson, C. J. (2017). Moral emotions and corporate psychopathy: A review. *Journal of Business Ethics, 141*(4), 797–810.

Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research, 23*(1), 3–43.

Walther, J. B. (1997). Group and interpersonal effects in international computer-mediated collaboration. *Human Communication Research, 23*(3), 342–369.

Walther, J. B., & Parks, M. R. (2002). Cues filtered out, cues filtered in: Computer-mediated communication and relationships. In M. L. Knapp, & J. A. Daly (Eds.), *Handbook of interpersonal communication* (3rd ed., pp. 529–563). Thousand Oaks, CA: Sage.

Walther, J. B., Van Der Heide, B., Hamel, L., & Shulman, H. (2009). Self-generated versus other-generated statements and impressions in computer mediated communication: A test of warranting theory using Facebook. *Communication Research, 36*, 229–253.

Weber, M., Ziegele, M., & Schnauber, A. (2013). Blaming the victim: The effects of extraversion and information disclosure on guilt attributions in cyberbullying. *Cyberpsychology, Behavior, and Social Networking, 16*(4), 254–259.

Welling, L. L. M., Jones, B. C., DeBruine, L. M., Conway, C. A., Law Smith, M. J., Little, A. C., et al. (2007). Raised salivary testosterone in women is associated with increased attraction to masculine faces. *Hormones and Behavior, 52*, 156–161.

Yoo, J., Choi, S., Choi, M., & Rho, J. (2012). Why people use Twitter: Social conformity and social value perspectives. *Online Information Review, 38*(2), 265–283. https://doi.org/10.1108/OIR-11-2012-0210