THE FACELESS: ANONYMITY AND PERSONALITY IN DIGITAL AGGRESSION

By

Mikayla Kim

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Psychology – Doctor of Philosophy

2023

ABSTRACT

Despite many benefits, one consequence of technology's proliferation and ritualized use is cyberbullying or digital aggression (DA). One key theory for understanding DA is the General Aggression Model. This model proposes that DA can be best understood as a consequence of interactions between personal and situational factors. Consistent with this theory, personality and anonymity have been identified as independent predictors of DA. Critically, however, virtually no research has sought to investigate whether and how these two factors may interact in predicting DA. The first study sought to bridge this gap by coding real-world DA behaviors on Twitter. Our findings indicated that individuals high in intellect/imagination used more antisocial words when they had anonymous Twitter accounts but not identifiable ones. The second study examined the effects of personality and technical self- and other-anonymity, including their interactions, on DA using a recently developed in-vivo experimental paradigm, the TAP-Chat. We found that anonymity moderated the relationship between extraversion and in-vivo DA, such that individuals high in extraversion used more antisocial words during the experiment when the participant and co-player were fully identified compared to when both were fully anonymous. Our findings collectively illuminate the roles of personality and anonymity in the prevalence of DA while indicating that these associations are measure- and context-specific. Such findings have key implications for the field's understanding of DA and, in doing so, inform the development of policy and prevention and intervention programs for DA.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to those who have been with me along this journey and my academic career. First, I could not have undertaken this journey without Dr. Alex Burt and her unwavering support, exemplary guidance, and valuable feedback. Her commitment and passion for her students' professional development and success have kept me grounded, comforted, and determined. I am also extremely grateful to my doctoral guidance committee members, Drs. Brent Donnellan, Morgan Ellithorpe, and Brooke Ingersoll, for taking the time to serve on my committee. Their expertise, valuable suggestions, and insightful comments were of immense help throughout my dissertation. I would also like to thank my cohort, colleagues, and friends for always being there for me. Last, but not least, I am deeply indebted to my parents, Sihwi Kim and Soomi Lee, and my brother, Edwin Kim. As my biggest source of support, my graduate school journey would not have been possible without their reassurance, enthusiasm, humor, prayers, and love. My parents are my constant source of inspiration and determination, whose unconditional love and unparalleled support are with me in whatever I pursue and wherever I go.

TABLE OF CONTENTS

INTRODUCTION

Continued advancement of technology has enabled a globally connected world, but it also has become another platform for bullying behaviors. The exact definition of digital aggression (DA) varies across studies, but DA can be defined as the use of information and communications technology (ICT) to intentionally inflict harm on others (Burt & Alhabash, 2018). DA has many forms, including harassment (i.e., repeatedly sending offensive messages), flaming (i.e., online arguments that include profanity and vulgar language), exclusion (i.e., isolating the target), impersonation (i.e., pretending to be the target or another individual to communicate negative information to or about the target), stalking (i.e., following the target and repeatedly sending threatening messages), outing (i.e., sharing the target's secrets or personal information without consent), and non-consensual sexting (i.e., distributing of nude pictures of the target without consent) (Kowalski et al., 2014; Langos, 2015; Vandebosch & Van Cleemput, 2008). DA also occurs through various types of ICTs, such as cell phones (i.e., phone calls or text messages), computers, or tablets, and it takes place across multiple platforms, including social networking sites, email, instant messenger, internet forums, or online games (Wang et al., 2009; Ybarra et al., 2012). Other commonly used terms are cyberbullying, online aggression, and electronic bullying (Mehari et al., 2014), although each of these constructs is somewhat narrower in scope than DA (e.g., cyberbullying requires a power imbalance between the victim and the perpetrator). For the current project, DA will be defined as any aggressive acts committed online and/or electronically using any form of ICTs.

Critically, DA is now recognized as a global public health crisis (Jang et al., 2016; Kraft, 2006), with lifetime victimization rates ranging from 4.9% to 65.0%, and lifetime perpetration rates ranging from 1.2% to 44.1% across studies worldwide (Brochado et al., 2017). In the

United States, Patchin (2021) found that almost half (46%) of middle and high school students reported experiencing DA in their lifetime, while 14% admitted to perpetrating DA toward others. This very high prevalence of DA belies its severity. Experiences of DA are associated with adverse health impacts, including both internalizing problems such as depression, anxiety, and even suicidality (Bonanno & Hymel, 2013; Mitchell et al., 2007; Molero et al., 2022; Perren et al., 2010; Schenk & Fremouw, 2012; Selkie et al., 2015; Wang et al., 2011; Wigderson & Lynch, 2013) and externalizing problems such as increased rule-breaking behaviors, aggression, and delinquency (Alhajji et al., 2019; Goebert et al., 2011; Jung et al., 2014; Ybarra et al., 2007; Ybarra & Mitchell, 2004). In short, young adults are thus at considerable and sustained risk for DA and its negative mental health consequences.

Anonymity

Efforts to identify factors that increase the risk for DA are thus a critical research objective. Extant studies have identified anonymity as one of the key technological features that increase DA (Barlett et al., 2017; Barlett & Gentile, 2012; Harrison, 2015; Ooi et al., 2019; Reason et al., 2016). One key element of anonymity is its point of view. Self-anonymity (*you can't see me*) refers to the inability of others to identify the perpetrator, while other-anonymity (*I can't see you*) refers to the inability of the perpetrator to identify others (e.g., targets of DA). Another key element relates to the type of anonymity. Hayne and Rice (1997) identified two categories: technical anonymity and social anonymity. Technical anonymity describes the removal of personal identification, whereas social anonymity refers to the perception of self and/or others as anonymous due to the absence of social cues. For instance, individuals may use their first name to play online games with strangers and yet perceive themselves as anonymous.

It has been argued that online self-anonymity provides "safe spaces" for DA perpetrators by minimizing accountability and altering their sense of morality. Online self-anonymity also reduces the likelihood of the perpetrators being detected and punished while offering a sense of perceived power to the perpetrators (Badiuk, 2006; Mishna et al., 2009). Taking advantage of this situation, DA perpetrators tend to feel less restrained and engage in these online behaviors that they typically would not exhibit in a real-world environment (Joinson, 2007; Martin & Vieaux, 2016). Consequentially, self-anonymity can increase the level of frustration, fear, and powerlessness for the victims (Dooley et al., 2009; Slonje & Smith, 2008; Sticca & Perren, 2013; Vandebosch & Van Cleemput, 2008). DA victims reported experiencing greater fear of being attacked online by a faceless individual who knows their identity and heightened distress when considering anyone could be the perpetrator, including their classmates, friends, or anybody in their lives (Badiuk, 2006; Mishna et al., 2009). One student stated that while it is disappointing when you know the perpetrator, "it's bad if you don't know who it is because then, in principle, it could be anyone" (Nocentini et al., 2010).

Furthermore, online anonymity affords an environment devoid of social cues, and the absence of these social cues has been associated with higher levels of deindividuation and decreased levels of inhibition (Siegel et al., 1986; Sproull & Kiesler, 1986). One of the essential social cues is the lack of eye contact. Eye contact is important in interpersonal communication by providing, regulating, and expressing emotions, and in its absence, DA perpetrators may feel less inhibited (Lapidot-Lefler & Barak, 2012). For example, one 13-year-old girl explained, "It's not face-to-face. It's easier to say more hurtful comments because sometimes you don't like to say things to people's faces, but when you do it for revenge on MSN or something, it might be easier to do because you do not see how much they are hurt by it" (Mishna et al., 2009). Thus,

perpetrators do not “see” the pain they caused. Instead, they can attribute their own interpretation of the event and distort the consequences of their actions (Runions & Bak, 2015).

To better understand how anonymity affects DA, the Barlett and Gentile Cyberbullying Model (BGCM; Barlett & Gentile, 2012) outlines that self-anonymity encourages perpetrators to aggress online, positing that as individuals engage in DA, they perceive themselves to be more and more anonymous. The subsequent bolstering of these perceptions through successful DA acts then leads to increased frequency and positive attitudes toward cyberbullying; in turn, the continued development of positive cyberbullying attitudes reinforces perpetration. This learning-based model of DA illustrates how a seemingly harmless act can eventually lead to continued and even more hostile and malicious acts of DA. Critically, however, it is equally clear that not all individuals aggress in anonymous environments, suggesting that individual difference variables may serve as important moderators of these effects.

Personality

Consistent with the previous point, an independent strand of literature has simultaneously sought to identify the personality characteristics of perpetrators. Most studies have focused on the associations between DA and the Big Five. The Big Five includes openness (indexes imagination, intellect, and liberalism), conscientiousness (indexes orderliness, self-discipline, and thoughtfulness; akin to low impulsivity), extraversion (indexes friendliness, assertiveness, and excitement seeking), agreeableness (indexes cooperation, trust, and altruism), and neuroticism (indexes sadness, moodiness, and emotional instability) (Costa & McCrae, 1985; Goldberg, 1999). For example, Festl and Quandt (2013) found cyber aggressors to be more extraverted but less conscientious and less agreeable. Other studies have replicated these findings for low agreeableness and conscientiousness (Kokkinos et al., 2013, 2016; Zezulka & Seigfried-

Spellar, 2016), but findings for extraversion, neuroticism, and openness have been more mixed (Kokkinos et al., 2013, 2016; van Geel et al., 2017; Zezulka & Seigfried-Spellar, 2016). Overall, however, this field of research generally indicates that DA perpetrators tend to be impulsive, antagonistic, and extraverted.

In addition, a handful of studies also reported positive associations between the Dark Tetrad and DA (Craker & March, 2016; Kircaburun et al., 2018; Pabian et al., 2015). The Dark Tetrad (Buckels et al., 2013; Paulhus & Williams, 2002) consists of Machiavellianism (tendency to manipulate others), narcissism (tendency to feel entitled and superior to others), psychopathy (tendency to lack remorse and to engage in impulsive and egotistical behavior), and sadism (tendency to enjoy the suffering of others). It remains unclear, however, whether these associations reflect their common overlap with the Big Five. Indeed, the Dark Tetrad has also been robustly associated with low agreeableness (Jakobwitz & Egan, 2006). For example, psychopathy can be understood as a factor with extremely low scores on some facets of Agreeableness and Conscientiousness and high and low scores on some facets of Neuroticism and Extraversion, respectively (Miller et al., 2001). Other studies have replicated associations with low agreeableness and low conscientiousness (Jakobwitz & Egan, 2006; Jonason & Webster, 2010; Miller et al., 2010; Paulhus & Williams, 2002), perhaps not surprisingly given that these traits are also closely linked to antisocial behaviors (Jones et al., 2011; Miller et al., 2008; Miller & Lynam, 2001). Other studies have also found that Machiavellianism was positively associated with neuroticism, while narcissism was positively associated with extraversion (Vernon et al., 2008; Veselka et al., 2012).

Current Studies

How should we make sense of these two strands of literature, one of which points to external environmental features and the other to characteristics of the perpetrators? The General Aggression Model (GAM; Anderson & Bushman, 2002) provides a very useful framework for answering this question, particularly as it has been frequently employed in previous aggressive behavior research (Gilbert & Daffern, 2011; Gilbert et al., 2017; Vannucci et al., 2012), including a small number of studies examining DA (Kowalski et al., 2014). The GAM utilizes a "person in the situation" approach (Allen & Anderson, 2017, p. 7) to explain aggressive behaviors. The model has three phases; inputs, routes, and outcomes (refer to Figure 1). Inputs refer to personological and situational variables that affect aggressive behaviors. For example, personological variables like gender, personality, values, socioeconomic status, technology knowledge, and other characteristics could increase the likelihood that an individual becomes an online aggressor. Situational factors, such as parental involvement, school climate, and (perceived) anonymity, then interact with these personological variables to either limit or encourage DA. These inputs then enter the second phase (routes) of the GAM to influence affect, cognition, and arousal to establish one's present internal state. The internal state subsequently affects the third phase (outcomes) of the GAM to influence appraisal and decision processes that lead to either thoughtful or impulsive action. Depending on the latter, one may then refrain or engage in DA. That behavior then influences a social encounter and loops back to the inputs, repeating the process.

As outlined by the GAM (Anderson & Bushman, 2002), personal factors are thought to interact with situational factors to influence the internal states of individuals, which then influences whether or not a person will engage in DA. It would thus be essential to examine both

the specific contexts afforded by technology and the ways in which those contexts interact with personal predispositions – something not done by any other study to date. The current studies aim to do just this while constructively replicating and extending prior work with coding of actual real-world DA and in-vivo experimental assessment. Among several 'inputs,' the current studies focus on the roles of personality and technical self-anonymity, including their interaction, in DA to better understand and elucidate the foundational importance of these two factors. For completeness' sake, technical other-anonymity is also evaluated in the second empirical study.

STUDY 1

Digital aggression (DA), more commonly known as cyberbullying, is defined as the use of Information and Communication Technologies (ICT) to intentionally inflict harm on others (Burt & Alhabash, 2018). Examples include sending, posting, or sharing negative, harmful, false, or mean content about an individual, and sharing personal or private information about an individual to cause embarrassment or humiliation. Predictably, given this definition, DA has emerged as quite harmful, with victims experiencing higher rates of depression, anxiety, and suicidality, among other things (Hinduja & Patchin, 2010; Na et al., 2015; Schenk & Fremouw, 2012; Schenk et al., 2013; Selkie et al., 2015). Unfortunately, DA is also quite frequent. The Cyberbullying Research Center surveyed more than 24,000 middle and high school students in the US from 2007 to 2021 and found that 29% of individuals had been cyberbullied and 16% had cyberbullied others at some point in their lifetime (Patchin, 2022). A similar frequency was observed for college students and emerging adults. Of 439 college students. 38% reported knowing someone who experienced cyberbullying, 22% self-reported being cyberbullied, and 9% disclosed engaging in cyberbullying behaviors (MacDonald & Roberts-Pittman, 2010).

While DA takes place across platforms, social networking sites (SNSs) have recently drawn significant attention as the most accessible and pervasive playgrounds for DA (Chan et al., 2021). Over the last decade, the average daily time spent on social media worldwide has drastically increased from only 90 minutes in 2012 to 147 minutes in 2022 (Statista, 2022). YouTube, Facebook (rebranded as Meta), and Instagram continue to be the most commonly used SNS among US adults, with 81%, 69%, and 40%, respectively, reporting using these sites at least once (Pew Research Center, April 2021). In addition, about one-quarter of adults reported using Snapchat, Twitter, WhatsApp, and TikTok (Pew Research Center, April 2021). Extant research

has found that SNS users, especially youth, seek information, resources, social support, social capital, and intimacy online (Ellison et al., 2007; Nabi et al., 2013). Most importantly, youth use SNS to develop and maintain friendships (Ellison et al., 2007; Nabi et al., 2013), a tendency that was exacerbated during the COVID-19 pandemic, where individuals had to rely on SNS to seek and maintain social support while social distancing (Saud et al., 2020; Wong et al., 2021).

Recent research has strongly suggested that time spent on SNS and higher SNS addiction scores predict higher rates of DA (Giordano et al., 2021; Giumetti & Kowalski, 2022). The high prevalence rates of DA on SNS may be explained, in part, by the unique features of the online SNS environment (Chan et al., 2021; Sticca & Perren, 2013). First, individuals are especially fearful of public acts of DA, in which negative and humiliating comments are publicized for everyone to read (Sticca & Perren, 2013; Waasdorp & Bradshaw, 2015). Second, harassing comments can be shared and repeated with an infinite audience (i.e., go viral) such that targets of DA re-live the experience as part of an ongoing cycle (Kowalski & Limber, 2007; Slonje et al., 2013; Sticca & Perren, 2013), especially with the use of hashtags and tags (Kane et al., 2014). Third, online SNS platforms allow DA to take place at anytime and anywhere, with or without the presence of DA targets, allowing easier access to perpetrators (Cassidy et al., 2013; Chan et al., 2019). Indeed, even if the victims log off or deactivate their accounts, the perpetrators can continue to post harassing comments and distribute them to other users. Lastly, online platforms afford anonymity to DA perpetrators such that they can hide behind pseudonyms, which can cause significant stress and fear for DA targets. Sticca and Perren (2013) found that adolescents perceived anonymous DA as more humiliating and threatening than traditional anonymous and non-anonymous bullying (both DA and in-person). The loss of perceived control over the situation may heighten the distress caused by anonymity. Of note, these features of the online

setting may be exacerbated on SNS due to a constant disclosure of personal information, leading to increased opportunities for individuals to experience personalized attacks (Kane et al., 2014; Menesini et al., 2012).

Digital aggression and anonymity

One feature of SNS that may be important for understanding links with DA is anonymity, which can take different forms (Hayne and Rice, 1997). Technical anonymity describes removing personal identification, such as one's name. In contrast, social anonymity refers to the perception of self and/or others as anonymous due to the absence of social cues. For the latter, *perception* is the key such that the individual may not truly be anonymous, but they *perceive* themselves to be so. Another key element of anonymity is the relevant point of view. Self-anonymity (*you can't see me*) refers to the inability of others to identify the perpetrator, while other-anonymity (*I can't see you*) refers to the inability of the perpetrator to identify others (e.g., targets of DA).

Self-anonymity, particularly technical self-anonymity, has emerged as an important predictor of DA. For example, a qualitative study with 695 undergraduate students found that 64% of the bullies and 72% of the bully-victims reportedly pretended to be someone else to perpetrate DA (Arıcak, 2009). Quantitative studies report similar results (Barlett et al., 2017; Barlett, 2015a, 2015b; Barlett et al., 2016, 2019; Barlett & Chamberlin, 2017; Barlett & Gentile, 2012; Barlett & Helmstetter, 2018; Barlett & Kowalewski, 2019; Dong, 2019; Wright, 2013) with anonymity as a significant predictor of DA in many studies. Similarly, more abusive posts were identified on more anonymous platforms such as Yik Yak (Liu & Sui, 2017) and with more anonymous social media accounts (Mondal et al., 2018; Moore et al., 2012). Anonymity has also been found to promote the development of positive DA attitudes, which predicted subsequent

DA involvement (Barlett et al., 2017; Barlett, 2015b; Barlett et al., 2016, 2019; Barlett & Gentile, 2012; Barlett & Helmstetter, 2018; Barlett & Kowalewski, 2019; Ooi et al., 2019). This suggests that when individuals learn to engage in online aggression anonymously without any consequences, DA is more likely to occur in the future. In short, online anonymity appears to provide "safe spaces" for DA perpetrators by minimizing accountability and altering their sense of morality, leading to more frequent and possibly more severe DA.

Qualitative studies have identified three processes by which anonymity may contribute to DA. First, anonymity allows individuals to act differently online than offline (*deindividuation*) (Brandtzæg et al., 2009; Mishna et al., 2009). The process of deindividuation can lead perpetrators to believe that their online persona does not represent "who they really are" and continue to separate their online actions from their "in-person" selves (Suler, 2014). Second, participants discussed the "culture" of online platforms and how aggressive behaviors were normalized and accepted (*depersonalization*), encouraging some to become more aggressive (McInroy & Mishna, 2017; Reason & Boyd, 2016). For example, some players did not view hostility in online gaming platforms as aggression but as part of the gaming experience (McInroy & Mishna, 2017). Third, the minimization of accountability was evident such that perpetrators were protected, or at the very least, perceived themselves to be protected. They often exploited the fact that they did not have to face any consequences of their actions (Harrison, 2015; Samoh et al., 2019). The lack of accountability makes it easier for perpetrators to continue engaging in aggressive behaviors with little regard for the harm their actions cause others.

Despite the clear empirical and theoretical links between self-anonymity and DA, it is nevertheless the case that not all individuals aggress in anonymous environments. Indeed, the General Aggression Model (GAM; Anderson & Bushman, 2002) posits that personal factors

(e.g., personality) interact with situational factors (e.g., anonymity) to influence the internal states of individuals, affecting decision-making processes regarding whether or not a person will engage in a specific behavior. For example, an individual may appraise the situation as one where aggression is inappropriate and opt to put down their phone and walk away, whereas another individual may appraise the same situation as one in which aggression is appropriate and impulsively send several mean SNS to another individual. Additionally, engaging in these types of encounters over time may be linked with distal outcomes (e.g., decreased popularity among peers or removal of access to SNS), which can, in turn, affect individual and situational factors.

Digital aggression and personality

Other lines of research have focused on identifying the personality characteristics of the perpetrators to better understand the phenomenon of DA. Personality refers to "relatively consistent patterns of thinking, feeling, and behaving manifested by individuals" (Jones et al., 2011, p.329), most frequently conceptualized via the Big Five. The Big Five includes openness (indexes imagination, intellect, and liberalism), conscientiousness (indexes orderliness, self-discipline, and thoughtfulness; akin to low impulsivity), extraversion (indexes friendliness, assertiveness, and excitement seeking), agreeableness (indexes cooperation, trust, and altruism), and neuroticism (indexes sadness, moodiness, and emotional instability) (Costa & McCrae, 1985; Goldberg, 1999).

The Big Five appear to predict DA in at least two ways. First, the Big Five appear to predict individual behavior on social media more generally. For example, Gil de Zúñiga and colleagues (2017) examined data from over 20,000 respondents from 20 countries and found that extraversion, agreeableness, conscientiousness, and openness positively predicted the frequency of media use, while emotional stability negatively predicted the frequency of media use. In

addition to predicting increased digital footprints on SNS, the Big Five traits have been found to predict specific behaviors on SNS. Individuals with high conscientiousness, for example, tend to be cautious in managing their profiles by posting fewer pictures (Amichai-Hamburger & Vinitzky, 2010), expressing fewer "likes," and engaging in less group activity on social media (Kosinski et al., 2014). On the other hand, individuals with high neuroticism were more likely to post more photos on their profile and self-disclose more personal information (Amichai-Hamburger & Vinitzky, 2010; Seidman, 2013) while using more negative words in their posts (Schwartz et al., 2013). Furthermore, individuals with high agreeableness (Schwartz et al., 2013) and high openness (Amichai-Hamburger & Vinitzky, 2010) tend to be more expressive on their profile, and they also "like" more content on social media (Bachrach et al., 2012) with larger networks (Quercia et al., 2012).

Second, the Big Five have been shown to predict DA in particular. For example, Festl and Quandt (2013) found that cyber aggressors were high on extraversion, low on conscientiousness, and low on agreeableness. However, neither neuroticism nor openness appeared to be associated with DA. Subsequent studies have replicated associations for low agreeableness and low conscientiousness/high impulsivity (Kim et al., 2020; Kokkinos et al., 2013, 2016; Zezulka & Seigfried-Spellar, 2016). One proposed explanation was that impulsive adolescents find it difficult to restrain themselves when online bullying opportunities arise as they are less likely to consider the consequences of their actions. However, there are mixed findings for extraversion, neuroticism, and openness, with some studies reporting significant associations (Kim et al., 2020; Kokkinos et al., 2013; Zezulka & Seigfried-Spellar, 2016) and others finding no associations (Kokkinos et al., 2016; van Geel et al., 2017).

Current Study

As outlined by the General Aggression Model (GAM; Anderson & Bushman, 2002), personal factors (including personality) are thought to interact with situational factors (such as anonymity) to influence the internal states of individuals, affecting whether or not a person will engage in DA. Despite substantial evidence that personality and anonymity are independent predictors of DA, virtually no research has sought to investigate whether and how these two factors may interact in the prediction of DA. The current study sought to bridge this gap by coding real-world DA behaviors on Twitter by examining the roles of personality, technical self-anonymity on Twitter, and their interactions.

The choice of platform requires additional explanation. Twitter was chosen as the platform of interest in light of prior work indicating that DA among college students primarily occurs via text messaging (56.8%) or Twitter (45.5%) (Whittaker & Kowalski, 2015), and because we can readily assess DA perpetrated on Twitter (Calvin et al., 2015; Whittaker & Kowalski, 2015). Indeed, key features of Twitter's versatility as a communication platform have made it a popular tool for both users and researchers to make connections, share research and resources, and track areas of interest (Choi et al., 2014). Moreover, due to the public nature of tweets and its broad user base, Twitter is also a convenient platform for DA (Calvin et al., 2015; Whittaker & Kowalski, 2015). Although not the norm, using anonymous accounts is rather common on Twitter since that platform does not require the use of real names from its users (Peddinti et al., 2017; Twitter, 2021). Twitter users are also allowed to create multiple accounts to "express different parts of [their] identity" (Twitter, 2021), enabling more users to adopt anonymous accounts. For instance, Peddiniti and colleagues (2014) found a significant correlation between content sensitivity and anonymity after classifying Twitter accounts as

highly identifiable, identifiable, partially anonymous, and anonymous. That is, accounts that posted high levels of sensitive or controversial tweets (i.e., sexual orientation, religious and racial hatred, and guns) had a relatively large percentage of anonymous followers compared to identifiable followers.

Given all this, the following research questions were explored in the current paper (specific hypotheses for each research question can be found in Table 1):

1) What personality traits predict self-anonymity on Twitter?
2) What personality traits predict DA on Twitter?
3) Does DA on Twitter differ across individuals with anonymous versus identifiable Twitter bio accounts?
4) Which, if any, personality traits moderate the relationship between technical self-anonymity and DA on Twitter?

Methods

Participants

Participants consist of undergraduate students at a large Midwestern Research University who participated in exchange for course credit or extra credit. Only those with active Twitter accounts with at least one public tweet were eligible to participate. The university's Institutional Review Board approved the research protocol before data collection. We collected two samples, Sample 1 between October 2017 and April 2019 and Sample 2 between October 2021 and April 2022. Sample 1 included 478 participants (male = 152, female = 326) aged between 18 and 24 years (M = 19.0, SD = 1.1). These participants self-identified as White non-Hispanic (70.9%), Black non-Hispanic (15.1%), Asian or Pacific Rim (6.5%), Hispanic (4.4%), and other races/ethnicities (3.1%). Sample 2 included 462 participants (male = 129, female = 316,

transgender = 4, self-described = 8, and five preferred not to answer) aged between 18 and 29 years (M = 19.5, SD = 1.5). The participants self-identified as White non-Hispanic (64.1%), Black non-Hispanic (16.2%), Asian or Pacific Rim (10.0%), Hispanic (5.6%), and other races/ethnicities (4.1%). There was no overlap in participants across samples, as participants in the subject pool could not participate in the same experiment a second time.

Procedure

Sample 1 participants completed the study in person. Upon arrival, research assistants confirmed that participants were eligible for the study (i.e., they had a Twitter account) and gathered their Twitter usernames. The participants then completed a series of questionnaires via computer, including a demographic questionnaire and personality measures described below.

Sample 2 participants completed the study online through the Psychology Department's online subject pool (SONA). They were given access to the Study URL with a series of questionnaires, including a demographic questionnaire and personality measures described below. To ensure eligibility for participation, the first question asked whether participants had Twitter accounts. If they answered 'no,' the study was terminated.

Quantitative Measures

Anonymity Coding of Twitter accounts. The anonymity of Twitter account bios was coded by a team of four trained research assistants using a 3-point scale: 0 for no anonymity (can clearly identify whom the account belongs to), 1 for partial anonymity (can *guess* whom the account belongs to), and 2 for complete anonymity (no information on whom the account belongs to). Our Twitter anonymity coding scheme is presented in Table 2. Each member of the coding team rated each Twitter account, which was then averaged across all four raters. The intraclass correlation (ICC) was 0.88 for Sample 1 and 0.85 for Sample 2. Sample 1 had mean

anonymity ratings of 0.39 (*SD* = 0.48), a median of 0.25, with a range of 0 to 2. However, only four individuals had fully anonymous accounts, and the anonymity data were positively skewed (skew = 1.2). Given this, we also examined anonymity as a dichotomous variable with the averaged values between at or near zero (0 – 0.25) as identifiable or non-anonymous accounts (*n* = 305; 63.8% of the sample). All other averaged values were coded as partially to completely anonymous (*n* = 173; 36.2%).

Sample 2 had mean anonymity ratings of 0.70 (*SD* = 0.60), a median of 0.50, with a range of 0 to 2. Of 462 accounts, 12 accounts were not accessible to accurately code their anonymity. In contrast to Sample 1, 40 individuals had fully anonymous accounts, but the anonymity data were still positively skewed (skew = 0.9). The same recoding process was implemented for Sample 2, with 185 accounts (41.1% of the sample) coded as identifiable or non-anonymous accounts and 265 accounts (58.9% of the sample) coded as partially to completely anonymous. Of note, however, we also analyzed anonymity as a continuous variable.

Digital Aggression on Twitter. Participants' public tweets in the past year were mined to measure DA on Twitter. We capped the number of tweets mined at 200 for especially prolific tweeters. As the participants of Sample 1 consented to longitudinal mining of their Twitter accounts, their tweets were mined on two separate occasions: within one week of their initial date of participation (Time 1) and re-mined in July 2020 (Time 2), focusing again on the first 200 tweets over the prior year (7/22/2019 to 7/22/2020). Despite the eligibility criteria, there were a number of participants whose tweets could not be mined due to (1) no available tweets posted on their public Twitter accounts, (2) posted tweets only included foreign languages or broken links, and (3) Twitter accounts set to private. At Time 1, a total of 20 accounts were considered inactive; thus, the tweets of 458 participants were available for analysis. At Time 2, a

total of 47 accounts were inactive, so the tweets of 431 participants were available for analysis. For Sample 2, their Twitter data was mined only once, also appropriately a week after their initial date of participation. There were no inactive accounts; thus, the tweets of 462 participants were available for analysis.

As done by Burt et al. (2020), we submitted the tweets to the Linguistic Inquiry and Word Count (LIWC) software (Pennebaker et al., 2015), a dictionary-based text analysis program that rapidly analyzes words in psychologically meaningful categories (Tausczik & Pennebaker, 2010). Among these categories, we specifically focused on **anger** (e.g., hate, annoyed) and **swear** (e.g., d*mn, sh*t) words, both of which have been strongly linked to DA (Al-garadi et al., 2016; Hosseinmardi et al., 2014; Samghabadi et al., 2017). Using the LIWC outputs, we then computed the percentage of tweets in which the participant used one of the two-word categories and averaged these together to create the Antisocial Word Index (AWI; DeWall et al., 2011). To validate this approach, Burt et al. (2020) had a team of four trained research assistants code 173,588 tweets in 843 participants using a 6-point scale ranging from 0 (not aggressive at all) to 5 (very aggressive). Each member of the coding team rated each tweet, and the score was then averaged across raters. Average coder ratings of Twitter DA were highly correlated with the easy-to-obtain LIWC ratings (r = .64 for swearing, .63 for anger, and .68 for the AWI). The LIWC codes thus appear to accurately capture DA on Twitter.

Personality. Participants in both samples completed the 50-item International Personality Item Pool-Five Factor Model (IPIP-FFM; Goldberg, 1999), a measure of the Big Five personality trait domains. Extraversion (α = .88 in Sample 1 and α = .89 in Sample 2) indexes friendliness, gregariousness, assertiveness, and exciting seeking. Agreeableness (α = .77 in Sample 1 and α = .80 in Sample 2) indexes cooperation, sympathy, and altruism.

Conscientiousness (α = .82 in Sample 1 and α = .81 in Sample 2) indexes orderliness, self-discipline, and cautiousness. Emotional stability (α = .84 in Sample 1 and α = .85 in Sample 2) is opposite to neuroticism and indexes calmness, composure, and unflappability. Intellect/Imagination (α = .77 in Sample 1 and α = .81 in Sample 2) assesses imagination, intellect, and liberalism. Each domain has 10 items, which are summed so that a high score indicates a high level of the trait.

Analyses

We employed multiple regression analyses to examine our first two research questions (i.e., What personality traits predict self-anonymity on Twitter?; What personality traits predict DA on Twitter?). To examine the third research question (i.e., Does DA on Twitter differ across individuals with identifiable versus anonymous Twitter bios?), we evaluated the differences in DA via *t*-tests. Finally, a four-step hierarchical multiple regression was conducted with Twitter DA as the dependent variable to examine the fourth research question (i.e., Which, if any, personality traits moderate the relationship between technical self-anonymity and DA on Twitter?). Prior to the analyses, personality traits were mean-centered to avoid multi-collinearity and clarify the regression coefficients (Irwin & McClelland, 2001). Demographic variables (age, gender, and race/ethnicity) were entered at step one, personality traits were entered at step two, followed by the addition of anonymity at step three, and statistical interaction terms between anonymity and personality at the final fourth step. As recommended by Aiken and West (1991), we plotted significant interaction effects at low (1 *SD* below the mean) and high (1 *SD* above the mean) levels of each personality trait. In addition, an inspection of Q-Q Plots revealed that standardized regression residuals were normally distributed with the values for skewness

between ± 2 and kurtosis between ± 7, which are considered acceptable to prove normal distribution (Hair et al., 2010) for both Sample 1 and Sample 2.

Results

Descriptive statistics and zero-order correlations between measured variables are provided in Table 3a and Table 3b for Sample 1 and Sample 2, respectively. For Sample 1, the results indicated that DA was relatively stable over time, such that DA at Time 1 was significantly correlated with DA at Time 2 (r = .47). Anonymity was also moderately correlated with DA at both Time 1 (continuous anonymity: r = .32; recoded anonymity: r = .34) and Time 2 (continuous anonymity: r = .21; recoded anonymity: r = .22). Low extraversion, low emotional stability, low agreeableness, and low conscientiousness were modestly associated with DA at Time 1 (rs ranged from – .09 to – .15). However, only one of these associations (low emotional stability) persisted over time at Time 2 (r = – .11). Interestingly, self-anonymity was also modestly associated with low extraversion, low emotional stability, and low agreeableness (rs ranged from – .10 to – .16). In Sample 2, however, self-anonymity was not significantly correlated with DA, but was modestly associated with low extraversion, low agreeableness, and low conscientiousness (rs ranged from – .11 to – .19). There were also no significant correlations between DA and personality traits.

RQ1. Personality Predictors of Twitter Anonymity.

Sample 1. A multiple regression was run to predict continuously assessed Twitter anonymity from personality traits (Table 4a); $F(5, 472) = 5.27$, $p < .001$, $R^2 = .05$. As we hypothesized, individuals with low extraversion (β = –0.15, p <.01), low agreeableness (β = –0.13, p <.01), and high intellect/imagination (β = 0.09, p = .05) had more anonymous Twitter account bios. When we examined the dichotomous index of anonymity using logistic regression,

results were similar, but not identical (Table 4b). Low extraversion, low agreeableness, and high intellect/imagination were associated with an increased likelihood of using anonymous Twitter account bios. The logistic regression model was statistically significant, $\chi^2(8) = 26.67$, $p < .001$. The model explained 6.7% (Nagelkerke R^2) of the variance in Twitter anonymity and correctly classified 63.8% of cases.

Sample 2. A multiple regression was run to predict continuously assessed Twitter anonymity from personality traits (Table 4c); $F(5, 444) = 6.76$, $p < .001$, $R^2 = .07$. Similar to sample 1, individuals with low extraversion ($\beta = –0.19$, $p < .01$) and high intellect/imagination ($\beta = 0.14$, $p < .01$) had more anonymous Twitter account bios. Unlike sample 1, however, we did not observe an association with agreeableness, but did observe an association with low conscientiousness ($\beta = –0.14$, $p < .01$). Similarly, the logistic regression model (Table 4d) was not statistically significant, $\chi^2(8) = 9.91$, $p = .27$, although the model explained 6.3% (Nagelkerke R^2) of the variance in Twitter anonymity and correctly classified 62.0% of cases. Low extraversion and low conscientiousness were associated with an increased likelihood of using anonymous Twitter account bios.

RQ2. Personality Predictors of DA on Twitter.

Sample 1. After controlling for age, gender, and race/ethnicity, we found that low extraversion ($\beta = – 0.11$, $p < .05$), low agreeableness ($\beta = – 0.12$, $p < .05$), and high intellect/imagination ($\beta = 0.13$, $p < .05$) predicted higher DA on Twitter at Time 1. However, only high intellect/imagination ($\beta = 0.14$, $p < .01$) persisted over time and continued to predict higher DA on Twitter at Time 2.

Sample 2. After controlling for age, gender, and race/ethnicity, we found that only high intellect/imagination ($\beta = 0.14$, $p < .01$) predicted higher DA on Twitter.

RQ3. Differences in DA across Anonymous and Identifiable Twitter Account Bios.

Sample 1. An independent *t*-test was run to examine differences in DA across individuals with identified and anonymous Twitter account bios. As we hypothesized, individuals with anonymous Twitter accounts (10.79 $\pm$ 6.61 at Time 1; 11.70 $\pm$ 7.77 at Time 2) used more antisocial words than those with identifiable Twitter accounts (6.78 $\pm$ 4.41 at Time 1; 8.27 $\pm$ 7.07 at Time 2) at Time 1 ($t(249.87) = -6.99$, $p < .001$; Table 5a) and at Time 2 ($t(300.60) = -4.54$, $p < .001$).

Sample 2. An independent *t*-test was run to examine differences in DA for individuals with identified and anonymous Twitter account bios. There were no significant differences between those with anonymous (8.20 $\pm$ 7.67) and identifiable (7.10 $\pm$ 7.27) Twitter accounts (Table 5b).

RQ4. Associations between Anonymity and Personality on Twitter DA.

Sample 1 at Time 1. For our final objective, we examined whether personality traits moderated the relationship between technical self-anonymity and DA on Twitter. The moderated regression results are reported in Table 6a. As seen there, recoded anonymity ($\beta = 0.30$, $p < .01$) and continuous anonymity ($\beta = 0.29$, $p < .01$) uniquely predicted DA, whereas the personality traits did not. That said, we did observe significant interactions between personality and recoded anonymity for emotional stability ($\beta = -0.12$, $p < .05$), agreeableness ($\beta = -0.18$, $p < .01$), and intellect ($\beta = 0.11$, $p < .05$). Specifically, individuals low in emotional stability (Figure 2a) and low in agreeableness (Figure 2b) used more antisocial words when they had anonymous accounts, but not when they had identifiable accounts. In contrast, individuals high in intellect/imagination (Figure 2c) used more antisocial words when they had anonymous accounts but not when they had identifiable accounts, although this finding was specific to the recoded

dichotomous indicator of anonymity and was not significant for continuously-assessed anonymity (although it remained positively signed). However, similar interaction effects were reported for individuals low in emotional stability (β = – 0.12, p < .05) and low agreeableness (β = – 0.17, p < .01). We also observed a significant interaction between continuous anonymity and high conscientiousness (β = 0.12, p < .05). Specifically, individuals high in conscientiousness used more antisocial words when they had more anonymous accounts, but not when they had more identifiable accounts.

When evaluating DA over time, we found that both recoded anonymity (β = 0.18, p <.001) and continuous anonymity (β = 0.19, p <.001) uniquely predicted DA (see Table 6b). Low conscientiousness (β = –0.12, p <.05) uniquely predicted DA when anonymity was dichotomized but not when continuous (β = –0.10, *ns*). Interaction effects were again observed for intellect/imagination (β = 0.12, p < .05) such that those high in this trait used more antisocial words when they had anonymous accounts but not when they had identifiable accounts (Figure 3). As above, however, this effect was not significant when examining continuously assessed anonymity (although it again remained positively signed).

Sample 2. Moderated regression results in sample 2 are reported in Table 6c. As seen there, gender was the only demographic variable that uniquely predicted DA (β = 0.14, p <.001 with recoded anonymity in the model; β = 0.15, p <.001 with continuous anonymity in the model). Similar to Sample 1, interaction effects were only observed for intellect/imagination (β = 0.18, p < .05) such that those high in this trait used more antisocial words when they had anonymous accounts but not when they had identifiable accounts (Figure 4), but only when anonymity was assessed dichotomously.

Discussion

The first goal of the current study was to examine personality predictors of anonymous Twitter bio accounts. As hypothesized, we found that individuals low on extraversion and high intellect/imagination were more likely to create and use anonymous accounts in both samples. Such findings are consistent with prior work indicating that introverted individuals express themselves more openly in anonymous environments (Amichai-Hamburger & Etgar, 2019; Zhang et al., 2022), and thus may be more likely to prefer online anonymity when interacting with others. Similarly, Hughes and colleagues (2012) found that Twitter users are typically higher on intellect/imagination in that they access Twitter for informational purposes, such as academic or political information. At the same time, Twitter users focus less on 'who you are' and your extant social circles, compared to Facebook, and more on what you think and wish to say (Huberman et al., 2008), which likely lends itself to increased online anonymity.

The second goal of the current study was to examine personality predictors of DA on Twitter. We focus here on those findings that replicated across samples. High intellect/imagination was found to generally predict DA on Twitter in our regression analyses, a finding that persisted across samples as well as over time. These findings was in line with our hypothesis, and may reflect the unique features of Twitter and, specifically, the need to be articulate and verbally adroit at gaining followers (Kim et al., 2020). In other words, DA on Twitter likely involves a degree of creativity and expressiveness related to the intellect/imagination factor. Low agreeableness also predicted higher DA in both samples, but did so only cross-sectionally. The findings for agreeableness did not persist to Time 2. Low extraversion predicted higher DA on Twitter but did so only at Time 1 in Sample 1, and thus is not discussed further.

As a third goal for our study, we were interested in whether DA on Twitter differed for individuals with anonymous and identifiable Twitter bio accounts. For Sample 1, individuals with anonymous Twitter accounts engaged in higher DA on Twitter than those with identifiable Twitter accounts. These findings are consistent with previous studies that utilized publicly-available social media data to examine the role of anonymity in DA and found that posts were more abusive and aggressive when users were more anonymous (Liu & Sui, 2017; Mondal et al., 2018; Moore et al., 2012). For instance, users with anonymous Twitter accounts (Mondal et al., 2018) and posts without forum identifiers (Moore et al., 2012) endorsed more hate, attacks, and aggressive posts. Critically, however, these findings did not replicate with Sample 2. These results suggest that while anonymity may lead to increased levels of deindividuation, lower self-awareness, and strengthened group conformity (Postmes et al., 1998; Spears & Lea, 1992) to increase instances of DA, it may not directly increase DA on Twitter.

Our final and central objective was to evaluate whether personality moderated the relationship between technical self-anonymity and DA on Twitter. Consistent with our hypotheses, results revealed that individuals high in intellect/imagination used more antisocial words when they had anonymous accounts but not identifiable ones. Moreover, this finding persisted both over time and to a second sample. These positive findings may reflect the fact that individuals high on intellect/imagination are more open to experience and tend to be more expressive in their interpersonal interactions (Zezulka & Seigfried-Spellar, 2016). They also report more imagination, curiosity, artistic talent, and diversity in interests (Costa & McCrae, 1985; John & Srivastava, 1999). They are also more likely to engage in blogging (Guadagno et al., 2008), use favorable features of new technologies (Kircaburun et al., 2020), and follow more groups on social media (Bachrach et al., 2012). These online behaviors may be especially

heightened on Twitter as the use of Twitter to socialize has been related to higher openness (Hughes et al., 2012). Since individuals high on intellect/imagination could be more expressive and active on Twitter, their increased online activities may expose them to more risks, including DA (Barlett et al., 2019; Park et al., 2014). With increased DA opportunities, they may normalize and perceive DA as less risky, especially when their identity can be concealed. Then each successful act of DA is positively reinforced, leading to more positive attitudes toward DA and increased frequency (Barlett & Gentile, 2012).

We also observed significant interactions for low emotional stability and low agreeableness in Sample 1 at Time 1, such that individuals low in these two traits used more antisocial words when they had anonymous accounts than those with identifiable accounts. Low agreeableness and low emotional stability (akin to neuroticism) indicate higher levels of irritability and hostile rumination, which are thought to foster moral disengagement and engagement in aggressive behaviors (Caprara et al., 2013). These results suggest that these effects may sometimes be more pronounced under anonymous conditions, although future work is needed to replicate these findings before any firm conclusions could be drawn. We also observed an unexpected significant interaction between continuous anonymity and high conscientiousness in Sample 1 at Time 1, such that those high in conscientiousness used more antisocial words when they had anonymous accounts but not when they had identifiable accounts. This was unexpected as less conscientious individuals tend to find it difficult to restrain themselves when opportunities arise to engage in DA without consideration of the possible consequences of their actions (Kokkinos et al., 2014; You & Lim, 2016). However, neither of these interactions replicated or persisted over time, so their significance remains unclear and needs further examination.

Strengths and Limitations

A few important limitations of the current study should be noted. First, we relied on a convenience sample of college students in the community. As such, our findings do not inform the understanding of DA of the general population, other age ranges, geographical regions, and those who do not have access to higher education. We also note that relatively few participants had fully anonymous Twitter accounts, with only four Sample 1 individuals (0.8% of the sample) and 40 Sample 2 individuals (8.9% of the sample) having fully anonymous accounts. This finding is consistent with a previous study that found only 6% of 50,173 Twitter accounts were fully anonymous (Cappos et al., 2017), and that at least one-quarter of Twitter users were partially-to-completely anonymous. The differences between the two samples may also be explained by a trend in which individuals, especially the Gen Z population, are increasingly opting for online anonymity over personal branding as more people become aware of the risks that come with sharing too much personal information (Bakker, 2022). Since Sample 1 included college students between 2017 and 2019 and Sample 2 between 2021 and 2022, the increased number of individuals who chose to remain fully anonymous may be part of this trend. Relatedly, because even stable behaviors can vary across days and months depending on personal, community, and world events. As such, it should be noted that tweets from Sample 1 were collected prior to the COVID-19 pandemic, while tweets from Sample 2 were collected during the pandemic. Given that individuals have expanded their use of social media to remain in contact with others while social distancing (Karmakar & Das, 2021), it is possible that the COVID-19 pandemic may have affected individuals' behaviors on Twitter.

Next, because our data were collected at only one to two timepoints, all of which were after the creation of their Twitter account, the direction of effect is ambiguous. It is thus unclear

whether those high in intellect/imagination created anonymous profiles to intentionally facilitate their engagement in DA, whether their use of anonymous Twitter profiles promoted DA after the fact, or whether both processes are at play (prior DA leads to increased use of self-anonymity, which increases DA, etc.). Prior theory presupposes the bidirectional process, but future studies should seek to clarify the direction of the association. It will also be important for future studies to increase our knowledge of how personality is relevant to behavior on Twitter, such as how people design and change their Twitter profiles over time. Such a longitudinal study would enable us to increase our understanding of the long-term interaction between personality and the Twitter dynamic.

Next, it is important to note that our results are specific to DA on Twitter and cannot be generalized to other social networking sites. Future research should analyze the association between anonymity and DA on other popular digital platforms, such as Snapchat and Discord. Finally, because our study specifically examined the role of technical self-anonymity in DA, it is unclear whether the findings generalize to other types of anonymity. This is important since it has been argued that perceived anonymity may be more important than actual or technical anonymity (Barlett & Gentile, 2012; Mishna et al., 2009); however, this remains more speculative than empirically tested.

Despite these limitations, the current study had several strengths. First, it is one of the first studies to our knowledge to examine whether and how the effects of anonymity on DA perpetration vary with the presence of specific personality traits. The current study also further illuminated associations between personality, anonymity, and DA by analyzing real-world behaviors on Twitter instead of relying on self-reported questionnaires. While users may not create anonymous Twitter accounts for the sole purpose of engaging in DA, anonymity may

offer a space free from societal norms, making them feel less restrained to engage in online behaviors that they may not exhibit in-person. Another strength of our study was that our sample consisted of racially and ethnically diverse college students, filling an important gap in the cyberbullying literature as most extant studies continue to focus on the K-12 school environment.

Implications

The current study has key implications for the field's understanding of DA. Our results indicate a clear need to identify additional technological factors that help us understand online behavior. While there was a trend of increased DA when using anonymous compared to identifiable profiles on Twitter, the effect sizes were small and inconsistently significant. Such findings suggest that while technical self-anonymity may play a role in DA, it may be somewhat less important in some online contexts than previously assumed. Future research should seek to identify other explanatory variables, such as motivation, frequency of technology use, and parental control and filtering that might be influential. Examining a broader base of variables can help improve our understanding of personal factors, situational factors, and their interactive effects on online behaviors.

Second, results revealed that individuals high in intellect/imagination engaged in higher levels of DA on Twitter when they had anonymous accounts than when they had identifiable accounts. How do we interpret this finding? The GAM provides a very useful framework. In the real world, individuals high on openness may engage in more effortful reappraisal processes in response to a negative social interaction, seeking out alternative outcomes (Anderson & Bushman, 2002) instead of engaging in an aggressive response. However, online anonymity may free these individuals from both social cues and moral transgressions (Patchin & Hinduja, 2006), hindering the reappraisal processes. As such, they are more likely to appraise the situation as one

in which aggression is appropriate (or more convenient) and engage in DA when using their anonymous Twitter accounts.

Finally, and building on the above, the current replicated findings of a person-by-environment interaction provide additional evidence in favor of the GAM as a model for understanding the development of aggression. The GAM argues that personal factors like personality may interact with situational factors like anonymity to influence the internal states of individuals, affecting decision-making processes regarding whether or not a person will engage in a specific behavior. For example, an online user may assess a situation and deem aggression appropriate and necessary, whereas another user may appraise the same situation as not demanding an aggressive response, so they choose to disengage and walk away. These differential behaviors and experiences accumulate over time and can loop back into the model to influence individual and situational factors. The current findings for intellect-imagination strongly support this model, suggesting that individuals high on the intellect-imagination personality trait engage in more DA on Twitter when they keep themselves anonymous than when they are identifiable. Importantly, however, the other personality traits did not align with the GAM in this study. These differences may reflect the possibility that individuals with certain personality traits (e.g., low agreeableness) may hold a unique baseline motivation to engage in DA, irrespective of anonymity, as it has been argued that DA often stems from the immediate appraisal that elicits impulsive actions (Savage & Tokunaga, 2017). Alternately, other situational factors (as mentioned above) may interact with these personality traits to provide additional pathways within the GAM to explain DA. Future research should seek to examine other person-by-environment interactions that may perpetuate DA. In doing so, it would be important to consider all aspects and processes of the GAM, whenever possible, to increase our

understanding of DA to better design prevention and intervention efforts aimed at targeting the individual in the situation (Mason, 2008).

STUDY 2

Digital aggression (DA) is defined as the use of Information and Communication Technologies (ICT) to intentionally inflict harm on others (Burt & Alhabash, 2018). Related terms like cyberbullying, online aggression, and electronic bullying (Mehari et al., 2014) include different forms and types of DA, such as harassment (i.e., repeatedly sending inappropriate or hurtful messages), flaming (i.e., using insults and profanity often as a reaction to provocation), exclusion (i.e., blocking an individual from contact), stalking (i.e., following an individual and sending targeted messages), outing (i.e., sharing an individual's secret or personal information without consent), and non-consensual sexting (i.e., distributing of nude pictures of an individual without consent) (Kowalski et al., 2014; Vandebosch & Van Cleemput, 2008). We will make use of a broad definition here, defining DA to encompass any aggressive acts committed online and/or electronically using any form of ICTs.

DA is very common, especially among college students and young adults. Indeed, previous studies report victimization rates between 22-55% and perpetration rates between 8-22% (Dilmaç, 2009; Francisco et al., 2015; MacDonald & Roberts-Pittman, 2010). A recent survey by the Pew Research Center similarly indicated that roughly two-thirds of adults under age 30 (64%) reported experiencing DA, making young adults the only adult age group in which a majority reported DA (Pew Research Center, January 2021). These high rates of DA are largely a function of their heavy and unsupervised use of technology, increased disclosure of personal lives on social media, experience seeking, and formation of tight cliques on social media (Jones & Scott, 2012, as cited in Kokkinos et al., 2014, p. 204). Despite this, most DA research to date has focused on school-aged children and adolescents, an approach that is

consistent with that in the broader bullying literature (Campbell, 2005; Espelage & Swearer, 2003; Raskauskas & Stoltz, 2007; Ybarra & Mitchell, 2004).

Unfortunately, DA appears to be just as harmful for young adults as it is for younger children and adolescents. College students who experienced DA victimization reported lower self-esteem (Na et al., 2015; Patchin & Hinduja, 2010) and higher levels of depression, anxiety, phobic anxiety, and paranoia (Schenk & Fremouw, 2012; Schenk et al., 2013; Selkie et al., 2015). Moreover, as the frequency of DA victimization increased, college students endorsed significantly more suicidal behaviors, including suicidal planning, attempts, and ideation (Na et al., 2015). This is consistent with a study of adolescents that also reported higher rates of suicidal ideation and attempts when experiencing DA victimization (Hinduja & Patchin, 2010). These adverse psychological outcomes are especially concerning given the proliferation of communication technologies and their ritualized daily use. For example, 97.5% of young adults in 2016 reported regularly using at least one social media site (Villanti et al., 2017) where DA most frequently occurs (Whittaker & Kowalski, 2015). In short, young adults are at considerable and sustained risk for DA and its negative mental health consequences. Efforts to understand the origins of DA among young adults are thus a critical public health objective.

Available models of the origins of physical aggression may provide a useful starting point for this work. The General Aggression Model (GAM; Allen & Anderson, 2017; Anderson & Bushman, 2002), for example, focuses on the complex interaction of personal and contextual factors to understand the origins of aggression (Kowalski et al., 2014). The GAM adopts a "person in the situation" approach to explain aggression (Allen & Anderson, 2017, p. 7) in various phases. First, personological and situational variables (inputs) influence affect, cognition, and arousal to establish one's present internal state (routes). The internal state is then linked with

appraisal and decision processes (outcomes) that lead to thoughtful or impulsive actions. The resulting action then influences the social encounter and feeds into the inputs, repeating the process as a feedback loop. Furthermore, the GAM proposes that each feedback loop can act as a learning trial that also influences distal causes and processes, which can, in turn, affect individual and situational factors (Allen & Anderson, 2017; Anderson & Bushman, 2002; Kowalski et al., 2014).

While the GAM has provided a practical and highly cited (Gilbert & Daffern, 2011; Gilbert et al., 2017) theoretical framework for research on other violent and aggressive behaviors, its application to DA has only recently gained traction (Kokkinos and Antoniadou, 2019). Applying the GAM to DA (Kowalski et al., 2014), we would hypothesize that DA perpetration starts with the person (e.g., gender, age, personality, etc.) and situational factors (e.g., school climate, provocation, technological availability and use, perceived anonymity, etc.), which then affect the present internal state of the individual. These internal states would then affect their appraisal and decision process (e.g., negative affect or heightened arousal would increase the likelihood that an aggressive online response was deemed appropriate). Enacted DA behavior would then feedback into the person and situational inputs, reinforcing aggressive tendencies and increasing the likelihood of future DA behaviors.

Although the literature is still small, available studies support some elements of the GAM as an explanation for the development of DA. First, several studies have found evidence of associations between DA and Big Five personality traits. The Big Five includes openness (indexes imagination, intellect, and liberalism), conscientiousness (indexes orderliness, self-discipline, and thoughtfulness; akin to low impulsivity), extraversion (indexes friendliness, assertiveness, and excitement seeking), agreeableness (indexes cooperation, trust, and altruism),

and neuroticism (indexes sadness, moodiness, and emotional instability) (Costa & McCrae, 1985; Goldberg, 1999). For example, Festl and Quandt (2013) found those high in cyber aggression to be more extraverted but less conscientious (or more impulsive) and less agreeable. Other studies have replicated these findings for low agreeableness and low conscientiousness/high impulsivity (Kokkinos et al., 2013, 2016; Zezulka & Seigfried-Spellar, 2016). However, there are mixed findings for extraversion, neuroticism, and openness, with some studies reporting significant associations (Kokkinos et al., 2013; Zezulka & Seigfried-Spellar, 2016) and others finding no associations (Kokkinos et al., 2016; van Geel et al., 2017).

In addition, a handful of studies also reported positive associations between the Dark Tetrad and DA (Craker & March, 2016; Kircaburun et al., 2018; Pabian et al., 2015). The Dark Tetrad (Buckels et al., 2013; Paulhus & Williams, 2002) consists of Machiavellianism (tendency to manipulate others), narcissism (tendency to feel entitled and superior to others), psychopathy (tendency to lack remorse and to engage in impulsive and egotistical behavior), and sadism (tendency to enjoy the suffering of others). It remains unclear, however, whether these associations reflect their common overlap with the Big Five. Indeed, the Dark Tetrad has also been robustly associated with low agreeableness (Jakobwitz & Egan, 2006). For example, psychopathy can be understood as a factor with extremely low scores on some facets of Agreeableness and Conscientiousness and high and low scores on some facets of Neuroticism and Extraversion, respectively (Miller et al., 2001). Other studies have replicated associations with low agreeableness and low conscientiousness (Jakobwitz & Egan, 2006; Jonason & Webster, 2010; Miller et al., 2010; Paulhus & Williams, 2002), perhaps not surprisingly given that these traits are also closely linked to antisocial behaviors (Jones et al., 2011; Miller et al., 2008; Miller & Lynam, 2001). More importantly, traditional in-person bullies are more likely to

possess dominant, impulsive, and "dark" personality characteristics to boast their power over their victims (Baughman et al., 2012; van Geel et al., 2017).

Critically, however, it remains as yet unknown whether these associations with DA are accentuated (or dampened) by features of the online or virtual environment as predicted by the GAM, an especially important consideration in light of the fact that person-environment interactions are thought to be a driving force behind the widespread effects of personality on behavior (Hardie, 2020; Rutter et al., 1997). One potentially important environment in the online context is the level of anonymity. Indeed, there is compelling data to indicate that DA is more likely when perpetrators perceive themselves as anonymous (Barlett et al., 2017; Barlett & Gentile, 2012; Harrison, 2015; Ooi et al., 2019; Reason et al., 2016). For example, a qualitative study by Harrison (2015) concluded that online anonymity made it difficult for individuals to recognize the implications of their online behaviors due to diminished accountability, with one male participant remarking, "[perpetrators] think [DA] is a victimless crime, nothing is going to happen to anyone, they don't see people getting hurt, so why not" (p. 279). Similarly, anonymity was an important predictor of the development of positive cyberbullying attitudes, which predicted later DA (Barlett et al., 2017). The role of anonymity is also recognized by the broader population, with prior qualitative work pointing to this very issue, "[perpetrators] feel like since nobody knows who they are, that they could say pretty much anything" (Reason et al., 2016, p. 2340).

There are two theories that seek to explain how anonymity in particular may encourage perpetrators to aggress online: the Social Identity Model of Deindividuation Effects (SIDE; Spears & Lea, 1992) theory and the Barlett and Gentile Cyberbullying Model (BGCM; Barlett & Gentile, 2012). The SIDE theory is centered around classic deindividuation theory; it posits that

anonymity leads to increased levels of deindividuation, lower self-awareness, and strengthened group conformity (Postmes et al., 1998; Spears & Lea, 1992). More recently, Barlett and Gentile (2012) proposed the learning-based BGCM to illuminate specific psychological processes that may underlie DA. The latter model postulates that as individuals engage in DA, they perceive themselves as more anonymous and believe that their physicality (i.e., height, weight, etc.) is irrelevant online (in contrast to the real world). These perceptions are bolstered by successful DA acts, eventually leading to increased frequency and more positive attitudes toward cyberbullying; in turn, the continued development of positive cyberbullying attitudes reinforces perpetration (Barlett & Gentile, 2012). This learning-based model of DA illustrates how a seemingly harmless act can eventually lead to more frequent, hostile, and malicious acts of DA.

Critically, however, neither anonymity model considers the possibility that some individuals may be more susceptible to the DA-promoting effects of anonymity given their preexisting personality traits. The SIDE and the BGCM both assume that the processes in question apply more or less equally to everyone. This is surprising, as it is clear that not all individuals aggress in anonymous environments, suggesting that individual difference variables may be important moderators of these effects. Indeed, as outlined by the GAM (Allen & Anderson, 2017; Anderson & Bushman, 2002), personal factors (including personality) are thought to interact with situational factors (such as anonymity) to influence the internal states of individuals, affecting whether a person will engage in DA.

There is thus a clear need for research that illuminates whether and how personality traits might interact with online anonymity to increase DA. In conducting this work, however, it would be critically important to precisely define anonymity. Indeed, relatively few studies have assessed anonymity in a rigorous way, relying instead upon self-reported items/questionnaires

measuring perceptions and/or beliefs about anonymity (akin to social anonymity) with sample items such as “I am confident that I would not be caught if I engaged in mean online behaviors” and “I am less likely to send mean e-mails or text messages if my name can be identified” (Barlett & Gentile, 2012; Wright, 2013). This is a less-than-ideal approach to assessing anonymity since it is well-known that self-reports can be plagued by issues such as social desirability bias and limited comprehension (Althubaiti, 2016; Pellegrini & Bartini, 2000). What is more, these self-report measures of anonymity have not been adequately validated (Wright, 2013), and thus it is unclear whether they were truly assessing anonymity or some other construct.

There are a few studies utilizing publicly-available social media data to examine the role of technical anonymity in DA. While these studies generally found online posts more abusive and aggressive when users were more anonymous (Liu & Sui, 2017; Mondal et al., 2018; Moore et al., 2012), results were inconsistent across studies (Rost et al., 2016). To our knowledge, only one study (Fox et al., 2015) has experimentally manipulated anonymity, randomly assigning 172 participants to either an anonymous or identified Twitter condition. They found that anonymous participants expressed significantly more sexist attitudes than those identified participants. More recently, Kim and Burt (2023) examined the role of technical anonymity on DA. In support of previous studies, college students with anonymous Twitter accounts were found to engage in higher DA than those with identifiable accounts.

Finally, the anonymity ‘reference point’ remains vague in nearly all studies on this topic. This is surprising since anonymity can take multiple forms: self-anonymity (*you can't see me*) or other-anonymity (*I can't see you*). Extant theory, as reviewed above, has focused implicitly on self-anonymity, but has yet to distinguish or consider self-versus other-anonymity. Moreover, to

our knowledge, the association between online aggression and technical self-anonymity has only been evaluated in one experimental study (Fox et al., 2015). This work did suggest that self-anonymity causes increases in sexist posts, but the sample size was small and did not examine the broader construct of DA. In sum, there is thus a clear need for research to be both more precise in their operationalization of anonymity, and to make use of an experimental design that allows for stronger causal inferences.

The Current Study

In the current study, I sought to advance our understanding of the origins of DA by experimentally testing key elements of the GAM in relation to DA using an experimental paradigm focused explicitly on technical self- and other- anonymity. I specifically examined the effects of personality (person factor) and technical anonymity (situational factor) and their interaction on DA. In doing so, I randomly assigned anonymous conditions to allow for stronger causal inferences and did so from both reference points: self-anonymity (*you can't see me*) or other-anonymity (*I can't see you*). I also made use of a recently developed in-vivo paradigm for assessing DA, the TAP-Chat. As described in detail below, the TAP-Chat is an adapted 'chat' version of the Taylor Aggression Paradigm (Taylor, 1967), that was designed to more closely resemble a social gaming format by providing participants with a chat function to communicate with their (fictitious) co-player.

Using this design, I answered the following research questions (specific hypotheses for each research question can be found in Table 1):

1) What personality traits predict DA on the TAP-Chat?
2) Does DA differ across experimentally manipulated anonymity conditions?

3) Do experimentally manipulated anonymity conditions moderate the relationship between personality traits and DA?

Methods

Participants

Participants were recruited directly by the company Qualtrics. Working with multiple research marketing firms worldwide, Qualtrics recruits participants from many panel sources, nationally seeking out individuals meeting study inclusion criteria. Our study inclusion criteria included young adults aged between 18-23 with a preferred gender breakdown of no more than 50% female or 50% male, with an approximately 5% cap on individuals who identify as transgender and those who prefer to self-identify. As Qualtrics recruits from various marketing panels, participants were provided with a wide array of potential compensation (i.e., vouchers, travel points, and cash) for participation in the study. Participants were informed of the compensation they may receive beforehand by the marketing agencies. Exact information surrounding compensation is unavailable to the researchers as this information is kept private by Qualtrics. All recruitment, participant contact, data collection, and compensation were handled by Qualtrics directly and did not involve the researchers, who only received the final anonymous data after it was collected and screened by Qualtrics.

The university's Institutional Review Board approved the research protocol before data collection. The sample collection began in December 2021 and ended in March 2022 when Qualtrics closed the project due to the slow participation rate. As a result, instead of the proposed 880 participants, our final sample included 553 participants (male = 123, female = 401, transgender = 15, self-described = 11, and three preferred not to answer) aged between 18 and 23 years (M = 20.7, SD = 1.6). The participants self-identified as White non-Hispanic (53.3%),

Black non-Hispanic (17.7%), Hispanic (15.4%), Asian or Pacific Rim (8.7%), and other races/ethnicities (4.9%).

Procedure

After participants provided consent to participate in the study, they were given access to the Study URL. They first completed a series of questionnaires, including demographic and personality measures described below. They were then redirected to participate in the TAP-Chat (Burt et al., 2020; detailed below).

Measures

Digital Aggression. Participants completed the TAP-Chat (see Figure 5), which included a tutorial followed by 24 trials of the reaction game, of which participants lost at least 50%. Similar to more recent versions of the TAP, participants were not required to respond to aggressive prompts. The mean chats sent to the participants were categorized by level of intensity (low vs. moderate vs. high). For example, low or no provocation (Level 0) chats included "lol sup?" and "Hey, how's it going?" Moderate provocation (Level 1) chats included "are you even trying! Lol!" and "beat you!" High provocation (Level 2) chats included "you SUCK at this game!" and "Who raised you to be this dumb?" To mimic an escalating feud, the TAP-Chat system was set to increase the intensity of the mean chat messages at specific intervals, with Trial 1 set to Level 0, Trials 2-13 set to Level 1, and Trials 14-24 set to Level 2. These "mean chats" used responsive design and were randomly selected from a pool of chats. To bolster the perception that they were playing a real opponent in the TAP-Chat, the program utilized automatic responses to participants' chats that contained words questioning the co-player (i.e., are you rigged, a bot, real, etc.). Participants were able to initiate chats at any time during the game, and they could also choose to remain non-responsive.

As in prior work (Burt et al., 2020), participant messages to their (fictitious) opponents were then coded for DA in two ways. First, they were coded by a team of four trained research assistants using a 6-point scale ranging from 0 (not aggressive at all) to 5 (very aggressive). Our DA coding scheme is presented in Table 7. Each member of the coding team rated each message, and these ratings were averaged across the four coders to yield an overall index of DA on the TAP-Chat (the intraclass correlation across all raters on all trials was 0.90). Second, participant messages were submitted to the LIWC software (Pennebaker et al., 2015) with a specific focus on anger (e.g., hate, kill) and swear (e.g., d*mn, piss) words (as in Study 1). As done by Burt et al. (2020), we dichotomized the two TAP-Chat LIWC variables as present versus absent and then added them to create a TAP-Chat AWI.

Anonymity. To model the effects of self- and other-anonymity, participants were randomly assigned to one of four conditions: (A) fully anonymous, (B) self-identified, other-anonymous, (C) self-anonymous, other-identified, or (D) fully identified (see Figure 6). For the identifiable condition, participants were shown the co-player's profile with a picture and a first name (a gender-neutral name Morgan was used). They were also asked to upload a recent photo of themselves taken within the last six months. For the anonymous condition, participants were shown the co-player's profile with the first name of Morgan and a generic symbol in place of a photo, and the participants were represented the same way.

Personality. Participants completed the 15-item Big Five Inventory Extra-Short Form with an added Big Ego question (BFI-2-XS; Soto & John, 2017) to measure five broad personality traits. Extraversion ($\alpha = .58$) indexes sociability, assertiveness, and energy level. Agreeableness ($\alpha = .59$) indexes compassion, respectfulness, and trust. Conscientiousness ($\alpha = .51$) indexes organization, productiveness, and responsibility. Negative emotionality ($\alpha = .65$)

indexes anxiety, depression, and emotional volatility. Open-mindedness ($\alpha = .41$; akin to Intellect/Imagination) assesses aesthetic sensitivity, intellectual curiosity, and creative imagination. Each domain has three items, except for Agreeableness, with four items, which are summed so that a high score indicates a high level of the trait. Although these reliabilities were lower than we hoped, we note that content validity is often prioritized over internal consistency for very short questionnaires such as this (John & Soto, 2007; Stanton et al., 2002).

Analyses

We began our analyses by reviewing the pictures uploaded by randomly assigned participants to the self-identified conditions. This was critical to ensure that the participants did not guise themselves under technical anonymity. If individuals had uploaded fictitious pictures (i.e., known celebrities, cartoons, and blank screens), they no longer fit the eligibility criteria for the self-identified conditions. Rather, they chose to participate in the TAP-Chat *anonymously* on their own accord. As such, participants with pictures that did not match their self-reported gender and ethnicity or participants who uploaded obviously fictitious pictures were recoded into the self-anonymous conditions from the self-identified conditions. The consequences of this decision for our sample sizes are seen in Table 8. Main analyses were conducted with these recoded sample sizes. Additional sensitivity analyses were also conducted to compare those in the self-identified conditions who uploaded their pictures to those who did not to evaluate potential differences between those who insisted on maintaining anonymity despite study instructions to the contrary versus those who allowed themselves to be identified. We also compared those who sent at least one chat during the TAP-Chat to those who did not send any chats to evaluate potential differences between those who actively participated in the TAP-Chat relative to those who did not.

To examine the first research question (i.e., What personality traits predict DA on the TAP-Chat?), we conducted a multiple regression analysis of the average coder ratings of DA and an ordinal logistic regression analysis for the AWI DA on the TAP-Chat. To examine the second research question (i.e., Does DA on the TAP-Chat differ for participants in the anonymous and identifiable conditions?), we evaluated the differences for each DA outcome via 2x2 factorial ANOVAs. Based on benchmarks suggested by Cohen (1988), Cohen's *d* effect sizes were interpreted as small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$). For both sets of analyses, gender was dummy-coded 1 for female and 0 for all other gender groups. Similarly, self-identified ethnicity was dummy-coded 1 for White and 0 for all other ethnic and racial groups.

Finally, a four-step hierarchical multiple regression was conducted to investigate the moderating role of anonymity on the relationship between personality traits and DA (the third research question). Prior to the analyses, personality traits were mean-centered to clarify the magnitudes of the regression coefficients (Irwin & McClelland, 2001). Each anonymity condition was dummy coded with the fully anonymous condition as the reference group. The dummy coding approach was employed due to the unbalanced sample size in each condition (Duke Global Health Institute, 2020). Interactions between personality traits and anonymity conditions were then specified using five interaction terms based on the anonymity dummy codes and mean-centered personality traits. Demographic variables (age, gender, and race/ethnicity) were entered in step one, personality traits were entered in step two, and the anonymity condition in step three. Statistical interaction terms between anonymity and personality in the final fourth step. As recommended by Aiken and West (1991), we plotted significant interaction effects at low (1 *SD* below the mean) and high (1 *SD* above the mean) levels of each personality trait. In addition, an inspection of Q-Q Plots revealed that standardized

regression residuals were normally distributed with the values for skewness between ± 2 and kurtosis between ± 7, which are considered acceptable to prove normal distribution (Hair et al., 2010) across anonymity conditions.

Results

Descriptive statistics and zero-order correlations between measured variables are provided in Table 9. Personality traits evidenced small-to-moderate correlations with one another, *r*s ranging from –.37 to .31. Notably, however, low agreeableness was the only personality trait to be significantly correlated with DA, an association that persisted across both indices of DA (TAP-Chat AWI: $r = –.10$; average TAP-Chat: $r = –.10$). There were no associations between personality traits and experimental conditions except for a small positive correlation (.09) between open-mindedness and self-identified condition.

Of the 211 participants randomly assigned to the self-identified conditions, 78 (37.0%) uploaded pictures that did not match their self-reported gender and ethnicity or were obviously not real (a cartoon picture). Independent *t*-tests were conducted to compare those who uploaded believable pictures ("real pictures") to those who did not ("fake pictures"). The results indicate that individuals high in extraversion ($d = 0.30$, $p < .05$) and open-mindedness ($d = 0.37$, $p < .05$) were more likely to upload their real pictures. However, there were no significant differences in demographic, other personality traits, or DA (*d* ranged from 0.02 to 0.24, all *ns*). As expected, participants who posted their pictures were more likely to send at least one chat on the TAP-Chat ("chatters") compared to those who posted their fake pictures ($d = 0.27$, $p < .05$).

Of the 533 participants, 455 (82.3%) sent at least one chat to their fictitious opponent. Independent *t-tests* comparing those who sent a chat ("chatters") to those who did not ("non-chatters") indicate that chatters were more likely to identify as male or transgender ($d = 0.36$, $p <$

.01) and identify as ethnic minority (d = 0.22, p < .05). Chatters were also slightly less likely to be conscientious (d = 0.24, p < .05). However, they did not differ on any other personality traits (d ranged from 0.09 to 0.22, all *ns*).

We also evaluated whether participants raised suspicions about their co-player or the reaction game (e.g., whether they believed their co-player was a real person). Among the 455 chatters, 94 participants (20.7%) sent chats that indicated suspicions regarding their co-player, using words such as *bot, fake,* or *not real*. Independent *t-tests* were conducted to compare those who questioned the co-player to those who did not. The results indicate that those who identify as male or transgender (d = 0.56, p < .001), White (d = 0.23, p < .05), or younger (d = 0.54, p < .05) were observed to question their co-player more often. They also tended to be lower in agreeableness (d = 0.33, p < .01) and conscientiousness (d = 0.30, p < .05), but there were no significant differences observed in other personality traits (d ranged from 0.04 to 0.15, all *ns*). Notably, those who questioned the co-player also used more antisocial words than those who did not question their co-player (d = 0.45, p < .001) but did not differ in their coder ratings of DA (d = 0.19, *ns*).

RQ1. Personality Predictors of in-vivo DA.

An ordinal logistic regression analysis was estimated to investigate whether personality traits predict in-vivo DA as indexed via the AWI. Together, the predictors accounted for a significant amount of variance in the outcome, with likelihood ratio $\chi^2(5) = 12.60$, $p < .05$. As hypothesized, agreeableness and open-mindedness significantly independently predicted the AWI (see Table 10a). Low agreeableness was associated with an increase in the odds of using more antisocial words on the TAP-Chat, with an odds ratio of 0.71 (95% CI, .55 to .92), Wald $\chi^2(1) = 7.00$, p < .01. An increase in open-mindedness was associated with an increase in the

odds of using more antisocial words on the TAP-Chat, with an odds ratio of 1.33 (95% CI, 1.02 to 1.71), Wald $\chi^2(1) = 4.63$, $p < .05$.

A multiple regression was then run to examine whether any of the Big Five personality traits predicted average coder ratings of DA. As hypothesized, individuals with low agreeableness ($\beta = –0.14$, $p < .01$) engaged in higher DA (Table 10b). By contrast, individuals with high open-mindedness did not engage in higher levels of DA by coder ratings ($\beta = –0.36$, $p = .45$).

RQ2. Differences in in-vivo DA across Anonymity Conditions.

We ran a 2x2 factorial ANOVA assessing the effects of anonymity conditions on each index of DA. In contrast to our hypotheses, there were no significant differences in the AWI of individuals in the self-anonymous ($M = .60$, $SD = .82$) versus self-identified conditions ($M = .70$, $SD = .85$), $F(1, 451) = 1.17$, $p = .28$, $d = 0.12$ nor between individuals in the other-anonymous ($M = .65$, $SD = .84$) versus the other-identified conditions ($M = .60$, $SD = .83$), $F(1, 451) = 0.94$, $p = .33$, $d = 0.06$. There was also no significant interaction between self- and other anonymity, $F(1, 451) = 0.55$, $p = .37$, although self-identified participants playing against anonymous bots ($M = .78$, $SD = .89$) used more antisocial words ($d = 0.21$) than anonymous participants playing against identified bots ($M = .59$, $SD = .84$).

Similarly, there were no significant differences in DA between individuals in the self-anonymous ($M = .73$, $SD = .74$) versus the self-identified conditions ($M = .76$, $SD = .69$), $F(1, 451) = 0.12$, $p = .73$, $d = 0.04$ nor between individuals in the other-anonymous ($M = .72$, $SD = .75$) versus the other-identified conditions ($M = .76$, $SD = .71$), $F(1, 451) = 0.09$, $p = .76$, $d = 0.05$. There was also no significant interaction between them, $F(1, 451) = 2.26$, $p = .13$, $d = 0.06$.

RQ3. Associations between Anonymity and Personality

For our final objective, we examined whether experimentally manipulated anonymity conditions moderated the relationship between personality traits and in-vivo DA. The moderated regression results for the AWI are reported in Table 11. As seen there, low agreeableness was the only personality trait that uniquely predicted the AWI in the self-anonymous, other-identified anonymity condition (β = –0.13, p <.05) and in the fully identified condition (β = –0.16, p <.01). We also observed significant interactions between extraversion and anonymity (β = 0.13, p <.05). The results of a simple slope analysis are shown in Figure 7. As can be seen, individuals high in extraversion used more antisocial words during the experiment when the participant and co-player were fully identified compared to when both were fully anonymous. In contrast to our hypothesis, anonymity did not significantly moderate the relationship between agreeableness and DA or negative emotionality and DA. Of note, however, there was a trend for negative emotionality (β = 0.11, p = .07). Consistent with this, a simple slope analysis revealed that individuals high in negative emotionality used more antisocial words when both the participant and co-player were fully identified compared to when both were fully anonymous.

The moderated regression results for average coder ratings of DA are reported in Table 12. As seen there, gender was the only demographic variable that uniquely predicted DA across different anonymity conditions. Similar to the AWI, low agreeableness uniquely predicted coder ratings of DA in the self-identified, other-anonymous anonymity condition (β = –0.12, p <.05), self-anonymous, other-identified anonymity condition (β = –0.16, p <.05), and fully identified condition (β = –0.14, p <.01). Interestingly, no significant interaction was reported between anonymity and personality traits in predicting average coder ratings of DA. However, similar to the AWI, there was a trend for negative emotionality (β = 0.11, p = .07), in which those high in

negative emotionality engaged in higher DA during the TAP-Chat when the participant and co-player were fully identified compared to when both were fully anonymous.

Discussion

The first goal of the current study was to examine personality predictors of in-vivo DA as assessed using the TAP-Chat. As hypothesized, low agreeableness and high open-mindedness predicted increased use of antisocial words during the experiment. Such findings echo prior literature's findings (Kokkinos & Antoniadou, 2019; Seigfried-Spellar & Lankford, 2018; van Geel et al., 2017; Zezulka & Seigfried-Spellar, 2016) that collectively argued for low agreeableness as a robust predictor of DA. This also aligns with previous studies that have established agreeableness as closely linked to antisocial behaviors (Jones et al., 2011; Miller et al., 2008; Miller & Lynam, 2001). Individuals with low agreeableness have been found to exhibit deviant interpersonal behaviors (Kokkinos et al., 2016) and lack the skills needed to effectively manage hostility and disagreement in interpersonal interactions (McCullough et al., 2001). These tendencies may be exacerbated in online settings, including making harmful comments on Facebook (Karl et al., 2010) or even taking revenge on others online (Baldasare et al., 2012). We also found that high open-mindedness predicted higher DA during the experiment, which again aligns with previous findings (Kim et al., 2020). Such findings suggest that highly open-minded individuals may have been more likely to engage in the TAP-Chat, perhaps because they are more open to experience and expressive in their interpersonal interactions (Zezulka & Seigfried-Spellar, 2016). Given that the TAP-Chat is a negatively-valenced interpersonal interaction by design (i.e., mean-provoking chats sent by the program when participants lose in the reaction game), this higher engagement may have resulted in higher use of antisocial words during the experiment.

As a second goal for our study, we were interested in examining whether in-vivo DA differed across individuals randomly assigned to different anonymity conditions. In contrast to our hypotheses, we did not observe any significant differences in DA between individuals in the self-anonymous versus the self-identified condition or between individuals in the other-anonymous and other-identified conditions. Despite this, our results did suggest that self-identified participants playing against anonymous bots used more antisocial words than self-anonymous participants playing against identified bots. These findings contradict previously reported positive associations between anonymity and DA, in which those who perceived themselves as anonymous engage in more DA (Barlett, 2015; Barlett et al., 2016, 2017, 2019; Barlett & Gentile, 2012; Wright, 2013). In contrast, a study by Rost et al. (2016) recognized a need to consider the perpetrators' motivation when studying the effects of anonymity on online aggression, as they observed higher levels of online aggression in non-anonymous comments posted to an online social-political platform, perhaps because those individuals perceived anonymity as a barrier to what they wanted to achieve. Similarly, it is possible that self-identified participants felt more justified in their aggressive responses and more motivated to share their opinions after receiving mean chats from their anonymous co-player on the TAP-Chat, perhaps because they felt more personally targeted.

As a final objective, we examined whether anonymity moderated the relationship between personality and in-vivo DA during the experiment. We found that anonymity moderated the relationship between extraversion and in-vivo DA, such that individuals high in extraversion used more antisocial words during the experiment when the participant and co-player were fully identified compared to when both were fully anonymous. Although interesting, this finding was not one we hypothesized. One possibility is that because extraverts are generally more motivated

to communicate online to seek social stimulation, explore their social nature, and enhance their social networks (Mark & Ganzach, 2014), both self- and other-disclosure (in the form of names and pictures in this experiment) fostered increased motivation and investment in the social exchange between the self and the co-player. Since the pre-programmed mean chats sent by the co-player were designed to provoke participants, they may have normalized DA in that context (*depersonalization*; McInroy & Mishna, 2017; Reason et al., 2016) and thus made DA appear less risky. For example, there is evidence that people experiencing hostility in online gaming platforms may not perceive it as aggression but rather as part of their gaming experiences (McInroy & Mishna, 2017). Moreover, as each successful act of DA was positively reinforced, this may have led to more positive attitudes toward DA and increased its frequency, as posited by the BGCM (Barlett & Gentile, 2012).

As an exploratory study, it is worth mentioning that there was a trend of interaction between negative emotionality and anonymity, such that individuals high in negative emotionality used more antisocial words during the experiment in the fully identified condition compared to the fully anonymous condition for both indices of DA. This finding is partly in line with our hypothesis given that individuals scoring high on neuroticism tend to be more sensitive to stress, more suspicious of others' motives, and have strong emotional reactions (Bolger & Zuckerman, 1995; Connolly & O'Moore, 2003). When these individuals are in "hot" emotional reactive states (mostly associated with irritability; Caprara et al., 2013, 2014), they may enter the phase of immediate appraisal of the GAM that elicits impulses to engage in DA (Anderson & Bushman, 2002; Savage & Tokunaga, 2017). When they are in "cold" emotional states (mostly associated with hostile rumination; Caprara et al., 2013, 2014), they may go through multiple appraisals and decision processes (Anderson & Bushman, 2002) and view DA as a legitimate

response to the provocations elicited by the co-player (Denissen & Penke, 2008; Milam et al., 2009; Suls & Martin, 2005). Furthermore, in the presence of negative affect and heightened arousal, irritability can be easily provoked, strengthening aggressive tendencies and encouraging anger-motivated aggressive behaviors (Anderson & Bushman, 2002).

Strengths and Limitations

There are several limitations to the current study. First, Qualtrics reported that many participants chose to terminate the study before and during the TAP-Chat, which was cited as the main reason to close the project due to the slow participation rate. Given this, certain characteristics may be shared by our pool of participants that chose to complete the study. For example, in contrast to our initial gender breakdown of up to 50% female and 50% male, our final sample included 73% female and 22% male. This means that a substantial number of male participants chose to terminate the study prematurely. The closing of the study also affected the unequal sample size across the four anonymity conditions.

Next, roughly 15% of our participants did not send any chats to their fictitious co-player. This is a notable improvement from a previous validation study (Burt et al., 2020), in which roughly a quarter to a third of the participants did not engage in the TAP-Chat. While modifications such as automatic bot responses are helping to reduce the non-responsiveness, continued modifications may help to increase participants' engagement.

Next, despite the random assignment of participants to different conditions, we observed a small positive association between open-mindedness and the self-identified condition. While we controlled for personality traits in our moderation analyses, it is important to account for the potential pre-existing differences among participants in our experimental study. Relatedly, among the participants that were randomly assigned to the self-identified conditions, about a

third of the participants did not upload their real pictures. As participants who posted their pictures were more likely to send at least one chat, there is a clear need to continue the development of the TAP-Chat to reduce participants' (perceived) anonymity. One example is taking photos of the participants, or at least making them believe their photos were taken before the TAP-Chat. Moreover, TAP-Chat deliberately sent mean chats to our participants following their loss in reaction games. While these games were minimally competitive, we may have evoked additional emotional reactions, such as competitiveness and revenge.

In addition, our personality measure had lower internal consistency reliability alpha values than ideal. While our study chose to administer a shorter 16-item BFI-2-XS to minimize the cost and time of the study, it may be informative to use more comprehensive personality measures such as the 50-item International Personality Item Pool-Five Factor Model (IPIP-FFM; Goldberg, 1999) or 44-item Big Five Inventory (BFI; John & Srivastava, 1999).

Despite these limitations, the current study had several strengths. It is one of the first studies to our knowledge to experimentally examine whether and how the association between personality traits and DA varies with the presence of self-anonymity and other-anonymity. Our study employed a brief laboratory-based paradigm to examine an individual's aggressive tendencies in a computer-mediated environment instead of relying on self-reported questionnaires. Although the questionnaires may be a useful measurement of DA in general, they do not capture actual instantiations of DA behaviors in laboratory and in-vivo contexts. Our study also contributed to existing knowledge of DA by investigating complex interactions between personal and contextual factors within the GAM framework. Another strength of our study was the use of a national sample of racially and ethnically diverse young adults recruited across the United States to increase generalizability. In this way, the study helped to address an

important gap in cyberbullying literature, as extant studies have largely focused on the K-12 school environment.

Implications

Our study sought to evaluate the roles of the technological environment (particularly anonymity) and individual factors (specifically personality traits) in the multifaceted origins of DA. We sought to lay the groundwork for future research investigating how technological affordances of computer-mediated communication and personality interact to influence DA. Results pointed to individuals high in extraversion using more antisocial words during the experiment when the participant and co-player were fully identified compared to when both were fully anonymous. Though not in line with our hypothesis, the results support the GAM as a useful framework. It suggests that not all personality traits interact with the technological environment, or at least with anonymity, to shape DA in the moment. Instead, personality traits may interact with other situational factors, such as the type of online platform and its climate and geographical location, to affect DA. Some personality traits (e.g., low agreeableness) may be associated with DA, irrespective of situational factors, since DA often stems from the immediate appraisal that elicits impulsive actions (Savage & Tokunaga, 2017). Thus, future research should seek to examine other person-by-environment interactions that may perpetuate DA.

Second, our results illustrated that anonymity is a less potent predictor of DA than we had hypothesized. At the very least, its consequences for DA appear more variable than we assumed. This may point to previous studies that emphasized the importance of social learning constructs where increased time spent online indirectly predicted DA through the development of (perceived) anonymity and positive cyberbullying attitudes (Barlett & Gentile, 2012; Barlett et al., 2019). Thus, rather than assuming anonymity as a by-product of technology and aggressive

individuals, it is crucial to consider how other factors may shape the effects of anonymity on DA. Future studies should consider various aspects and processes of the GAM to consider additional pathways to increase our understanding of DA and ultimately aid in developing DA prevention and intervention efforts.

Finally, with the persistent and intensified DA in online environments, our field must continue to examine various factors in which these negative and socially harmful behaviors perpetuate. Our findings also highlight the importance of measuring the DA construct in multiple ways. Instead of relying on self-reported measures, assessing DA across different contexts is important. This is especially important as social networking sites afford different anonymous conditions. For instance, Snapchat servers delete their messages 24 hours after users have viewed them or one week after the message was sent, whichever is sooner (Snapchat, 2023). In contrast, Facebook (rebranded as Meta) endorses the real-name system policy for their user profiles such that everyone uses the name they go by in everyday life (Meta, 2023). Thus, each platform may need to be studied individually rather than examined for a general effect of DA across platforms.

GENERAL DISCUSSION

Although extant evidence had pointed to personality and technical anonymity as important predictors of DA, no study had yet to examine whether and how anonymity might interact with personal predispositions to influence DA. The current study aimed to address this gap in literature. Findings for personality, anonymity, and their interactions are detailed below in turn.

Personality. Both studies found that high intellect/imagination (akin to open-mindedness) predicted DA, results which align nicely with previous findings (Kim et al., 2020). Such findings suggest that individuals with creativity and intellectual curiosity may be more likely to engage in DA across multiple platforms, perhaps because such individuals are more open to experience and expressive in their interpersonal interactions (Zezulka & Seigfried-Spellar, 2016). Indeed, extant work has indicated that Twitter users tend to be more intellectually curious with a wide range of interests and that they seek cognitively-oriented information more than socially-oriented information (e.g., they are more interested in the news than in people) (Hughes et al., 2012). Our results suggest that this higher engagement may also result in higher use of antisocial words as they come across increased incidents of DA on Twitter and the TAP-Chat (a negatively-valenced interpersonal interaction by design). Moreover, we found that low agreeableness also predicted DA on the TAP-Chat, echoing previous studies that identified low agreeableness as the trait most closely linked to antisocial behaviors (Jones et al., 2011; Miller et al., 2008; Miller & Lynam, 2001). The lack of emotion regulation and social skills exhibited by individuals with low agreeableness (Kokkinos et al., 2016; McCullough et al., 2001) may be exacerbated in online settings, resulting in increased DA.

Technical anonymity. Technical anonymity was not a clear predictor of DA in either of our studies, a finding that is inconsistent with prior research. In the first study, we found that individuals with anonymous Twitter accounts engaged in higher DA on Twitter than those with identifiable Twitter accounts, but these findings did not replicate with a second sample. We also failed to find evidence of significant differences in DA between individuals in the self-anonymous vs. self-identified or other-anonymous vs. other-identified conditions on the TAP-Chat in Study 2. Instead, we observed a trend for self-identified participants playing against anonymous bots using more antisocial words than self-anonymous participants playing against identified bots. Thus, while anonymity may be a contributing factor of DA in some online contexts, it does not seem to directly increase DA on Twitter or the TAP-Chat.

Personality x Technical Anonymity. Although anonymity did not directly seem to predict DA in our studies, our results did suggest that certain personality traits may interact with technical anonymity to increase DA. In particular, and consistent with our hypotheses, individuals high in intellect/imagination engaged in higher DA on Twitter when they had anonymous accounts than when they had identifiable accounts, results that persisted over time and were replicated in a second sample. Given that individuals high on intellect/imagination are more expressive and open to experience (Zezulka & Seigfried-Spellar, 2016), they may be engaging in more online activities and interpersonal interactions, which may also expose them to increased DA opportunities. This continued exposure may act to normalize and desensitize Twitter users to DA, especially when their identities are hidden, leading to more positive attitudes toward DA and increased frequency (Barlett & Gentile, 2012).

Although this finding for intellect-imagination replicated across samples and over time in Study 1, it did not persist to our experimental manipulation of in-vivo DA in Study 2. By

contrast, individuals high in extraversion used more antisocial words on the TAP-Chat when the participant and co-player were fully identified compared to when both were fully anonymous, which was the opposite of what we had hypothesized. As extraverts are more motivated and invested in seeking social stimulation (Mark & Ganzach, 2014), self- and other-disclosure (in the form of names and pictures in Study 2) may have created a more personal social exchange on the TAP-Chat. Given that the pre-programmed mean chats are designed to provoke the participants, the fully-identified participants may have reacted more negatively, perceived the mean chats as more personal, and engaged in reciprocal DA when compared to the anonymous participants.

Implications and future research. Although we cannot make assumptions about the direction of causation, the current study has important implications for the field's understanding of DA. Foremost, our findings collectively serve to further illuminate the roles of personality and anonymity in the prevalence of DA while also indicating that these associations are measure- and context-specific. Namely, our positive and replicated findings of the interaction between intellect-imagination and technical anonymity were specific to Twitter and did not extend to our in-vivo assessment of DA. Similarly, positive findings of an extraversion-by-anonymity interaction were specific to the TAP-Chat and were not observed on Twitter. In some ways, this disconnect makes sense. Twitter captures real-world perpetration of DA occurring over days to months during interactions with a vast number of users, while the TAP-Chat is designed to elicit negative responses over approximately 10 minutes. Nevertheless, these inconsistencies collectively suggest that personality-by-anonymity interactions may be context- or platform-specific. Given this, we suggest that researchers studying the origins and correlates of DA focus on documenting and understanding differences in risk predictors across platforms. Inherent in doing this work would be a shift away from questionnaires and towards real-world behavior on

social networking sites and experimental assessments. Future research needs to be more attentive and mindful of different measures of DA across various contexts.

Relatedly, because the current set of studies focused only on the Big Five broad domains of personality, future research should consider facets of personality. This would be important in light of prior studies highlighting the importance of examining personality at more specific facet-levels beyond the domains (De Young et al., 2007; Vize et al., 2018). For example, facets related to interpersonal antagonism (e.g., low Altruism and Compliance) and disinhibition (e.g., low Deliberation and Dutifulness) demonstrated the most consistent associations with broad antisocial behaviors (Derefinko et al., 2011; Miller et al., 2012; Vize et al., 2018). Similarly, future studies of anonymity should make efforts to conceptualize and assess anonymity as a multi-faceted construct, with direct consideration of both the type of anonymity (technical versus social) and the point of view (self versus other). In a given online context, several different combinations of anonymity may be at play, and yet prior studies assessing anonymity have only rarely considered this. Future work should seek more precision in their conceptualizations of anonymity.

Next, future studies will need to account for the speed at which technology is evolving and advancing. As the field moves forward, it must continuously work to keep up with technological advancement. If not, the field will always lag behind the current trends with possibly irrelevant intervention strategies and outdated policies (David-Ferdon & Hertz, 2007). For example, participants were asked about DA occurring through the internet and email in the early to mid-2000s, but as technology evolves, researchers must update their measures to reflect current trends among youth to measure DA occurring on different platforms (Bauman & Bellmore, 2015). In short, researchers should constantly consult and seek feedback from

individuals to ensure that measures of DA do not become outdated (Mishna et al., 2009). It is also important to specify the time period in which the data were collected. In our case, for example, our assessments of DA on Twitter are restricted to the historical version of Twitter that preceded its acquisition by Elon Musk. It thus remains unclear how DA on modern-day Twitter may relate to anonymity, personality, and their interaction.

All that said, the current findings are important, in that they illuminate, for the first time, how personality traits interact with anonymity to influence DA on specific platforms. Such findings dovetail very well with prior theory in the GAM (Anderson & Bushman, 2002), which hypothesizes that personal factors interact with situational factors to influence the internal states of individuals, which then influences their engagement in DA. Future studies should continue to examine these kinds of interactions to help broaden the field's understanding of DA, with the ultimate goal of informing prevention and intervention efforts aiming to reduce DA. Ideally, such efforts will view DA holistically so that intervention programs also attend to what might also be taking place in an individual's "real world."

REFERENCES

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Al-garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*, *63*, 433–443. https://doi.org/10.1016/j.chb.2016.05.051

Alhajji, M., Bass, S., & Dai, T. (2019). Cyberbullying, mental health, and violence in adolescents and associations with sex and race: Data from the 2015 youth risk behavior survey. *Global Pediatric Health*, *6*. https://doi.org/10.1177/2333794x19868887

Allen, J. J., & Anderson, C. A. (2017). General Aggression Model. In P. Roessler, C. A. Hoffner, & L. van Zoonen (Eds.) *International Encyclopedia of Media Effects* (pp. 1-15). Wiley-Blackwell. https://doi.org/10.1002/9781118783764.wbieme0078

Althubaiti, A. (2016). Information bias in health research: Definition, pitfalls, and adjustment methods. *Journal of Multidisciplinary Healthcare*, 211. https://doi.org/10.2147/JMDH.S104807

Amichai-Hamburger, Y., & Etgar, S. (2019). Personality and Internet use: The case of introversion and extroversion. In A. Attrill-Smith, C. Fullwood, M. Keep, & D. J. Kuss (Eds.), *The Oxford Handbook of Cyberpsychology* (pp. 57–73). Oxford University Press.

Amichai-Hamburger, Y., & Vinitzky, G. (2010). Social network use and personality. *Computers in Human Behavior*, *26*(6), 1289–1295. https://doi.org/10.1016/j.chb.2010.03.018

Anderson, C. A., & Bushman, B. J. (2002). Human Aggression. *Annual Review of Psychology*, *53*, 27-51.

Arıcak, O. T. (2009). Psychiatric symptomatology as a predictor of cyberbullying among university students. *Eurasian Journal of Educational Research (EJER)*, *34*, 167-184.

Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., & Stillwell, D. (2012, June). Personality and patterns of Facebook usage. *Proceedings of the 4th Annual ACM Web Science Conference* (pp. 24-32). https://doi.org/10.1145/2380718.2380722

Badiuk, B. B. (2006). Cyberbullying in the global village: The worldwide emergence of high-tech as a weapon for bullies. In A. Green (Ed.), *Education students' anthology (Vol. 9)* (pp. 12–16). Winnipeg, MB: Faculty of Education.

Bakker, S. (2022, June 30). *Gen Z chooses anonymity over fame*. FreedomLab. Retrieved January 21, 2023, from https://www.freedomlab.com/posts/gen-z-chooses-anonymity-over-fame

Baldasare, A., Bauman, S., Goldman, L., & Robie, A. (2012). Chapter 8 Cyberbullying? Voices of College Students. In L. A. Wankel & C. Wankel (Eds.), *Cutting-Edge Technologies in Higher Education* (pp. 127–155). Emerald Group Publishing Limited. https://doi.org/10.1108/S2044-9968(2012)0000005010

Barlett, C. P. (2015a). Anonymously hurting others online: The effect of anonymity on cyberbullying frequency. *Psychology of Popular Media Culture*, *4*(2), 70–79. http://dx.doi.org.proxy2.cl.msu.edu/10.1037/a0034335

Barlett, C. P. (2015b). Predicting adolescent's cyberbullying behavior: A longitudinal risk analysis. *Journal of Adolescence*, *41*, 86–95. https://doi.org/10.1016/j.adolescence.2015.02.006

Barlett, C. P., & Chamberlin, K. (2017). Examining cyberbullying across the lifespan. *Computers in Human Behavior*, *71*, 444–449. https://doi.org/10.1016/j.chb.2017.02.009

Barlett, C., Chamberlin, K., & Witkower, Z. (2017). Predicting cyberbullying perpetration in emerging adults: A theoretical test of the Barlett Gentile Cyberbullying Model. *Aggressive Behavior*, *43*(2), 147–154. https://doi.org/10.1002/ab.21670

Barlett, C. P., & Gentile, D. A. (2012). Attacking others online: The formation of cyberbullying in late adolescence. *Psychology of Popular Media Culture*, *1*(2), 123–135. https://doi.org/10.1037/a0028113

Barlett, C. P., Gentile, D. A., & Chew, C. (2016). Predicting cyberbullying from anonymity. *Psychology of Popular Media Culture*, *5*(2), 171–180. https://doi.org/10.1037/ppm0000055

Barlett, C. P., & Helmstetter, K. M. (2018). Longitudinal relations between early online disinhibition and anonymity perceptions on later cyberbullying perpetration: A theoretical test on youth. *Psychology of Popular Media Culture*, *7*(4), 561–571. https://doi.org/10.1037/ppm0000149

Barlett, C. P., & Kowalewski, D. A. (2019). Learning to cyberbully: An extension of the Barlett Gentile cyberbullying model. *Psychology of Popular Media Culture*, *8*(4), 437–443. https://doi.org/10.1037/ppm0000183

Barlett, C. P., Madison, C. S., Heath, J. B., & DeWitt, C. C. (2019). Please browse responsibly: A correlational examination of technology access and time spent online in the Barlett Gentile Cyberbullying Model. *Computers in Human Behavior*, *92*, 250–255. https://doi.org/10.1016/j.chb.2018.11.013

Baughman, H. M., Dearing, S., Giammarco, E., & Vernon, P. A. (2012). Relationships between bullying behaviours and the Dark Triad: A study with adults. *Personality and Individual Differences*, *52*(5), 571-575. https://doi.org/10.1016/j.paid.2011.11.020

Bauman, S., & Bellmore, A. (2015). New directions in cyberbullying research. *Journal of School Violence, 14*, 1-10. https://doi.org/10.1080/15388220.2014.968281

Bolger, N., & Zuckerman, A. (1995). A framework for studying personality in the stress process. *Journal of Personality and Social Psychology*, *69*(5), 890–902. https://doi.org/10.1037/0022-3514.69.5.890

Bonanno, R. A., & Hymel, S. (2013). Cyber Bullying and Internalizing Difficulties: Above and Beyond the Impact of Traditional Forms of Bullying. *Journal of Youth and Adolescence*, *42*(5), 685–697. https://doi.org/10.1007/s10964-013-9937-2

Brandtzæg, P. B., Staksrud, E., Hagen, I., & Wold, T. (2009). Norwegian Children's Experiences of Cyberbullying When Using Different Technological Platforms. *Journal of Children and Media*, *3*(4), 349–365. https://doi.org/10.1080/17482790903233366

Brochado, S., Soares, S., & Fraga, S. (2017). A Scoping Review on Studies of Cyberbullying Prevalence Among Adolescents. *Trauma, Violence, & Abuse*, *18*(5), 523–531. https://doi.org/10.1177/1524838016641668

Buckels, E. E., Jones, D. N., & Paulhus, D. L. (2013). Behavioral Confirmation of Everyday Sadism. *Psychological Science*, *24*(11), 2201–2209. https://doi.org/10.1177/0956797613490749

Burt, S. A., & Alhabash, S. (2018). Illuminating the nomological network of digital aggression: Results from two studies. *Aggressive Behavior*, *44*(2), 125–135. https://doi.org/10.1002/ab.21736

Burt, S. A., Kim, M., & Alhabash, S. (2020). A novel in vivo measure of cyberaggression. *Aggressive Behavior*, *46*(5), 449–460. https://doi.org/10.1002/ab.21911

Calvin, A. J., Bellmore, A., Xu, J. M., & Zhu, X. (2015). # bully: Uses of hashtags in posts about bullying on Twitter. *Journal of School Violence*, *14*(1), 133-153. https://doi.org/10.1080/15388220.2014.966828

Campbell, M. A. (2005). Cyber bullying: An old problem in a new guise? Australian Journal of Guidance and Counselling, 15(1), 68-76.

Cappos, J., Peddinti, S. T., & Ross, K. W. (2017, October 12). *User anonymity on Twitter*. InfoQ. Retrieved January 15, 2023, from https://www.infoq.com/articles/user-anonymity-twitter/#:~:text=Quantifying%20User%20Anonymity,first%20or%20a%20last%20name

Caprara, G. V., Alessandri, G., Tisak, M. S., Paciello, M., Caprara, M. G., Gerbino, M., & Fontaine, R. G. (2013). Individual Differences in Personality Conducive to Engagement in Aggression and Violence. *European Journal of Personality*, *27*(3), 290–303. https://doi.org/10.1002/per.1855

Caprara, G. V., Tisak, M. S., Alessandri, G., Fontaine, R. G., Fida, R., & Paciello, M. (2014). The contribution of moral disengagement in mediating individual tendencies toward aggression and violence. *Developmental psychology*, *50*(1), 71-85. https://doi.org/10.1037/a0034488

Cassidy, W., Faucher, C., & Jackson, M. (2013). Cyberbullying among youth: A comprehensive review of current international research and its implications and application to policy and practice. *School psychology international*, *34*(6), 575-612. https://doi.org/10.1177/0143034313479697

Chan, T. K., Cheung, C. M., & Wong, R. Y. (2019). Cyberbullying on social networking sites: the crime opportunity and affordance perspectives. *Journal of Management Information Systems*, *36*(2), 574-609. https://doi.org/10.1080/07421222.2019.1599500

Chan, T. K. H., Cheung, C. M. K., & Lee, Z. W. Y. (2021). Cyberbullying on social networking sites: A literature review and future research directions. *Information & Management*, *58*(2), 103411. https://doi.org/10.1016/j.im.2020.103411

Choi, D., Hwang, M., Kim, J., Ko, B., & Kim, P. (2014). Tracing trending topics by analyzing the sentiment status of tweets. *Computer Science and Information Systems*, *11*(1), 157-169. https://doi.org/10.2298/CSIS130205001C

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic.

Connolly, I., & O'Moore, M. (2003). Personality and family relations of children who bully. *Personality and Individual Differences*, *35*(3), 559–567. https://doi.org/10.1016/S0191-8869(02)00218-0

Costa, P. T., & McCrae, R. R. (1985). *The NEO personality inventory*. 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. https://doi.org/10.1016/j.paid.2016.06.043

David-Ferdon, C., & Hertz, M. F. (2007). Electronic Media, Violence, and Adolescents: An Emerging Public Health Problem. *Journal of Adolescent Health*, *41*(6), S1–S5. https://doi.org/10.1016/j.jadohealth.2007.08.020

Denissen, J. J. A., & Penke, L. (2008). Neuroticism predicts reactions to cues of social inclusion. *European Journal of Personality*, *22*(6), 497–517. https://doi.org/10.1002/per.682

Derefinko, K., DeWall, C. N., Metze, A. V., Walsh, E. C., & Lynam, D. R. (2011). Do different facets of impulsivity predict different types of aggression? *Aggressive behavior*, *37*(3), 223-233. https://doi.org/10.1002/ab.20387

DeWall, C. N., Buffardi, L. E., Bonser, I., & Campbell, W. K. (2011). Narcissism and implicit attention seeking: Evidence from linguistic analyses of social networking and online presentation. *Personality and Individual Differences*, *51*(1), 57–62. https://doi.org/10.1016/j.paid.2011.03.011

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of personality and social psychology*, *93*(5), 880-896. https://doi.org/10.1037/0022-3514.93.5.880

Dilmaç, B. (2009). Psychological needs as a predictor of cyber bullying: A preliminary report on college students. *Educational Sciences: Theory and Practice*, *9*(3), 1307-1325.

Dong, Y. (2019). The effect of traditional bullying-victimization on behaviour cyberbullying among college students: Based on the structural equation mode. *International Journal of Social Psychology*, 35(1), 175-199. https://doi.org/10.1080/02134748.2019.1687969

Dooley, J. J., Pyżalski, J., & Cross, D. (2009). Cyberbullying Versus Face-to-Face Bullying: A Theoretical and Conceptual Review. *Zeitschrift Für Psychologie / Journal of Psychology*, *217*(4), 182–188. https://doi.org/10.1027/0044-3409.217.4.182

Duke Global Health Institute. (2020, August 27). *Core Guide: Dummy and Effect Coding in the Analysis of Factorial Designs*. Research Design & Analysis Core. Retrieved April 24, 2023, from https://sites.globalhealth.duke.edu/rdac/wp-content/uploads/sites/27/2020/08/Core-Guide_Dummy-and-Effect-Coding_16-03-20.pdf

Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of computer-mediated communication*, *12*(4), 1143-1168. https://doi.org/10.1111/j.1083-6101.2007.00367.x

Espelage, D. L., & Swearer, S. M. (2003). Research on school bullying and victimization: What have we learned and where do we go from here?. *School psychology review*, *32*(3), 365-383.

Festl, R., & Quandt, T. (2013). Social Relations and Cyberbullying: The Influence of Individual and Structural Attributes on Victimization and Perpetration via the Internet. *Human Communication Research*, *39*(1), 101–126. https://doi.org/10.1111/j.1468-2958.2012.01442.x

Fox, J., Cruz, C., & Lee, J. Y. (2015). Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. *Computers in Human Behavior*, *52*, 436–442. https://doi.org/10.1016/j.chb.2015.06.024

Francisco, S. M., Veiga Simão, A. M., Ferreira, P. C., & Martins, M. J. das D. (2015). Cyberbullying: The hidden side of college students. *Computers in Human Behavior*, *43*, 167–182. https://doi.org/10.1016/j.chb.2014.10.045

Gil de Zúñiga, H., Diehl, T., Huber, B., & Liu, J. (2017). Personality Traits and Social Media Use in 20 Countries: How Personality Relates to Frequency of Social Media Use, Social Media News Use, and Social Media Use for Social Interaction. *Cyberpsychology, Behavior, and Social Networking*, *20*(9), 540–552. https://doi.org/10.1089/cyber.2017.0295

Gilbert, F., & Daffern, M. (2011). Illuminating the relationship between personality disorder and violence: Contributions of the General Aggression Model. *Psychology of Violence*, *1*(3), 230. https://doi.org/10.1037/a0024089

Gilbert, F., Daffern, M., & Anderson, C. A. (2017). The General Aggression Model and its application to violent offender assessment and treatment. *The Wiley handbook of violence and aggression*, 1-13. https://doi.org/10.1002/9781119057574.whbva037

Giordano, A. L., Prosek, E. A., & Watson, J. C. (2021). Understanding Adolescent Cyberbullies: Exploring Social Media Addiction and Psychological Factors. *Journal of Child and Adolescent Counseling*, *7*(1), 42–55. https://doi.org/10.1080/23727810.2020.1835420

Giumetti, G. W., & Kowalski, R. M. (2022). Cyberbullying via social media and well-being. *Current Opinion in Psychology*, *45*, 101314. https://doi.org/10.1016/j.copsyc.2022.101314

Goebert, D., Else, I., Matsu, C., Chung-Do, J., & Chang, J. Y. (2011). The Impact of Cyberbullying on Substance Use and Mental Health in a Multiethnic Sample. *Maternal and Child Health Journal*, *15*(8), 1282–1286. https://doi.org/10.1007/s10995-010-0672-x

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe, 7,* 7-28.

Guadagno, R. E., Okdie, B. M., & Eno, C. A. (2008). Who blogs? Personality predictors of blogging. *Computers in Human Behavior*, *24*(5), 1993-2004. https://doi.org/10.1016/j.chb.2007.09.001

Hair, J., Black, W. C., Babin, B. J. & Anderson, R. E. (2010). *Multivariate data analysis (7th Ed.)*. Upper Saddle River, New Jersey: Pearson Educational International.

Hardie, B. (2020). *Studying Situational Interaction: Explaining Behaviour By Analysing Person-Environment Convergence*. Springer Nature. https://doi.org/10.1007/978-3-030-46194-2

Harrison, T. (2015). Virtuous reality: Moral theory and research into cyber-bullying. *Ethics and Information Technology*, *17*(4), 275–283. https://doi.org/10.1007/s10676-015-9382-9

Hayne, S. C., & Rice, R. E. (1997). Attribution accuracy when using anonymity in group support systems. *International Journal of Human-Computer Studies*, *47*(3), 429-452. https://doi.org/10.1006/ijhc.1997.0134

Hinduja, S., & Patchin, J. W. (2010). Bullying, Cyberbullying, and Suicide. *Archives of Suicide Research*, *14*(3), 206–221. https://doi.org/10.1080/13811118.2010.494133

Hosseinmardi, H., Ghasemianlangroodi, A., Han, R., Lv, Q., & Mishra, S. (2014, August 21). *Towards Understanding Cyberbullying Behavior in a Semi-Anonymous Social Network* [Paper presentation]. 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014). https://arxiv.org/pdf/1404.3839.pdf

Huberman, B., Romero, D. M., & Wu, F. (2008). Social networks that matter: Twitter under the Microscope. *First Monday*. https://doi.org/10.5210/fm.v14i1.2317

Hughes, D. J., Rowe, M., Batey, M., & Lee, A. (2012). A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior*, *28*(2), 561–569. https://doi.org/10.1016/j.chb.2011.11.001

Irwin, J. R., & McClelland, G. H. (2001). Misleading heuristics and moderated multiple regression models. *Journal of Marketing Research*, *38*, 100 –109. https://doi.org/10.1509/jmkr.38.1.100.18835

Jakobwitz, S., & Egan, V. (2006). The dark triad and normal personality traits. *Personality and Individual Differences*, *40*(2), 331–339. https://doi.org/10.1016/j.paid.2005.07.006

Jang, Y. J., Kim, H. W., & Jung, Y. (2016). A mixed methods approach to the posting of benevolent comments online. *International Journal of Information Management*, *36*(3), 414-424. https://doi.org/10.1016/j.ijinfomgt.2016.02.001

John, O. P., & Soto, C. J. (2007). The importance of being valid. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 461-494). New York, NY: Guilford.

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

Joinson, A. N. (2007). Disinhibition and the Internet. In J. Gackenbach (Ed.), *Psychology and the Internet: Intrapersonal, interpersonal, and transpersonal implications* (2nd ed., pp. 75-92). San Diego, CA: Academic Press.

Jonason, P. K., & Webster, G. D. (2010). The dirty dozen: A concise measure of the dark triad. *Psychological Assessment*, *22*(2), 420–432. https://doi.org/10.1037/a0019265

Jones, S. E., Miller, J. D., & Lynam, D. R. (2011). Personality, antisocial behavior, and aggression: A meta-analytic review. *Journal of Criminal Justice*, *39*(4), 329–337. https://doi.org/10.1016/j.jcrimjus.2011.03.004

Jones, J. C., & Scott, S. (2012). Cyberbullying in the university classroom: A multiplicity of issues. In *Misbehavior Online in Higher Education* (Vol. 5, pp. 157-182). Emerald Group Publishing Limited.

Jung, Y.-E., Leventhal, B., Kim, Y. S., Park, T. W., Lee, S.-H., Lee, M., Park, S. H., Yang, J.-C., Chung, Y.-C., Chung, S.-K., & Park, J.-I. (2014). Cyberbullying, Problematic Internet Use, and Psychopathologic Symptoms among Korean Youth. *Yonsei Medical Journal*, *55*(3), 826. https://doi.org/10.3349/ymj.2014.55.3.826

Kane, G. C., Alavi, M., Labianca, G., & Borgatti, S. P. (2014). What's different about social media networks? A framework and research agenda. *MIS quarterly*, *38*(1), 275-304. https://www.jstor.org/stable/26554878

Karl, K., Peluchette, J., & Schlaegel, C. (2010). Who's posting facebook faux pas?: A cross-cultural examination of personality differences. International Journal of Selection and Assessment, 18(2), 174;186174;186. https://doi.org/10.1111/j.1468-2389.2010.00499.x

Karmakar, S., & Das, S. (2021). Understanding the rise of Twitter-based cyberbullying due to COVID-19 through comprehensive statistical evaluation. *Proceedings of the Annual Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2021.309

Kim, M., Clark, S. L., Donnellan, M. B., & Burt, S. A. (2020). A multi-method investigation of the personality correlates of digital aggression. *Journal of Research in Personality*, *85*, 103923. https://doi.org/10.1016/j.jrp.2020.103923

Kim, M., & Burt, S. A. (2023). The Faceless: Anonymity and Personality in Digital Aggression on Twitter [Manuscript submitted for publication]. Department of Psychology, Michigan State University.

Kircaburun, K., Jonason, P. K., & Griffiths, M. D. (2018). The Dark Tetrad traits and problematic social media use: The mediating role of cyberbullying and cyberstalking. *Personality and Individual Differences*, *135*, 264–269. https://doi.org/10.1016/j.paid.2018.07.034

Kircaburun, K., Alhabash, S., Tosuntaş, Ş. B., & Griffiths, M. D. (2020). Uses and Gratifications of Problematic Social Media Use Among University Students: A Simultaneous Examination of the Big Five of Personality Traits, Social Media Platforms, and Social Media Use Motives. *International Journal of Mental Health and Addiction*, *18*(3), 525–547. https://doi.org/10.1007/s11469-018-9940-6

Kokkinos, C. M., & Antoniadou, N. (2019). Cyber-bullying and cyber-victimization among undergraduate student teachers through the lens of the General Aggression Model. *Computers in Human Behavior*, *98*, 59-68. https://doi.org/10/1016/j.chb.2019.04.007

Kokkinos, C. M., Antoniadou, N., Dalara, E., Koufogazou, A., & Papatziki, A. (2013). Cyber-Bullying, Personality and Coping among Pre-Adolescents: *International Journal of Cyber Behavior, Psychology and Learning*, *3*(4), 55–69. https://doi.org/10.4018/ijcbpl.2013100104

Kokkinos, C. M., Antoniadou, N., & Markos, A. (2014). Cyber-bullying: An investigation of the psychological profile of university student participants. *Journal of Applied Developmental Psychology*, *35*(3), 204–214. https://doi.org/10.1016/j.appdev.2014.04.001

Kokkinos, C. M., Baltzidis, E., & Xynogala, D. (2016). Prevalence and personality correlates of Facebook bullying among university undergraduates. *Computers in Human Behavior*, *55*, 840–850. https://doi.org/10.1016/j.chb.2015.10.017

Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D., & Graepel, T. (2014). Manifestations of user personality in website choice and behaviour on online social networks. *Machine learning*, *95*(3), 357-380. https://doi.org/10.1007/s10994-013-5415-y

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin*, *140*(4), 1073–1137. https://doi.org/10.1037/a0035618

Kowalski, R. M., & Limber, S. P. (2007). Electronic Bullying Among Middle School Students. *Journal of Adolescent Health*, *41*(6), S22–S30. https://doi.org/10.1016/j.jadohealth.2007.08.017

Kraft, E. (2006). Cyberbullying: A worldwide trend of misusing technology to harass others. *WIT Transactions on Information and Communication Technologies*, *36*, 155-166.

Langos, C. (2015). Cyberbullying: The Shades of Harm. *Psychiatry, Psychology and Law*, *22*(1), 106–123. https://doi.org/10.1080/13218719.2014.919643

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

Liu, C., & Sui, D. (2017). Exploring the Spatiotemporal Pattern of Cyberbullying with Yik Yak. *The Professional Geographer*, *69*(3), 412–423. https://doi.org/10.1080/00330124.2016.1252273

MacDonald, C. D., & Roberts-Pittman, B. (2010). Cyberbullying among college students: Prevalence and demographic differences. *Procedia - Social and Behavioral Sciences*, *9*, 2003–2009. https://doi.org/10.1016/j.sbspro.2010.12.436

Mark, G., & Ganzach, Y. (2014). Personality and Internet usage: A large-scale representative study of young adults. *Computers in Human Behavior*, *36*, 274-281. https://doi.org/10.1016/j.chb.2014.03.060

Martin, R., & Vieaux, L. (2016). The digital rage: How anger is expressed online. In G. Riva, B. K. Wiederhold, & P. Cipresso, (Eds.), *The Psychology of Social Networking: Identity and Relationships in Online Communities* (Vol. 2, pp. 117-127). Warsaw, Poland: De Gruyter Open.

Mason, K. L. (2008). Cyberbullying: A preliminary assessment for school personnel. Psychology in the Schools, 45, 323–348. https://doi.org/10.1002/pits.20301

McCullough, M. E., Bellah, C. G., Kilpatrick, S. D., & Johnson, J. L. (2001). Vengefulness: Relationships with Forgiveness, Rumination, Well-Being, and the Big Five. *Personality*

and Social Psychology Bulletin, *27*(5), 601–610. https://doi.org/10.1177/0146167201275008

McInroy, L. B., & Mishna, F. (2017). Cyberbullying on Online Gaming Platforms for Children and Youth. *Child and Adolescent Social Work Journal*, *34*(6), 597–607. https://doi.org/10.1007/s10560-017-0498-0

Mehari, K. R., Farrell, A. D., & Le, A.-T. H. (2014). Cyberbullying among adolescents: Measures in search of a construct. *Psychology of Violence*, *4*(4), 399–415. https://doi.org/10.1037/a0037521

Menesini, E., Nocentini, A., Palladino, B. E., Frisén, A., Berne, S., Ortega-Ruiz, R., Calmaestra, J., Scheithauer, H., Schultze-Krumbholz, A., Luik, P., Naruskov, K., Blaya, C., Berthaud, J., & Smith, P. K. (2012). Cyberbullying Definition Among Adolescents: A Comparison Across Six European Countries. *Cyberpsychology, Behavior, and Social Networking*, *15*(9), 455–463. https://doi.org/10.1089/cyber.2012.0040

Meta. (2023). *Names allowed on Facebook*. Retrieved June 13, 2023, from https://www.facebook.com/help/229715077154790

Milam, A. C., Spitzmueller, C., & Penney, L. M. (2009). Investigating individual differences among targets of workplace incivility. *Journal of occupational health psychology*, *14*(1), 58-69. https://doi.org/10.1037/a0012683

Miller, J. D., Dir, A., Gentile, B., Wilson, L., Pryor, L. R., & Campbell, W. K. (2010). Searching for a Vulnerable Dark Triad: Comparing Factor 2 Psychopathy, Vulnerable Narcissism, and Borderline Personality Disorder: Vulnerable Dark Triad. *Journal of Personality*, *78*(5), 1529–1564. https://doi.org/10.1111/j.1467-6494.2010.00660.x

Miller, J. D., Lyman, D. R., Widiger, T. A., & Leukefeld, C. (2001). Personality Disorders as Extreme Variants of Common Personality Dimensions: Can the Five Factor Model Adequately Represent Psychopathy? *Journal of Personality*, *69*(2), 253–276. https://doi.org/10.1111/1467-6494.00144

Miller, J. D., & Lynam, D. (2001). Structural models of personality and their relation to antisocial behavior: A meta-analytic review. *Criminology*, *39*(4), 765–798. https://doi.org/10.1111/j.1745-9125.2001.tb00940.x

Miller, J. D., Lynam, D. R., & Jones, S. (2008). Externalizing Behavior Through the Lens of the Five-Factor Model: A Focus on Agreeableness and Conscientiousness. *Journal of Personality Assessment*, *90*(2), 158–164. https://doi.org/10.1080/00223890701845245

Miller, J. D., Zeichner, A., & Wilson, L. F. (2012). Personality correlates of aggression: Evidence from measures of the five-factor model, UPPS model of impulsivity, and BIS/BAS. *Journal of interpersonal violence*, *27*(14), 2903-2919. https://doi.org/10.1177/0886260512438279

Mishna, F., Saini, M., & Solomon, S. (2009). Ongoing and online: Children and youth's perceptions of cyber bullying. *Children and Youth Services Review*, *31*(12), 1222–1228. https://doi.org/10.1016/j.childyouth.2009.05.004

Mitchell, K. J., Ybarra, M., & Finkelhor, D. (2007). The Relative Importance of Online Victimization in Understanding Depression, Delinquency, and Substance Use. *Child Maltreatment*, *12*(4), 314–324. https://doi.org/10.1177/1077559507305996

Molero, M. M., Martos, Á., Barragán, A. B., Pérez-Fuentes, M. C., & Gázquez, J. J. (2022). Anxiety and depression from Cybervictimization in adolescents: A metaanalysis and meta-regression study. *The European Journal of Psychology Applied to Legal Context*, *14*(1), 42–50. https://doi.org/10.5093/ejpalc2022a5

Mondal, M., Silva, L. A., Correa, D., & Benevenuto, F. (2018). Characterizing usage of explicit hate expressions in social media. *New Review of Hypermedia and Multimedia*, *24*(2), 110–130. https://doi.org/10.1080/13614568.2018.1489001

Moore, M. J., Nakano, T., Enomoto, A., & Suda, T. (2012). Anonymity and roles associated with aggressive posts in an online forum. *Computers in Human Behavior*, *28*(3), 861–867. https://doi.org/10.1016/j.chb.2011.12.005

Na, H., Dancy, B. L., & Park, C. (2015). College student engaging in cyberbullying victimization: Cognitive appraisals, coping strategies, and psychological adjustments. *Archives of psychiatric nursing*, *29*(3), 155-161. https://doi.org/10.1016/j.apnu.2015.01.008

Nabi, R. L., Prestin, A., & So, J. (2013). Facebook friends with (health) benefits? Exploring social network site use and perceptions of social support, stress, and well-being. *Cyberpsychology, behavior, and social networking*, *16*(10), 721-727. https://doi.org/10.1089/cyber.2012.0521

Nocentini, A., Calmaestra, J., Schultze-Krumbholz, A., Scheithauer, H., Ortega, R., & Menesini, E. (2010). Cyberbullying: Labels, Behaviours and Definition in Three European Countries. *Australian Journal of Guidance and Counselling*, *20*(2), 129–142. https://doi.org/10.1375/ajgc.20.2.129

Ooi, K.-B., Lee, V.-H., Hew, J.-J., & Lin, B. (2019). Mobile Social Cyberbullying: Why are Keyboard Warriors Raging? *Journal of Computer Information Systems*, 1–12. https://doi.org/10.1080/08874417.2019.1679685

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. https://doi.org/10.1016/j.paid.2014.11.015

Park, S., Na, E.-Y., & Kim, E. (2014). The relationship between online activities, netiquette and cyberbullying. *Children and Youth Services Review*, *42*, 74–81. https://doi.org/10.1016/j.childyouth.2014.04.002

Patchin, J. W. (2021, June 1). *2021 cyberbullying data*. Cyberbullying Research Center. Retrieved March 11, 2023, from https://cyberbullying.org/2021-cyberbullying-data

Patchin, J. W. (2022, June 22). *Summary of Our Cyberbullying Research (2007-2021)*. Cyberbullying Research Center. https://cyberbullying.org/summary-of-our-cyberbullying-research

Patchin, J. W., & Hinduja, S. (2006). Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying. *Youth Violence and Juvenile Justice*, *4*(2), 148–169. https://doi.org/10.1177/1541204006286288

Patchin, J. W., & Hinduja, S. (2010). Cyberbullying and Self-Esteem. *Journal of School Health*, *80*(12), 614–621. https://doi.org/10.1111/j.1746-1561.2010.00548.x

Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*(6), 556–563. https://doi.org/10.1016/S0092-6566(02)00505-6

Peddinti, S. T., Ross, K. W., & Cappos, J. (2014). "on the internet, nobody knows you're a dog". *Proceedings of the Second ACM Conference on Online Social Networks*. https://doi.org/10.1145/2660460.2660467

Peddinti, S. T., Ross, K. W., & Cappos, J. (2017, February 1). *Mining anonymity: Identifying sensitive accounts on Twitter*. arXiv.org. Retrieved December 5, 2022, from http://arxiv.org/abs/1702.00164

Pellegrini, A. D., & Bartini, M. (2000). An empirical comparison of methods of sampling aggression and victimization in school settings. *Journal of Educational Psychology, 92*(2), 360-366. https://doi.org/10.1037/0022-0663.92.2.360

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. Austin, TX: University of Texas at Austin.

Perren, S., Dooley, J., Shaw, T., & Cross, D. (2010). Bullying in school and cyberspace: Associations with depressive symptoms in Swiss and Australian adolescents. *Child and Adolescent Psychiatry and Mental Health*, *4*(1), 28. https://doi.org/10.1186/1753-2000-4-28

Pew Research Center. (2021, January 13). *The State of Online Harassment*. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment

Pew Research Center. (2021, April 7). *Social Media Use in 2021*. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021

Postmes, T., Spears, R., & Lea, M. (1998). Breaching or building social boundaries? SIDE-effects of computer-mediated communication. *Communication research*, *25*(6), 689-715.

Quercia, D., Lambiotte, R., Stillwell, D., Kosinski, M., & Crowcroft, J. (2012). The personality of popular Facebook users. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (pp. 955-964). https://doi.org/10.1145/2145204.2145346

Raskauskas, J., & Stoltz, A.D. (2007). Involvement in traditional and electronic bullying among adolescents, Developmental Psychology, 43(3), 564-575. https://doi.org/10.1037/0012-1649.43.3.564

Reason, L., Boyd, M., & Reason, C. (2016). Cyberbullying in rural communities: Origin and processing through the lens of older adolescents. *The Qualitative Report*, *21*(12), 2331-2348.

Rost, K., Stahel, L., & Frey, B. S. (2016). Digital Social Norm Enforcement: Online Firestorms in Social Media. *PLOS ONE*, *11*(6), e0155923. https://doi.org/10.1371/journal.pone.0155923

Runions, K. C., & Bak, M. (2015). Online Moral Disengagement, Cyberbullying, and Cyber-Aggression. *Cyberpsychology, Behavior, and Social Networking*, *18*(7), 400–405. https://doi.org/10.1089/cyber.2014.0670

Rutter, M., Dunn, J., Plomin, R., Simonoff, E., Pickles, A., Maughan, B., Ormel, J., Meyer, J. & Eaves, L. (1997). Integrating nature and nurture: Implications of person–environment correlations and interactions for developmental psychopathology. *Development and psychopathology*, *9*(2), 335-364. https://doi.org/10.1017/S0954579497002083

Samghabadi, N. S, Maharjan, S., Sprague, A., Diaz-Sprague, R., & Solorio, T. (2017). Detecting Nastiness in Social Media. *Proceedings of the First Workshop on Abusive Language Online*, 63–72. https://doi.org/10.18653/v1/W17-3010

Samoh, N., Boonmongkon, P., Ojanen, T. T., Samakkeekarom, R., Jonas, K. J., & Guadamuz, T. E. (2019). *'It's an ordinary matter'*: Perceptions of cyberbullying in Thai youth culture. *Journal of Youth Studies*, *22*(2), 240–255. https://doi.org/10.1080/13676261.2018.1495835

Saud, M., Mashud, M. I., & Ida, R. (2020). Usage of social media during the pandemic: Seeking support and awareness about COVID-19 through social media platforms. *Journal of Public Affairs*, *20*(4), e2417. https://doi.org/10.1002/pa.2417

Savage, M. W., & Tokunaga, R. S. (2017). Moving toward a theory: Testing an integrated model of cyberbullying perpetration, aggression, social skills, and Internet self-efficacy. *Computers in Human Behavior*, *71*, 353–361. https://doi.org/10.1016/j.chb.2017.02.016

Schenk, A. M., & Fremouw, W. J. (2012). Prevalence, psychological impact, and coping of cyberbully victims among college students. *Journal of school violence*, *11*(1), 21-37. https://doi.org/10.1080/15388220.2011.630310

Schenk, A. M., Fremouw, W. J., & Keelan, C. M. (2013). Characteristics of college cyberbullies. *Computers in Human Behavior*, *29*(6), 2320–2327. https://doi.org/10.1016/j.chb.2013.05.013

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, *8*(9). https://doi.org/10.1371/journal.pone.0073791

Seidman, G. (2013). Self-presentation and belonging on Facebook: How personality influences social media use and motivations. *Personality and individual differences*, *54*(3), 402-407. https://doi.org/10.1016/j.paid.2012.10.009

Seigfried-Spellar, K. C., & Lankford, C. M. (2018). Personality and online environment factors differ for posters, trolls, lurkers, and confessors on Yik Yak. *Personality and Individual Differences*, *124*, 54–56. https://doi.org/10.1016/j.paid.2017.11.047

Selkie, E. M., Kota, R., Chan, Y.-F., & Moreno, M. (2015). Cyberbullying, Depression, and Problem Alcohol Use in Female College Students: A Multisite Study. *Cyberpsychology, Behavior, and Social Networking*, *18*(2), 79–86. https://doi.org/10.1089/cyber.2014.0371

Siegel, J., Dubrovsky, V., Kiesler, S., & McGuire, T. W. (1986). Group processes in computer-mediated communication. *Organizational behavior and human decision processes*, *37*(2), 157-187.

Slonje, R., & Smith, P. K. (2008). Cyberbullying: Another main type of bullying? *Scandinavian Journal of Psychology*, *49*(2), 147–154. https://doi.org/10.1111/j.1467-9450.2007.00611.x

Slonje, R., Smith, P. K., & Frisén, A. (2013). The nature of cyberbullying, and strategies for prevention. *Computers in Human Behavior*, *29*(1), 26–32. https://doi.org/10.1016/j.chb.2012.05.024

Snapchat. (2023). *When does Snapchat delete Snaps and Chats?*. Retrieved April 1, 2023, from https://help.snapchat.com/hc/en-us/articles/7012334940948-When-does-Snapchat-delete-Snaps-and-Chats-

Soto, C. J., & John, O. P. (2017). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality*, *68*, 69–81. https://doi.org/10.1016/j.jrp.2017.02.004

Spears, R., & Lea, M. (1992). Social influence and the influence of the "social" in computer-mediated communication. In M. Lea (Ed.), *Contexts of computer-mediated communication* (pp. 30-65). London: HarvesterWheatsheaf.

Sproull, L., & Kiesler, S. (1986). Reducing Social Context Cues: Electronic Mail in Organizational Communication. *Management Science*, *32*(11), 1492–1512. https://doi.org/10.1287/mnsc.32.11.1492

Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology, 55,* 167-194. https://doi.org/10.1111/j.1744-6570.2022.tb00108.x

Statista. (2022, January). *Daily time spent on social networking by internet users worldwide4 from 2012 to 2022.* https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide

Sticca, F., & Perren, S. (2013). Is Cyberbullying Worse than Traditional Bullying? Examining the Differential Roles of Medium, Publicity, and Anonymity for the Perceived Severity of Bullying. *Journal of Youth and Adolescence*, *42*(5), 739–750. https://doi.org/10.1007/s10964-012-9867-3

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior*, *7*(3), 321-326.

Suls, J., & Martin, R. (2005). The Daily Life of the Garden-Variety Neurotic: Reactivity, Stressor Exposure, Mood Spillover, and Maladaptive Coping. *Journal of Personality*, *73*(6), 1485–1510. https://doi.org/10.1111/j.1467-6494.2005.00356.x

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Taylor, S. P. (1967). Aggressive behavior and physiological arousal as a function of provocation and the tendency to inhibit aggression. *Journal of Personality*, *35*(2), 297-310. https://doi.org/10.1111/j.1467-6494.1967.tb01430.x

Twitter. (2021). *Privacy Policy: Basic Account Information*. Retrieved April 24, 2021, from https://twitter.com/en/privacy

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*, 231–235. https://doi.org/10.1016/j.paid.2016.10.063

Vandebosch, H., & Van Cleemput, K. (2008). Defining Cyberbullying: A Qualitative Research into the Perceptions of Youngsters. *CyberPsychology & Behavior*, *11*(4), 499–503. https://doi.org/10.1089/cpb.2007.0042

Vannucci, M., Nocentini, A., Mazzoni, G., & Menesini, E. (2012). Recalling unpresented hostile words: False memories predictors of traditional and cyberbullying. *European Journal of Developmental Psychology*, *9*(2), 182-194. https://doi.org/10.1080/17405629.2011.646459

Vernon, P. A., Villani, V. C., Vickers, L. C., & Harris, J. A. (2008). A behavioral genetic investigation of the Dark Triad and the Big 5. *Personality and Individual Differences*, *44*(2), 445–452. https://doi.org/10.1016/j.paid.2007.09.007

Veselka, L., Schermer, J. A., & Vernon, P. A. (2012). The Dark Triad and an expanded framework of personality. *Personality and Individual Differences*, *53*(4), 417–425. https://doi.org/10.1016/j.paid.2012.01.002

Villanti, 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

Vize, C. E., Miller, J. D., & Lynam, D. R. (2018). FFM facets and their relations with different forms of antisocial behavior: An expanded meta-analysis. *Journal of Criminal Justice*, *57*, 67-75. https://doi.org/10.1016/j.jcrimjus.2018.04.004

Waasdorp, T. E., & Bradshaw, C. P. (2015). The Overlap Between Cyberbullying and Traditional Bullying. *Journal of Adolescent Health*, *56*(5), 483–488. https://doi.org/10.1016/j.jadohealth.2014.12.002

Wang, J., Iannotti, R. J., & Nansel, T. R. (2009). School Bullying Among Adolescents in the United States: Physical, Verbal, Relational, and Cyber. *Journal of Adolescent Health*, *45*(4), 368–375. https://doi.org/10.1016/j.jadohealth.2009.03.021

Wang, J., Nansel, T. R., & Iannotti, R. J. (2011). Cyber and Traditional Bullying: Differential Association With Depression. *Journal of Adolescent Health*, *48*(4), 415–417. https://doi.org/10.1016/j.jadohealth.2010.07.012

Whittaker, E., & Kowalski, R. M. (2015). Cyberbullying Via Social Media. *Journal of School Violence*, *14*(1), 11–29. https://doi.org/10.1080/15388220.2014.949377

Wigderson, S., & Lynch, M. (2013). Cyber- and traditional peer victimization: Unique relationships with adolescent well-being. *Psychology of Violence*, *3*(4), 297–309. https://doi.org/10.1037/a0033657

Wong, A., Ho, S., Olusanya, O., Antonini, M. V., & Lyness, D. (2021). The use of social media and online communications in times of pandemic COVID-19. *Journal of the Intensive Care Society*, *22*(3), 255-260. https://doi.org/10.1177/1751143720966280

Wright, M. F. (2013). The Relationship Between Young Adults' Beliefs About Anonymity and Subsequent Cyber Aggression. *Cyberpsychology, Behavior, and Social Networking*, *16*(12), 858–862. https://doi.org/10.1089/cyber.2013.0009

Ybarra, M. L., Boyd, D., Korchmaros, J. D., & Oppenheim, J. (2012). Defining and Measuring Cyberbullying Within the Larger Context of Bullying Victimization. *Journal of Adolescent Health*, *51*(1), 53–58. https://doi.org/10.1016/j.jadohealth.2011.12.031

Ybarra, M. L., Diener-West, M., & Leaf, P. J. (2007). Examining the Overlap in Internet Harassment and School Bullying: Implications for School Intervention. *Journal of Adolescent Health*, *41*(6), S42–S50. https://doi.org/10.1016/j.jadohealth.2007.09.004

Ybarra, M. L., & Mitchell, K. J. (2004). Online aggressor/targets, aggressors, and targets: A comparison of associated youth characteristics. *Journal of Child Psychology and Psychiatry*, *45*(7), 1308–1316. https://doi.org/10.1111/j.1469-7610.2004.00328.x

You, S., & Lim, S. A. (2016). Longitudinal predictors of cyberbullying perpetration: Evidence from Korean middle school students. *Personality and Individual Differences*, *89*, 172–176. https://doi.org/10.1016/j.paid.2015.10.019

Zezulka, L., & Seigfried-Spellar, K. (2016). Differentiating Cyberbullies and Internet Trolls by Personality Characteristics and Self-Esteem. *Journal of Digital Forensics, Security and Law*. https://doi.org/10.15394/jdfsl.2016.1415

Zhang, S., Su, W., Han, X., & Potenza, M. N. (2022). Rich Get Richer: Extraversion Statistically Predicts Reduced Internet Addiction through Less Online Anonymity Preference and Extraversion Compensation. *Behavioral Sciences*, *12*(6), 193. https://doi.org/10.3390/bs12060193

APPENDIX A:

TABLES

Table 1. Research questions and hypotheses.

Research Question	Hypothesis
RQ1: What is the effect of the technical self-anonymity and personality on DA on Twitter?	
RQ1a: What personality traits predict self-anonymity on Twitter?	**H1a**: Individuals low in extraversion and high intellect/imagination will be more likely to use anonymous Twitter bios.
RQ1b: What personality traits predict DA on Twitter?	**H1b**: Individuals low in conscientiousness and high in intellect/imagination will engage in higher levels of DA on Twitter.
RQ1c: Does DA on Twitter differ across individuals with anonymous versus identifiable Twitter bio accounts?	**H1c**: Individuals with anonymous Twitter bios will engage in higher levels of DA on Twitter than those with identifiable Twitter bios.
RQ1d: Which, if any, personality traits moderate the relationship between technical self-anonymity and DA on Twitter?	**H1d**: Individuals with low agreeableness and high intellect/imagination will engage in higher levels of DA in the presence of anonymity.
RQ2: What is the effect of the perpetrator self- and other- anonymity on DA assessed experimentally?	
RQ2a: What personality traits predict DA on the TAP-Chat?	**H2a**: Individuals with low agreeableness and high open-mindedness will engage in higher levels of DA on the TAP-Chat.
RQ2b: Does DA differ across experimentally manipulated anonymity conditions?	**H2b**: Individuals in the self-anonymous conditions will engage in higher DA on the TAP-Chat than those in the self-identified conditions. Individuals in the other-identified conditions will engage in higher DA on the TAP-Chat than those in the other-anonymous conditions.
RQ2c: Do experimentally manipulated anonymity conditions moderate the relationship between personality traits and DA?	**H2c**: Individuals with low agreeableness and high negative emotionality will engage in higher levels of DA on the TAP-Chat in the presence of either anonymity (self or other).

Table 2. Twitter anonymity coding scheme.

0 – No Anonymity (You can clearly identify who the account belongs to)	• Full name • Profile picture • Banner includes the user • Location and/or university → check for identifiable information in the bio such as major, graduation year, or hometown • Links to personal websites or Facebook, Instagram, or other social medias
1 – Partial Anonymity (You can guess who the account belongs to)	• Full name (or first name only) • Group profile picture ○ If profile picture is not a group image but also not of the user (i.e., a picture of a cartoon or actor/actress), look at the accounts images to see if the user has any pictures of themselves • Banner includes the user • Location and/or university → check for identifiable information in the bio such as major, graduation year, or hometown
2 – Complete Anonymity (No information on who the account belongs to)	• First name only (or nicknames) • No profile picture (i.e., pictures of scenery, quotes, etc.,) ○ No images of self • Bio has no identifiable information such as those listed above or links to other social medias

Table 3a. Descriptive statistics and zero-order correlations for Sample 1.

	M	*SD*	1	2	3	4	5	6	7	8
1. Twitter AWI Time 1	8.23	5.65	—							
2. Twitter AWI Time 2	9.52	7.51	**.47****	—						
3. Continuous anonymity	.39	0.48	**.32****	**.21****	—					
4. Recoded anonymity	.36	0.48	**.34****	**.22****	**.87****	—				
5. Extraversion	0	7.79	**–.12***	–.07	**–.16****	**–.14****	—			
6. Emotional Stability	0	7.34	**–.15****	**–.11***	**–.10***	**–.10***	**.27****	—		
7. Agreeableness	0	5.20	**–.10***	–.04	**–.14****	**–.11***	**.17****	.03	—	
8. Conscientiousness	0	6.39	**–.09***	–.08	.08	–.04	–.01	**.19****	**.18****	—
9. Intellect/Imagination	0	5.37	.04	.07	.00	.06	**.29****	.06	**.28****	**.13****

Table 3b. Descriptive statistics and zero-order correlations for Sample 2.

	M	*SD*	1	2	3	4	5	6	7
1. Twitter AWI	7.72	7.47	—						
2. Continuous anonymity	.70	.60	.06	—					
3. Recoded anonymity	.59	.49	.07	**.74****	—				
4. Extraversion	0	8.35	–.05	**–.19****	**–.18****	—			
5. Emotional Stability	0	7.43	–.06	–.01	.00	**.17****	—		
6. Agreeableness	0	5.66	–.05	**–.11***	**–.11***	**.25****	–.08	—	
7. Conscientiousness	0	6.56	–.04	**–.14****	**–.11***	.08	**.21****	**.18****	—
8. Intellect/Imagination	0	5.97	.09	.04	–.03	**.23****	.01	**.34****	**.20****

Note: Recoded anonymity was dichotomized 0 for identifiable/non-anonymous Twitter bios and 1 for partially-to-completely anonymous Twitter bios. Personality variables were mean-centered prior to analyses. Correlations significant at $p < .05$ are highlighted in bold.
Abbreviation: AWI, antisocial word index.
$^{*}p < .05$. $^{**}p < .01$.

Table 4a. Multiple regression results for analysis predicting Twitter anonymity as a function of personality traits for Sample 1.

	b	*SE*	*β*	R^2
Intercept	.387	.022	—	.053**
Extraversion	–.009	.003	**–.152****	
Emotional Stability	–.003	.003	–.046	
Agreeableness	–.012	.004	**–.131****	
Conscientiousness	–.005	.004	–.064	
Intellect/Imagination	.009	.004	**.094***	

Note: *$p < .05$. **$p < .01$.

Table 4b. Logistic regression results for analysis predicting Twitter anonymity as a function of personality traits for Sample 1.

	β	*SE β*	Wald's χ^2	Odds ratio (e^{β})
Intercept	**–.596****	.098	36.755	.551
Extraversion	**–.041****	.014	8.897	.960
Emotional Stability	–.017	.014	1.443	.983
Agreeableness	**–.049***	.020	5.987	.952
Conscientiousness	–.009	.016	.291	.991
Intellect/Imagination	**.057****	.020	7.964	1.059

Note: *$p < .05$. **$p < .01$.

Table 4c. Multiple regression results for analysis predicting Twitter anonymity as a function of personality traits for Sample 2.

	b	*SE*	*β*	R^2
Intercept	.702	.027	—	.071**
Extraversion	–.014	.004	**–.193****	
Emotional Stability	.003	.004	.042	
Agreeableness	–.009	.004	–.087	
Conscientiousness	–.013	.004	**–.144****	
Intellect/Imagination	.014	.005	**.144****	

Note: *$p < .05$. **$p < .01$.

Table 4d. Logistic regression results for analysis predicting Twitter anonymity as a function of personality traits for Sample 2.

	β	*SE* β	Wald's χ^2	Odds ratio (e^{β})
Intercept	**.388****	.099	15.482	1.475
Extraversion	**–.045****	.013	11.885	.956
Emotional Stability	.015	.014	1.220	1.016
Agreeableness	–.022	.020	1.313	.978
Conscientiousness	**–.033***	.016	4.357	.968
Intellect/Imagination	.019	.018	1.034	1.019

Note: $^{*}p < .05$. $^{**}p < .01$.

Table 5a. Mean differences in Twitter DA for identifiable and anonymous Twitter bio accounts for Sample 1.

	n	*M*	*SD*	*t*	*df*	*p*	*d*
Time 1							
Identifiable / Anonymous	292 / 166	6.777 / 10.792	4.410 / 6.615	–6.987	249.873	**<.001**	.714
Time 2							
Identifiable / Anonymous	274 / 157	8.276 / 11.697	7.076 / 7.773	–4.541	300.597	**<.001**	.460

Table 5b. Mean differences in Twitter DA for identifiable and anonymous Twitter bio accounts for Sample 2.

	n	*M*	*SD*	*t*	*df*	*p*	*d*
Identifiable / Anonymous	185 / 265	7.097 / 8.199	7.270 / 7.669	–1.532	448	.123	.148

Table 6a. Moderated regression results predicting Twitter AWI as a function of personality and anonymity for Sample 1 at Time 1.

	Recoded Anonymity				Continuous Anonymity			
Predictors	Step 1	Step 2	Step 3	Step 4	Step 1	Step 2	Step 3	Step 4
	Beta				Beta			
Age	.068	.076	.058	.060	.068	.076	.048	.058
Gender	.045	.073	.067	.045	.045	.073	.068	.055
Ethnicity/Race	.064	.052	.005	.008	.064	.052	.010	.006
Extraversion (EXT)		**–.108***	–.060	–.060		**–.108***	–.066	–.104
Emotional Stability (EMO)		–.095	–.079	–.014		–.095	–.085	–.013
Agreeableness (AGR)		**–.118***	–.085	.038		**–.118***	–.085	.016
Conscientiousness (CON)		–.071	–.067	–.108		–.071	–.052	–.113
Intellect/Imagination (INT)		**.126***	.081	.010		**.126***	**.099***	.047
Anonymity			**.306****	**.295****			**.277****	**.289****
EXT x Anonymity				.015				.074
EMO x Anonymity				**–.121***				**–.124***
AGR x Anonymity				**–.179****				**–.168****
CON x Anonymity				.084				**.125***
INT x Anonymity				**.111***				.092
R^2	.010	**.061****	**.148****	**.176****	.010	**.061****	**.131****	**.160****
ΔR^2		**.051****	**.087****	**.029****		**.051****	**.070****	**.029****

Note: $^{*}p < .05$. $^{**}p < .01$.

Table 6b. Moderated regression results predicting Twitter AWI as a function of personality and anonymity for Sample 1 at Time 2.

Predictors	Recoded Anonymity				Continuous Anonymity			
	Step 1	Step 2	Step 3	Step 4	Step 1	Step 2	Step 3	Step 4
	Beta				Beta			
Age	.031	.043	.034	.033	.031	.043	.027	.028
Gender	.076	**.110***	.105	.083	.076	**.110***	.103	.093
Ethnicity/Race	.039	.033	.007	.017	.039	.033	.008	.011
Extraversion (EXT)		–.090	–.061	–.009		–.090	–.061	–.027
Emotional Stability (EMO)		–.051	–.041	–.070		–.051	–.044	–.060
Agreeableness (AGR)		–.072	–.051	.000		–.072	–.051	–.031
Conscientiousness (CON)		–.086	–.083	**–.121***		–.086	–.077	–.103
Intellect/Imagination (INT)		**.144****	**.118***	.045		**.144****	**.127***	.080
Anonymity			**.192****	**.184****			**.188****	**.186****
EXT x Anonymity				–.083				–.052
EMO x Anonymity				.046				.025
AGR x Anonymity				–.068				–.032
CON x Anonymity				.059				.040
INT x Anonymity				**.123***				.078
R^2	.008	**.042***	**.076****	**.092****	.087	**.204***	**.273****	**.283****
ΔR^2		**.034***	**.034****	.016		**.034***	**.033****	.005

Note: $^{*}p < .05$. $^{**}p < .01$.

Table 6c. Moderated regression results predicting Twitter AWI as a function of personality and anonymity for Sample 2.

Predictors	Recoded Anonymity				Continuous Anonymity			
	Step 1	Step 2	Step 3	Step 4	Step 1	Step 2	Step 3	Step 4
	Beta				Beta			
Age	.013	.008	.007	.012	.013	.008	.009	.013
Gender	**.141****	**.144****	**.144****	**.142****	**.141****	**.144****	**.143****	**.150****
Ethnicity/Race	.090	.082	.075	.073	.090	.082	.080	.080
Extraversion (EXT)		–.031	–.024	–.073		–.031	–.028	–.085
Emotional Stability (EMO)		–.010	–.012	.025		–.010	–.011	–.006
Agreeableness (AGR)		**–.103***	–.101	–.054		**–.103***	–.102	–.078
Conscientiousness (CON)		–.050	–.045	–.029		–.050	–.047	–.048
Intellect/Imagination (INT)		**.145****	**.143****	.003		**.145****	**.142****	.070
Anonymity			.043	.042			.017	.021
EXT x Anonymity				.072				.084
EMO x Anonymity				–.045				–.009
AGR x Anonymity				–.049				–.024
CON x Anonymity				–.020				.008
INT x Anonymity				**.178***				.095
R^2	**.028****	**.052****	**.053****	**.068****	**.028****	**.052****	**.052****	**.059***
ΔR^2		.023	.002	.015		.023	.000	.007

Note: $^{*}p < .05$. $^{**}p < .01$.

Table 7. TAP digital aggression coding scheme.

		Examples
0	**Not aggressive at all**	hi, lol, haha trying my best!
1	**Very Low**	ha very funny, r u human, no reason to have beef
2	**Low**	harsh, yikes, *you're
3	**Moderate**	your mom, you really are 12 with these horrible disses, wow that was rude…
4	**High**	who raised you to be so mean???, why so mad about a damn game chill, u suck
5	**Very High**	bitch, we're pressing a fucking button…, dude I am on some old ass computer right now from the 90's trust me with this lag I won't win any

Table 8. Sample Size for each of the anonymity conditions on the TAP-Chat.

		Other-Identified		
		No	Yes	Total
Self-Identified	No	180	162	342
	Yes	98	113	211
	Total	289	275	553

Note: Initial sample size **before** reviewing uploaded pictures by participants assigned to Self-Identified conditions.

		Other-Identified		
		No	Yes	Total
Self-Identified	No	209	211	420
	Yes	69	64	133
	Total	289	275	553

Note: Final sample size **after** reviewing uploaded pictures by participants assigned to Self-Identified conditions.

Table 9. Descriptive statistics and zero-order correlations.

	M	*SD*	1	2	3	4	5	6	7	8
1. TAP-Chat AWI	.62	.83	—							
2. Average TAP-Chat	.74	.73	**.43****	—						
3. Self-ID Condition	.24	.43	.05	.02	—					
4. Other-ID Condition	.50	.50	–.03	.02	–.02	—				
5. Extraversion	0	.92	–.00	.00	.06	.01	—			
6. Agreeableness	0	.78	**–.10***	**–.10***	–.06	–.02	**–.09***	—		
7. Conscientiousness	0	.86	–.03	.01	–.07	–.00	**.21****	**.31****	—	
8. Negative Emotionality	0	.95	.08	–.02	–.02	.01	**–.34****	–.01	**–.37****	—
9. Open-mindedness	0	.75	.07	–.06	**.09***	–.03	.05	**.20****	**.11***	.06

Note: Average TAP-Chat is the average ratings of digital aggression on TAP-Chat across the four coders. Self-ID Condition: 0 for participants who were self-anonymous and 1 for who were self-identified on TAP-Chat. Other-ID Condition: 0 for participants who interacted with bot-anonymous and 1 for those who interacted with bot-identified. Personality variables were mean-centered prior to analyses. Correlations significant at $p < .05$ are highlighted in bold.
Abbreviation: AWI, antisocial word index.
$^{*}p < .05$. $^{**}p < .01$.

Table 10a. Ordinal logistic regression results for analysis predicting TAP-Chat AWI as a function of personality traits.

	Odds Ratios	*SE*	Wald's χ^2	95% Confidence Interval for Odds Ratios	
				Lower	Upper
Extraversion	.987	.110	.013	.796	1.225
Agreeableness	**.712****	.128	7.004	.554	.916
Conscientiousness	1.082	.125	.402	.848	1.381
Negative Emotionality	1.186	.114	2.239	.949	1.482
Open-mindedness	**1.326***	.131	4.631	1.025	1.714

Note: $^{*}p < .05$. $^{**}p < .01$.

Table 10b. Multiple regression results for analysis predicting Average TAP-Chat as a function of personality traits.

	b	*SE*	*β*	R^2
Intercept	.199	.435	—	**.044****
Extraversion	–.014	.039	–.018	
Agreeableness	–.130	.047	**–.141****	
Conscientiousness	.020	.045	.024	
Negative Emotionality	–.018	.040	–.024	
Open-mindedness	–.035	.047	–.036	

Note: Demographic variables were added as covariates in the model. $^{*}p < .05$. $^{**}p < .01$.

Table 11. Moderated regression results predicting TAP AWI as a function of personality and anonymity conditions.

Predictors / Anonymity Conditions (AC)	Self-identified, Other-anonymous				Self-anonymous, Other-identified				Fully identified			
	Step 1	Step 2	Step 3	Step 4	Step 1	Step 2	Step 3	Step 4	Step 1	Step 2	Step 3	Step 4
	Beta				Beta				Beta			
Age	–.044	–.025	–.025	–.021	–.044	–.025	–.026	–.025	–.044	–.025	–.025	–.038
Ethnicity/Race	–.057	–.065	–.064	–.063	–.057	–.065	–.066	–.062	–.057	–.065	–.064	–.071
Gender	.030	.035	.034	.033	.030	.035	.035	.037	.030	.035	.035	.033
EXT		.005	.004	.035		.005	.006	–.027		.005	.006	–.035
AGR		**–.132***	**–.130***	–.100		**–.132***	**–.131***	**–.131***		**–.132***	**–.133***	**–.158****
CON		.036	.031	.048		.036	.039	–.002		.036	.035	.020
NEM		.076	.074	.069		.076	.078	.074		.076	.075	.046
OPM		.083	.078	.081		.083	.083	.080		.083	.083	.099
Anonymity[a]			.065	.074			–.028	–.029			–.009	.014
EXT x AC				–.079				.047				**.132***
AGR x AC				–.050				–.003				.067
CON x AC				–.039				.072				.067
NEM x AC				.006				.000				.110
OPM x AC				–.008				.015				–.067
R^2	.006	.029	.034	.046	.006	.029	.030	.036	.006	.029	.030	.051
ΔR^2		.023	.004	.012		.023	.001	.006		.023	.000	.021

Note: $^{*}p < .05$. $^{**}p < .01$. Personality variables were mean-centered prior to analysis. Gender was dummy-coded 1 for female and 0 for all other gender groups. Self-identified ethnicity/race was dummy-coded 1 for White and 0 for all other ethnic and racial groups. [a]Anonymity condition was dummy coded with the fully anonymous condition as the reference group in each model. EXT=Extraversion; AGR=Agreeableness; NEM=Negative Emotionality; OPM = Open-mindedness.

Table 12. Moderated regression results predicting Average TAP coder ratings as a function of personality and anonymity conditions.

Predictors / Anonymity Conditions (AC)	Self-identified, Other-anonymous				Self-anonymous, Other-identified				Fully identified			
	Step 1	Step 2	Step 3	Step 4	Step 1	Step 2	Step 3	Step 4	Step 1	Step 2	Step 3	Step 4
	Beta				Beta				Beta			
Age	.039	.049	.049	.060	.039	.049	.050	.053	.039	.049	.048	.049
Ethnicity/Race	–.086	–.084	–.083	–.083	–.086	–.084	–.082	–.083	–.086	–.084	–.082	–.090
Gender	**.120***	**.151****	**.150****	**.143****	**.120***	**.151****	**.150****	**.149****	**.120***	**.151****	**.151****	**.149****
EXT		–.018	–.019	.012		–.018	–.019	–.060		–.018	–.017	–.030
AGR		**–.141****	**–.139****	**–.118***		**–.141****	**–.144****	**–.158***		**–.141****	**–.142****	**–.135***
CON		.024	.021	.028		.024	.020	.036		.024	.021	.006
NEM		–.024	–.026	.003		–.024	–.028	–.013		–.024	–.026	–.057
OPM		–.036	–.040	–.027		–.036	–.034	–.025		–.036	–.035	–.021
Anonymity[a]			.049	.064			.042	.044			–.022	–.012
EXT x AC				–.072				.063				.052
AGR x AC				–.053				.019				–.025
CON x AC				–.013				–.026				.062
NEM x AC				–.069				–.031				.109
OPM x AC				–.037				–.013				–.058
R^2	**.023***	**.044****	**.047***	**.056***	**.023***	**.044****	**.046***	.049	**.023***	**.044****	**.057***	**.054***
ΔR^2		.021	.002	.010		.021	.002	.003		.021	.000	.009

Note: $^{*}p < .05$. $^{**}p < .01$. Personality variables were mean-centered prior to analysis. Gender was dummy-coded 1 for female and 0 for all other gender groups. Self-identified ethnicity/race was dummy-coded 1 for White and 0 for all other ethnic and racial groups. [a]Anonymity condition was dummy coded with the fully anonymous condition as the reference group in each model. EXT=Extraversion; AGR=Agreeableness; NEM=Negative Emotionality; OPM = Open-mindedness.

APPENDIX B:

FIGURES

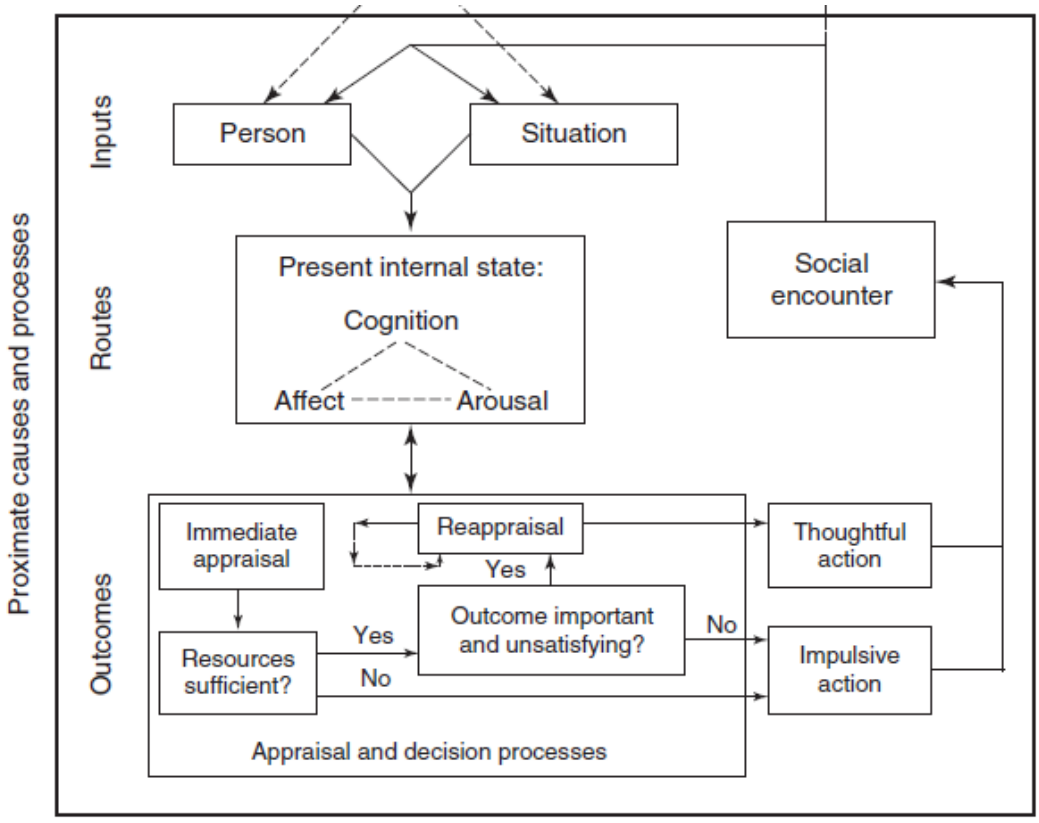


Figure 1. Overview of the General Aggression Model (GAM).

Note. The GAM has three different proximate causes and processes: inputs, routes, and outcomes. The current studies specifically focus on personality traits for the *person* inputs and (perceived) anonymity for the *situation* inputs. The diagram was reproduced from Allen and Anderson (2017).

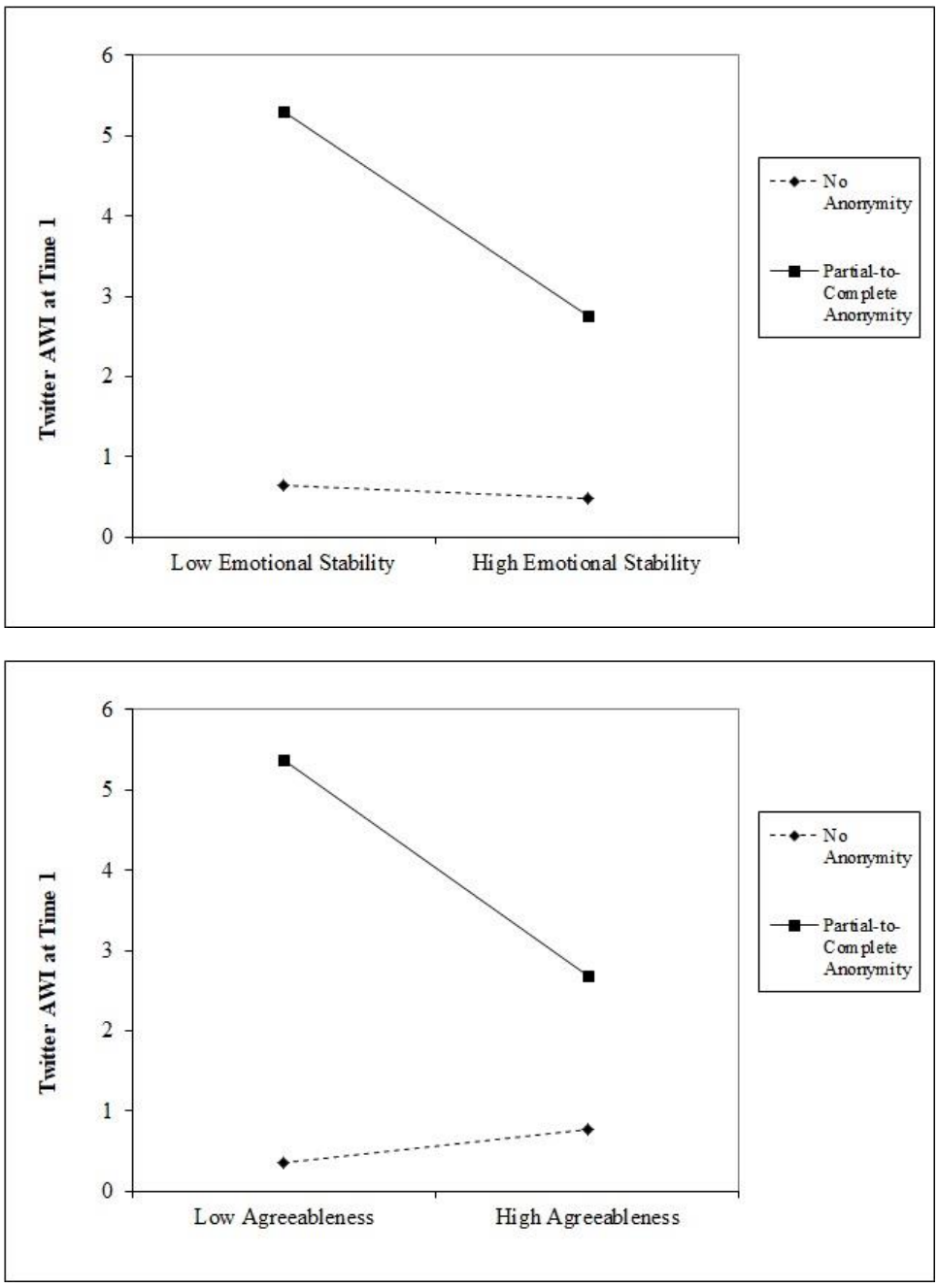


Figure 2. The moderating effects of anonymity on the association between Twitter AWI and (a) emotional stability, (b) agreeableness, and (c) intellect/imagination for Sample 1 at Time 1.

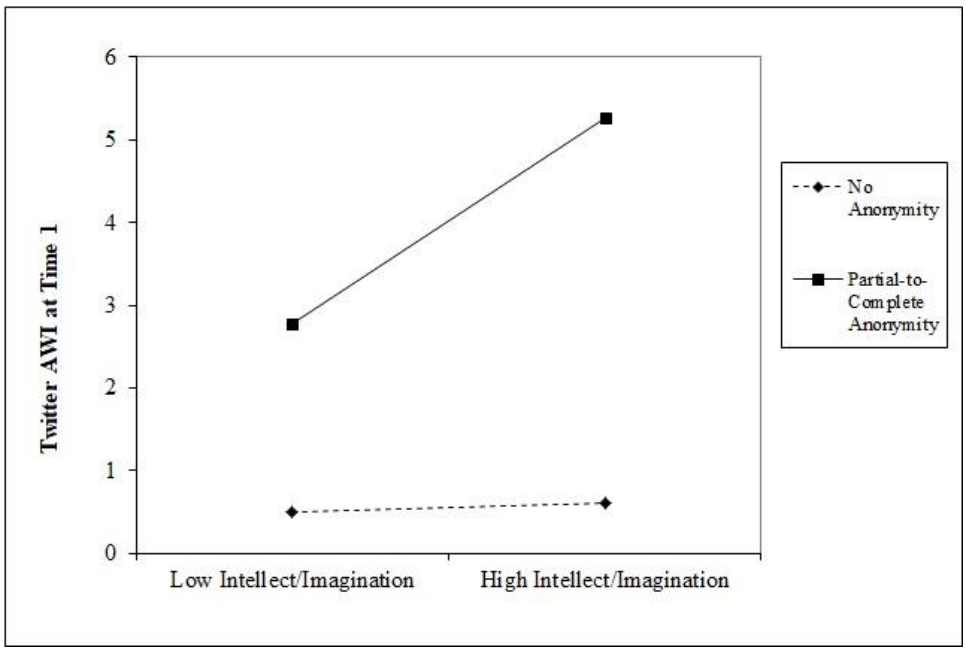


Figure 2 (cont'd).

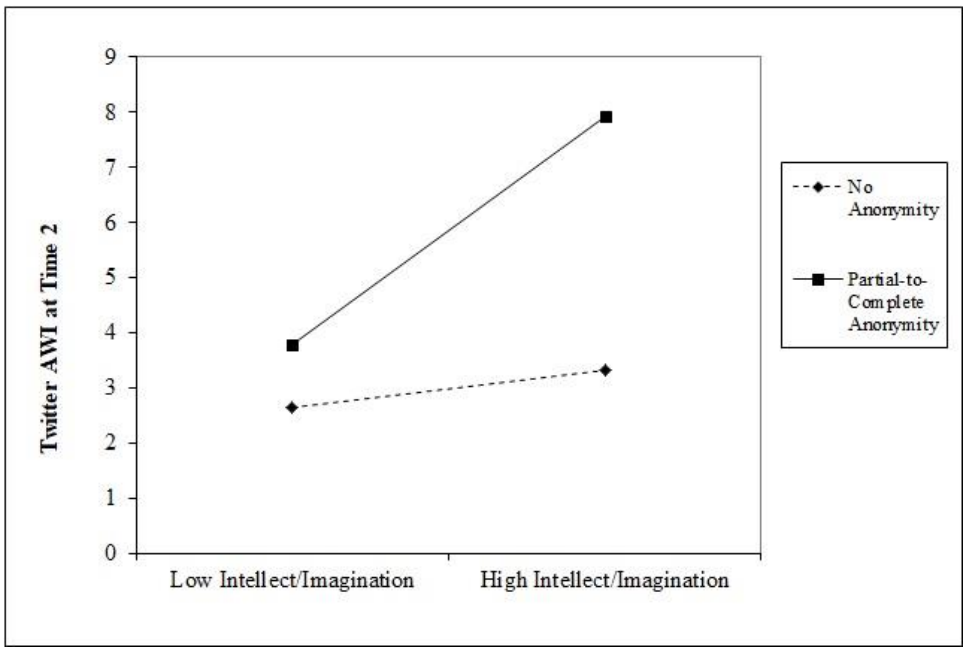


Figure 3. The moderating effects of anonymity on the association between Twitter AWI and intellect/imagination for Sample 1 at Time 2.

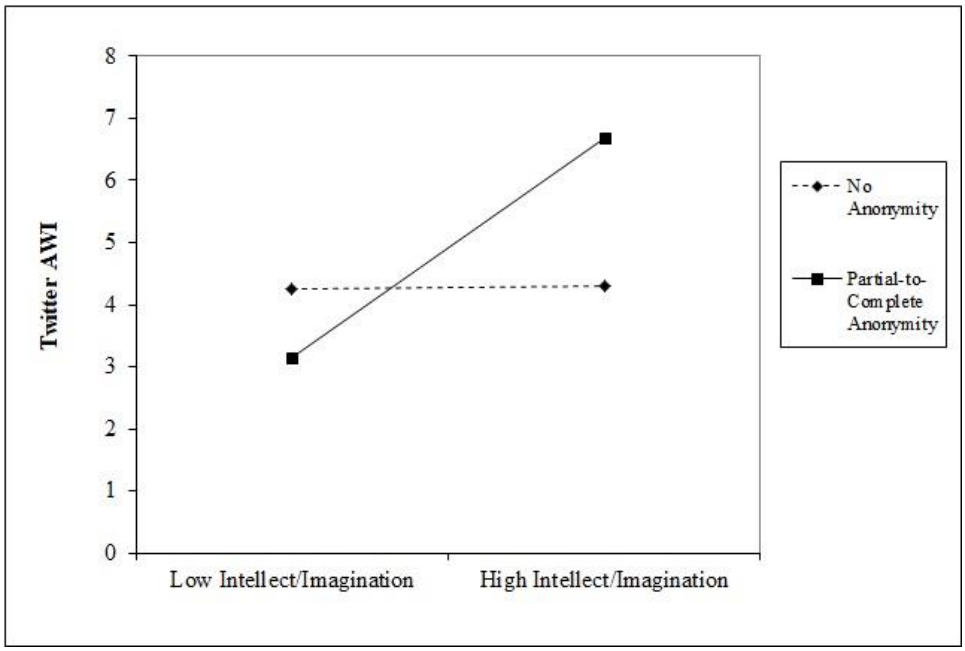


Figure 4. The moderating effects of anonymity on the association between Twitter AWI and intellect/imagination for Sample 2.

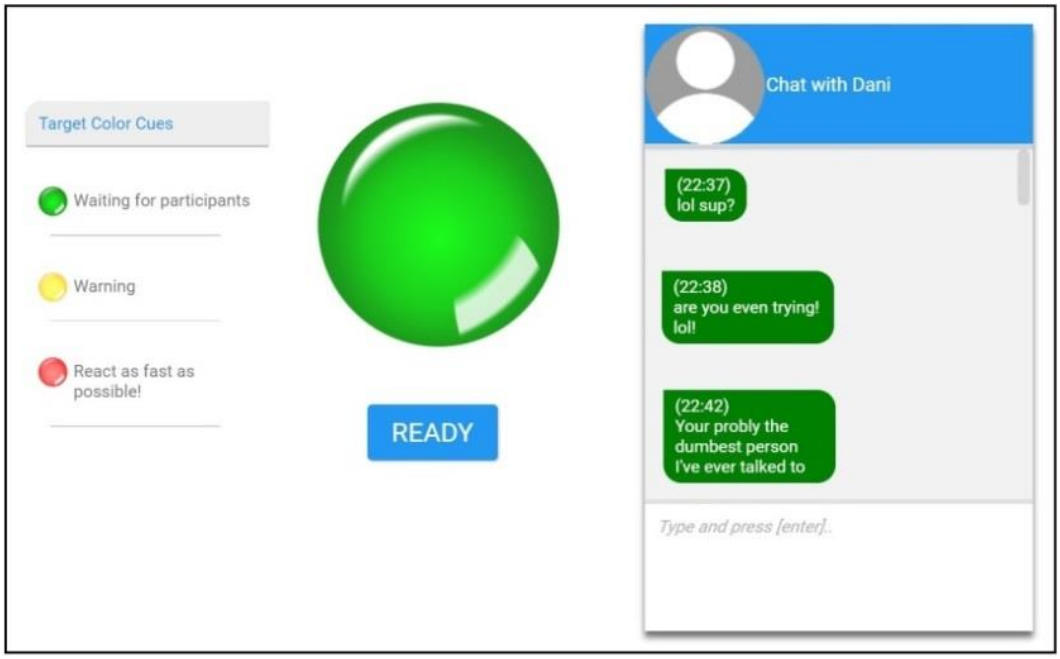


Figure 5. General schematic of the TAP-Chat.

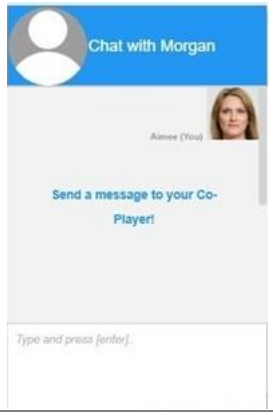
	Self-Anonymous	Self-Identified
Other-Anonymous	Condition A 	Condition B 
Other-Identified	Condition C 	Condition D 

Figure 6. Different conditions of anonymity on TAP-Chat.

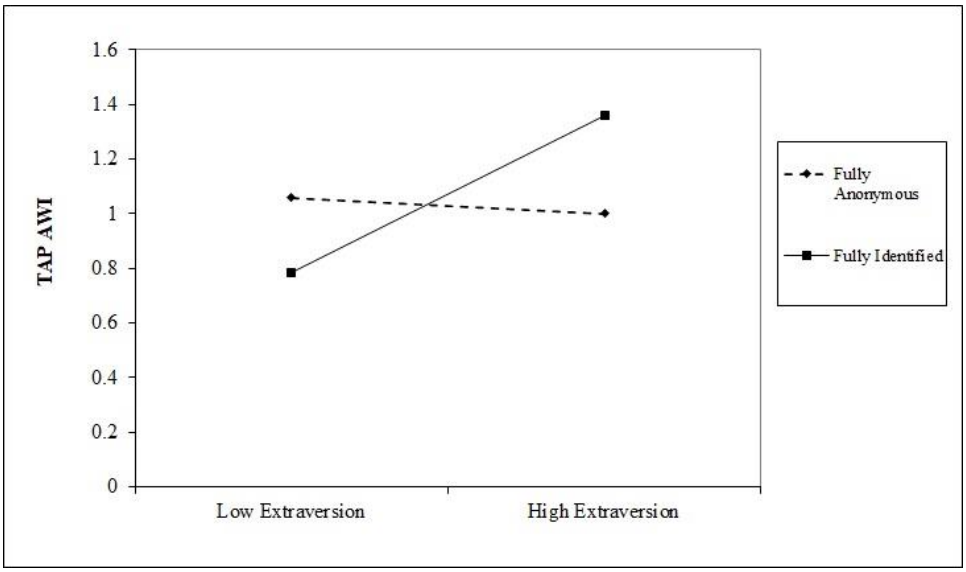


Figure 7. The moderating effects of anonymity conditions on the association between TAP AWI and extraversion.