Using Bayesian Methods to Determine Truth-Telling in an Online-Based Survey on Aggression

A thesis presented in partial fulfilment of the requirements for the degree of

Master of Arts

In

Psychology

At Massey University, Manawatu, New Zealand.

Kayla Gray

2024

Abstract

The desire to obtain truthful responses in research on sensitive matters has been explored widely over a number of years, and in a variety of fields. Researchers have regarded the Bayesian Truth Serum (BTS) and its application of incentive-based methods as the lead framework for reducing the effect of social desirability bias (SDB) on accurate and honest responses in research. BTS operates on the assumption we always portray ourselves to be better than we realistically are. By offering a financial incentive to provide honest answers, the expectation is people will share more sensitive and socially undesirable information about themselves, as motivated by financial rewards. 289 participants were recruited via Prolific and asked to complete an online-based survey regarding their aggressive attitudes and behaviours. Participants also provided estimates regarding the percentage of people in the general population who they believed would share the same perspective. Results indicate that while BTS may have encouraged participants to respond truthfully about the aggressive tendencies of themselves and others, the results were not significant. However, some findings were consistent with previous research in that males were found more likely to endorse the use of aggression in certain situations and were more likely to have engaged in aggressive behaviour (based on their tendency to respond positively to the use of aggression). In comparison, females were more likely to respond angrily when provoked, which lends support to existing research. In conclusion, our findings indicated the application of a financial-based incentive (BTS) did not produce significant differences between groups regarding the measurement of aggressive behaviour and perceptions. Despite this, results indicated the *way* participants responded to statements about aggression differed depending on the group they were assigned to. Discussed are future directions which may improve the success of BTS methodology in the field of aggression, and implications for future research in this area.

Preface and Acknowledgements

The purpose of this research was to obtain truthful responses about aggressive behaviour with the idea that results obtained could better inform the development of targeted interventions for such behaviours. Better yet, we also hoped our research would provide an accurate reflection of views within the general population which would translate to an increased understanding and awareness of the impact aggressive behaviour can have.

To my first supervisors, Peter Cannon and Aaron Drummond - thank you for sharing your knowledge and time, and for challenging me to stretch beyond my comfort zone. I must also acknowledge my supervisor, Ute Kreplin, for her guidance and support at the time that mattered most. I would also like to acknowledge Rina Parry, who provided her unwavering positivity and mathematical genius from beginning to end. And finally, to my partner Trevor, thank you. You were my motivation when I lost mine and stood by me to ensure I made it to the end. After a tumultuous four years, it is incredibly relieving to finally have this research on paper, and I cannot thank all my supervisors and supports enough for their contribution to the content of this thesis project.

Approval for the research has been obtained from the Massey University Human Ethics Committee (Application Number: 400021260) for the experiments described.

Table of Contents

List of Tables

List of Figures

Acronyms

ABS – Aggressive Behaviour Scale
Aggression-ES-A – Aggression Experience Sampling
BAQ – Brief Aggression Questionnaire
BDHI – Buss-Durkee Hostility Inventory
BPAQ Buss Perry Aggression Questionnaire
BPAQ-SF – Buss Perry Aggression Questionnaire Short Form
BTS – Bayesian Truth Serum
CMAI – Cohen-Mansfield Agitation Inventory
CNTR – Control
CV – Contingent Valuation
EXP - Experimental
FFM – Five Factor Model
GAM – General Aggression Model
HDHQ – Hostility and Direction of Hostility Questionnaire
IAT – Implicit Association Test
IED – Intermittent Explosive Disorder
IM – Impression Management
IP – Internet Protocol
IPV – Intimate Partner Violence
MMPI – Minnesota Multiphasic Personality Inventory
POAS – Perception of Aggression Scale
SDB – Social Desirability Bias
STAS – State Anger Scale
UK – United Kingdom
US – United States of America
WHO – World Health Organisation
WTA – Willingness to Accept
WTP – Willingness to Pay

Using Bayesian Methods to Determine Truth-Telling in an Online-Based Survey on Aggression

Introduction

There are many negative consequences associated with the use of aggression, including physical injury and the need for medical intervention (Munoz-Rivas et al., 2007) and psychological implications such as anxiety, emotional fatigue, and symptoms consistent with burnout syndrome (Bernaldo-De-Quiros et al., 2015). Becoming a victim of psychological aggression (such as insults, use of derogatory names, and acting out of spite) also contributes to negative emotional responses for the victims of such aggression (Shorey et al., 2012). The use of physical aggression against children in the context of parental discipline contributes to maladaptive coping strategies and an increased use of aggression among children (Weiss et al., 1992). A positive relationship also appears to exist between the use of physical aggression and towards children and children's own aggression, indicating that more severe use of aggression encourages more severe aggressive responses by children (Weiss et al., 1992). Boys with a parental figure incarcerated have also been shown to demonstrate physically aggressive behaviour at an increased rate whereas the same has not been observed for girls (Wildeman, 2010). Aggression arising within an institutional facility such as prison, also gives rise a greater incidence of violent recidivism in the community (Mooney & Daffern, 2015). This suggests a relationship may exist between negative attitudes towards aggressive behaviour and one's engagement in such behaviour, which can ultimately lead to a period of incarceration and an increased propensity for violence. Aggressive behaviour and attitudes can be present in a number of situations, and at times, can contribute to an increased risk of criminal convictions in adulthood. Interactions between environmental and biological factors, genetic predispositions and personality factors are considered to contribute to the development of aggressive behaviour. The catalyst model has been influential in synthesising these factors to create a framework from which to base the study of aggression.

This study is largely motivated by the desire to effect positive change in our communities, and especially within an offender population where aggressive behaviour is often prevalent. Aggressive behaviour can be dependent on a number of variables such as negative emotionality, delinquency in childhood, lack of support, exposure to domestic

violence and the development of maladaptive attitudes towards aggression (Bonta & Andrews, 2017; Wildeman, 2010). Aggression is however a difficult phenomenon to measure, as research within this field defines aggression in several ways, measures aggression using a variety of strategies, and the role environmental, genetic, and individual factors are posited to play in the development of aggression also differs (Longino, 2001). Aggression has been defined as the act of harm against another with immediate intent to cause damage (Anderson & Bushman, 2002), an externalised behaviour with the goal of hurting others who want to avoid such harm (Bushman & Huesmann, 2010), and must be observable, intentional, and involve physical beings (Allen & Anderson, 2017). An important consideration in the definition of aggression is also the target's motivation to avoid becoming a victim of aggressive behaviour (Allen & Anderson, 2017).

In a comparison of self- and partner reports on aggressive behaviour, high correlations were found between both report methods, especially on measures of physical aggression (O'Connor et al., 2001). Agreeance on factors of physical aggression were also found to correlate with the endorsement of reacting aggressively to certain scenarios with the use of self-report methods (O'Connor et al., 2001). Face-to-face interviews have also been used to measure aggression, which have demonstrated positive reports of both experiences of indirect aggression as the victim or perpetrator (Forrest et al., 2005).

Research into the genetic influence on the development of aggressive behaviour posits that almost 50% of the variance in aggression studies is accounted for by genetics and heritability (Craig & Halton, 2009). Environmental factors are also found to contribute to individual differences in aggression, whilst high-stress and abusive environments tend to promote the development of antisocial behaviours; a factor of aggression (Craig & Halton, 2009). Individual beliefs about the use of aggression have also been found to influence one's actual engagement in aggressive behaviours across time (Barchia & Bussey, 2011). People who believe they are entitled to act aggressively and are justifiable in such actions, tend to enact aggression at a much higher rate than those without the same beliefs. Children who disengage from what is considered morally ethical also tend to become more aggressive over time, which also demonstrates the stable nature of aggression (Barchia & Bussey, 2011).

Participants were asked to complete an adapted version of the Buss-Perry Aggression Questionnaire Short Form (BPAQ-SF) which will include further questions from the full version to obtain a more robust measure of physical aggression; physical aggression being a factor of aggression we anticipate will have more direct association with aggressive behaviour than other facets of aggression which are measured by the BPAQ. In this study,

participants will also complete another measure of aggression - the Perception of Aggression Scale (POAS). This questionnaire was included due to the wide use of the tool as a stable and accurate measure of individual's perceptions of aggressive behaviour.

The primary aim of this study was to determine whether individuals would respond differently, and thus be more truthful, if they were informed a 'truth-telling algorithm' would be applied to their answers. We applied the Bayesian Truth Serum (BTS) method to this study for the purpose of maximising the truthfulness of answers obtained from the aggression questionnaires. Applying the BTS method to online surveys has demonstrated an increase in honest responses thus ensuring more accurate and increased quality of response outcomes, and subsequent applicability in the 'real' world (Frank et al., 2017). When individuals are asked to provide information regarding their own beliefs about a certain topic (such as aggressive behaviour), as well as make predictions about others' beliefs, our judgement is often biased by our own beliefs. Thus, utilising BTS within an online survey should improve reliability of participants' responses.

The use of the BTS method within aggression research is limited. The BPAQ is a widely used measure for aggression however there are a limited number of studies which validate the tool on any testing populations other than college students. As such, developing a new method for detecting truthfulness in an online-based survey of aggressive behaviour and attitudes with a general testing population may allow for increased applicability to other populations. Obtaining truthful responses would also paint a clearer picture of how aggression and associated behaviours are perceived which is likely to make a substantial contribution to the research literature in this area. This study aims to determine how truthful participants are about their own aggressive 'traits', and how they compare themselves in relation to other people by utilising two key measures of aggression – the BPAQ and POAS. It is expected the information obtained as a result of this study will inform the development of a practice tool and/or rehabilitative programme targeted to address known facets of aggressive behaviour in the general population.

Literature Review

Aggression is defined as a behaviour in which the perpetrator of the aggressive behaviour has both the intent and the belief their actions will cause harm to another individual (Anderson & Bushman, 2002). It is also considered an important facet of negative emotionality which can contribute to aggressive behaviour and an inability to control such

behaviour, delinquency in childhood and an increased risk of criminal convictions in adulthood (Bonta & Andrews, 2017). Violence on the other hand is a form of aggressive behaviour with more severe intended consequences such as death (Anderson & Bushman, 2002). The difference between aggression and violence is that aggressive behaviour does not involve extreme pain or injury to another, whereas violent behaviour involves malicious intent to cause serious physical harm to another (Busman & Huesmann, 2010). Aggression also manifests itself in different forms – physical, verbal, or relational. Each form of aggression is exhibited in different ways however it is generally direct or indirect, verbal, or physical, active, or passive. Differences also exist between the purpose of, and motives behind, aggressive behaviour. There is a tendency for such behaviour to be perceived as reactive or proactive; with reactive aggression often motivated by angry emotionality and the desire to inflict harm to another, whereas the latter is deliberate and intentional behaviour which is motivated by other factors such as finances (Busman & Huesmann, 2010).

Individual differences in personality factors and fantasies about aggression have been shown to have an impact on one's engagement in proactive and reactive aggression (Huesmann, Eron & Lefkowitz, 1984). Narcissistic tendencies have also been found to contribute towards one's likelihood of becoming involved in antisocial behaviour later in life, with such individuals more likely to react aggressively when offended (Tharsini, 2019). Tharsini (2019) further reported children who experience maltreatment and who struggle to control feelings of anger in childhood tend to exhibit physically aggressive behaviour in adulthood and can also manifest itself in criminal convictions. However, childhood neglect is not limited to informing the development of aggressive behaviour later in life, as it has also been found to correlate with the development of a range of psychiatric disorders (Buchmann, Hohmann, Brandeis, Banaschewski & Poustka, 2014). Excessive use of punishment, inconsistent parenting styles, and the unavailability of a positive parental role model, combined with a distinct lack of positive reinforcement are also noted as contributing factors towards aggression and conduct problems in adulthood (Buchmann et al., 2014). Similar findings were reported by Hodgins, Cree, Alderton and Mak (2008) who found individuals exhibiting symptoms of conduct disorder prior to the age of 15 tended to suffer from severe mental illness in adulthood and were also more likely to engage in violent offending behaviour, resulting in a criminal conviction. Thus, it appears that having aggressive inclinations and challenging behaviour as a child, as indicated by a conduct disorder diagnosis, is associated with acting aggressively and committing violent crimes as an adult (Hodgins, Cree, Alderton & Mak, 2008).

Negative emotionality is another factor of aggression which can remain static over the course of an offender's life, thereby promoting the development of criminal behaviour and consistently maintaining it across the lifespan. A longitudinal study conducted by Huesmann, Eron and Lefkowitz (1984) demonstrated correlations between early displays of aggressive behaviour and criminal behaviour later in life, especially amongst males. Negative emotionality coupled with emotion dysregulation has also been found to contribute towards aggressive tendencies amongst violent offenders, and specifically, the display of physical aggression (Garofalo & Velotti, 2017). Anger is another facet of physical aggression shown to impact on the manner in which aggressive behaviour is displayed, which is often compounded by the presence of negative emotionality. Thus, individuals who are presented with a distressing or upsetting situation, and who are prone to negative emotionality tend to behave in a physically aggressive manner in comparison to individuals affected by positive emotionality. Such individuals tend to adapt better when confronted with an emotionally arousing situation whilst maintaining control of their behaviour, thus indicating the individual differences which contribute towards aggressive tendencies (Garofalo & Velotti, 2017).

As discussed earlier, anger has been identified as a contributing factor towards one's decision to engage in an aggressive manner and appears more often in strong and physically attractive individuals (Wyckoff, 2016). The presence of anger has been found to predict the use of both direct and indirect aggression, and impact on one's self-reported desire to engage in aggressive behaviour against another individual (Wyckoff, 2016). Furthermore, individuals who are considered strong and physically attractive, both self-perceived and perceived by others, were found to be more likely to aggress directly than individuals low in strength and attractiveness. Individuals in the latter category were considered more likely to utilise indirect aggression towards others in the presence of anger, which was perceived to be a tactical use of aggression (Wyckoff, 2016). Trait anger specifically, a type of anger associated with neuroticism, has been found to have a positive relationship with both intrinsic and extrinsic motivation to disengage or engage in an angry and aggressive manner (Smits & Kuppens, 2005).

Individuals with a tendency to engage in outward anger displays also tend to exhibit lower levels of self-control, and therefore differ in disinhibition levels, resulting in a higher incidence of physical aggression (Smits & Kuppens, 2005). Individuals with a tendency to act aggressively also tend to exhibit a lack of consideration for the consequences of using aggression, which contributes to higher incidences of aggressive behaviour (Bushman et al., 2012). Alcohol use and intoxication only serves to exacerbate aggressive behaviour in

individuals who display a distinct inability to consider the future consequences of their behaviour. Consequently, aggression is observed more frequently among intoxicated individuals in comparison to those who remain sober, especially among men (Bushman et al., 2012). When it comes to women, powerlessness, being disrespected, life stressors and low self-esteem are identified as factors associated with anger and aggression (Thomas, 2005). Interestingly, anger among African American women is different, and more acceptable within the culture of such a society for women to act as aggressively as men as a means to protect oneself from relentless racism (Thomas, 2005). Among children, observing and learning from interpersonal relationships can contribute to a higher incidence of anger and aggression and an inhibited perspective of their own aggressive tendencies (Lochman et al., 2010). A poor awareness of internalised aggressive tendencies can influence the development of poor social and problem-solving skills, abuse of substances and engagement in antisocial behaviour.

The impact aggressive and maladaptive behaviour can have on others indicates the importance of research in this field. Children who were perceived to be more aggressive in childhood were found to score more highly on specific factors related to the measurement of aggression in the Minnesota Multiphasic Personality Inventory (MMPI) scales in adulthood; indicating aggression tends to be a stable personality characteristic across time (Huesmann et al., 1984). Aggressive behaviour was also found to manifest in traffic infractions, alcohol-related driving offenses, domestic violence, physical aggression, severe and inflexible child discipline, and convictions for criminal offenses (Huesmann et al., 1984). Similarly, both indirect and physical aggression has been found to remain stable across childhood (Vaillancourt et al., 2003). Both boys and girls were also found to be consistent in their displays of aggression, however it remained particular stable for physically aggressive behaviours. Aggression during childhood has also been linked to criminality as an adult, albeit not as a guaranteed pathway but more as a contributing factor (Huesmann et al., 2009). Highly aggressive boys have also been found to maintain patterns of aggression across time, while highly aggressive girls are more likely to lessen in levels of aggression over time (Huesmann et al., 2009). Environmental and social circumstances can also contribute to the stability of aggression over time. This manifests itself in higher incidences of being arrested for engagement in criminal behaviour, traffic infractions, presence of intimate partner aggression, substance abuse, mental illness and poor interpersonal relationships (Huesmann et al., 2009).

The consideration of genetic, environmental and societal factors as components of causation are likely to be crucial in the prediction of one’s propensity to engage in violent or

aggressive behaviour across the lifespan. Individual characteristics such as personality factors, genetic propensities and attitudes towards aggressive behaviour remain stable across time, largely attributed to the establishment of schematic structures from learned experiences (Anderson & Bushman, 2002). Research on behavioural genetics suggests inclinations for violence and establishing a criminal history are more consistent among identical twins in comparison to fraternal twins (Stangor et al., 2014). Research within the genetics field provides evidence to support the idea that specific genetic factors are at play in the development of aggression over time, with highly aggressive infants found to still be aggressive as adults. Importantly, children who are mistreated appear to be more affected by particular genetic factors, which is thought to contribute to poor inhibition of aggressive behaviours. These findings provide support for the idea that both genetic and environmental factors appear to be at play in the manifestation and development of aggression (Stangor et al., 2014).

Social circumstances such as exposure to violent role models and violent forms of media, being excluded from one's social group or class and indulgence in substances, illegal or otherwise are all considered to be contributory factors to the development and engagement in aggressive behaviour (Warburton & Anderson, 2015). Furthermore, individuals with predispositions to engage in aggressive behaviour, developed from an intertwined web of factors and circumstance, tend to be increasingly aggressive under the influence of substances given the depressant effect substance use has on aggressive tendencies (Warburton & Anderson, 2015). Exposure to violence as a youth has also been found to influence youth engagement in violent offending behaviour, increased rates of truancy, use of illicit substances, and issues with attention, thoughts and somatic symptoms of illness (Kirk & Hardy, 2014). Engagement in offending behaviour as a youth also appears to foster an environment in which youth are continually exposed to violence and aggression due to an increased likelihood of involvement with antisocial peers and participation in risk-taking behaviours. A lack of social support, relaxed supervision from a parent, and unsafe and impoverished neighbourhoods were also identified as contributing factors to aggression amongst youth populations (Kirk & Hardy, 2014).

Cultural factors also bear relevance in the measurement of aggression given that aggression is perceived differently depending on the culture which individuals prescribe to (Severance et al., 2013). Within Pakistani culture for example, disrespecting one's image or status by way of aggressive means is considered an extended version of damaging one's honour, which is perceived as a violation of self-worth. Honour is inextricably linked to one's

self-worth and level of self-esteem within a Middle Eastern context which is not observed in other cultures (Severance et al., 2013). Among an independent and Westernised culture such as that of the United States, greater weight is instead placed on material wealth and professional status. Threats to undermine those aspects of one's life by aggressing tends to be perceived as incredibly damaging (Severance et al., 2013).

Attitudes towards aggression are also formed in early childhood, as evidenced by the study completed by Barnes, Howell, Thurston and Cohen (2016). The development of maladaptive attitudes toward aggression were found to be influenced by several factors, including childhood depression, and witnessing significant others engaging in aggressive behaviour (Barnes et al., 2016). Furthermore, children displaying negative emotional behaviours such as anger and aggression tended to receive a lack of support from significant others, thus reinforcing the continuation of aggressive behaviours. While no significant results were found to support the influence of positive peer responses on negative emotionality in children, the study provides evidence supporting the influence of social factors on the development of relational and physical aggression as a perceived means of effective conflict resolution in children (Barnes et al., 2016).

Other social factors associated with parental incarceration such as reduced financial resources, changes in family structure due to divorce or separation of parents and altered disciplinary practices were also found to contribute to poor wellbeing and aggressive behaviour in children (Wildeman, 2010). In reference to school shooters, which provides an example of an extreme display of violence and aggression against others; such individuals are posited to possess a number of psychosocial characteristics pointing to a prescribed 'pathway' to aggression (Watson et al., 2004). Such a pathway is hypothesised to be characterised by an individual with low self-esteem, a vulnerability to bullying behaviour, a victim of social exclusion, dysfunctional family units, an alignment with aggressive ideologies and the consumption of violent media to imitate the use of weapons. While the consumption of violent media is considered to contribute towards aggression in children, exposure to 'real-life' violence and aggression has also been found to instil particular beliefs about the use of aggression among children (Mitrofan et al., 2014). Children being raised in a low-income environment tend to display higher levels of aggression, which is exacerbated by 'real-life' examples of the use of aggression by exposure to domestic violence within the home, through neighbours or on the streets. As a result, children grow up with a skewed perspective of aggressive behaviour as the 'norm' from which to act; maintained through one's own experiences of aggression across childhood (Mitrofan et al., 2014).

Another important factor to consider in the research on aggression is gender. Gender differences in the development and use of aggression have long been present in the aggressive research fields; differences which appear to develop from an early age. In children, boys are more likely to use physical aggression (hitting, punching) to settle conflict with others, whilst girls typically internalise conflict and utilise relational aggressive tactics such as gossiping (Barnes, Howell, Thurston & Cohen, 2016). Similar findings were reported by Wildeman (2010) which examined the effect of paternal incarceration on children's aggressive behaviours. Boys were found to exhibit more physically aggressive behaviours whereas no significant influence was found between paternal incarceration and girls' physically aggressive behaviours. That is not to say that aggression in girls does not exist; however, the type of aggression that girls display is often more indirect as opposed to directed physical aggression which is often seen in boys (Bjorkqvist, 2018).

Hormones and biological factors offer some explanation on differences between genders. Research has suggested a correlation between the length of a child's second and fourth fingers and aggressive behaviour, which is determined during pregnancy and influenced by hormones (Bjorkqvist, 2018). Bjorkqvist (2018) also reported a low ratio between the second and fourth finger lengths was found to influence physical aggression in men whereas the same ratio has been found to correlate with indirect aggression in women. Furthermore, indirect aggression is characteristically associated more with females due to a higher degree of social intelligence whereas males have a preference to engage in physical aggression (Bjorkqvist, 2018). Similar correlations were reported by McDermott (2015) who noted the tendency for males to engage in physical aggression as a means to assert their dominance, achieve a greater social status and to appear superior to other males. Women on the other hand are more inclined to engage in verbal aggression as their incentive to maintain reproductive capacities is more pronounced; an incentive which engagement in physical aggression may jeopardise (McDermott, 2015) An example of this can be seen in the percentage of female vs. male murderers; murder being an act of physical aggression and violence. Typically, most murders are committed by males and can be seen as a way to achieve clout and reproductive status over others (McDermott, 2015). In the US, homicide rates for men are significantly higher than those of women, and women are more likely to become victims of murder at the hands of a male than the reverse (Kellermann & Mercy, 1992). In comparison to males, females are also more likely to be a victim of violent offending, such as sexual and aggravated assault, as perpetrated by an intimate partner (Catalano et al., 2009). There is also a greater incidence of female murder victims when

compared to males, which has remained consistent since 1993 in the US. Prevalence rates of violence in the US also suggest males are less likely to become the victim of violence and are more likely to perpetuate the use of such violence (Catalano et al., 2009).

Differences in how aggression manifests itself depending on the gender of the individual is thought to be partially explained by mating behaviour, thus indicating the influence of evolutionary factors on the development of aggression (McDermott, 2015). Along with homicide, males also dominate the statistics for recorded aggravated assaults, with males seen to engage in this type of aggressive behaviour more so than their female counterparts (Anderson & Bushman, 2002).

Models of Aggression

The social factors described above are considered fundamental in the development of aggression in childhood and through into adulthood however such factors are only considered on an ambiguous level in the widely used theoretical general aggression model (GAM) when explaining aggressive behaviour (Ferguson et al., 2008). According to the GAM, aggressive behaviour is learnt from viewing violent media and other aggressive stimuli, and the greater the consumption of such media, the more aggressive the individual. Exposure to violent media contributes to the development of aggressive cognitive scripts which are automatically activated upon viewing violent stimuli, such as that which appears in violent video games (Anderson & Dill, 2000).

Exposure to violent media is also thought to impact upon internal biases and beliefs, reinforce skewed expectations of others' aggressive' behaviour, generate attitudes which support violent means of conflict resolution and inflicts changes upon the viewer's personality. Individual factors such as high self-esteem, gender, beliefs about one's ability to execute aggressive actions, specific attitudes towards engagement in violent behaviour, morals and values, and goals to be wealthy or be a formidable member of society (such as that seen in gangs) can all influence the development of and engagement in aggressive behaviour (Anderson & Bushman, 2002). While these factors are considered in the GAM, they are solely proposed to be attributed to underlying schematic scripts formed from personal experiences.

The GAM is influenced by social learning theory and operates on the principle that repeated exposure to violent media leads to increased engagement in violent behaviour. Interestingly, participants found to be higher in trait aggression as measured by the Buss

Perry Aggression Questionnaire (BP-AQ) also appeared to be influenced more by exposure to violent media in comparison to participants low in trait aggression. Such a finding infers the possibility of a causal relationship between continual exposure to violent media and increased aggressive behaviour, which subsequently heightens one's desire to consume such media (Anderson & Dill, 2000). The GAM provides a structure of knowledge which helps to identify the factors that influence and contribute to aggressive behaviour between intimate partners and groups, within the context of climate change and among victims of suicide (DeWall et al.,2011). Thus, the GAM is a framework that provides the tools to understand and explain aggressive behaviour in a variety of contexts. In the context of intimate partner violence (IPV) for example, individuals who held more desirable attitudes towards acting aggressively within an intimate relationship were more likely to execute such attitudes in the form of physical harm against their partner (DeWall et al., 2011). The GAM also illustrates how individuals become inclined to respond aggressively when faced with certain situations (Allen et al., 2018). For example, an individual who has predilections which are more favourable towards violence is more likely to undertake an immediate negative appraisal of the situation with a desire to harm the person responsible for the situation. Overall, the GAM is able to elucidate the ways in which individual and environmental factors contribute to thoughts and feelings associated with aggression and anger; provides a context for the influence of arousal levels on the decision to engage in aggressive behaviour; and explains how the repetitive cycle of such processes acts as a knowledge pathway to the development of aggressive structures (Allen,et al., 2018). When all of these factors are combined, the result is posited to be an aggressive personality.

However, given the GAM tends to minimise exposure to family violence and biological and genetic predispositions as critical contributory elements to aggressive and criminal behaviour, the value of such model to predict aggressive behaviour remains equivocal (Ferguson et al., 2008). Although widely used, the GAM does not fully recognise the impact external variables have on aggressive behaviour, which have been identified as crucial in the understanding of aggressive behaviour (Ferguson et al., 2008).

To account for the limitations of the GAM is the catalyst model; a model which acknowledges the interaction between environmental factors (exposure to media violence), biological and genetic predispositions and personality factors as influential on the desire to engage in aggressive behaviour (Wildeman, 2010). The catalyst model considers violent media as a medium for violent behaviour to occur, hereby rendering violent media as a 'role model' for those who consume it. Products of the social environment such as financial

stressors, divorce, separation, and family violence are perceived as catalytic rather than causal agents in the development of violent behaviour; factors which were echoed by Wildeman (2010) as facilitators of aggression in children. The catalyst model has largely been utilised amongst the violent media field of research and appears less frequently within the aggressive behaviour field. In Ferguson et al.'s (2008) study, an aggressive personality, violence-seeking behaviour, and exposure to family violence were considered characteristics and predictors of trait aggression which the catalyst model proposes as the most direct influence on the undertaking of violent crime.

As described earlier, the combination of the consumption of violent media, coupled with internal predispositions and external influences such as the function of the family unit and exposure to aggressive role models is hypothesised to contribute to aggressive tendencies among children (Mitrofan et al., 2014). These factors are considered best explained by the catalyst model due to its' collaborative approach in understanding the causative influences on aggression. Individuals with genetic predispositions to engage in violence and aggression tend to engage in such behaviour at a higher rate than those without such motives (Ferguson et al., 2008). Such individuals are thus more likely to seek out violent media which contributes to learned patterns and forms of violence and aggression maintained through the use of "stylistic catalysts" (Ferguson et al., 2008, pp. 315). These 'catalysts' are not necessarily considered causes of aggression but are more so considered contributory factors in the development of aggressive tendencies, such as high-stress environments, exposure to abuse and neglect within the domestic setting and the consumption of violent media. A consequence of the amalgamation of these factors is the development of one's propensity to engage in violent criminal acts (Ferguson et al., 2008).

In a study measuring the prevalence of violent crime and aggression among a prisoner population based on contributory factors as outlined in the catalyst model, the results indicated the proclivity to engage in criminal behaviour and the means of committing such crimes was influenced by exposure to violent forms of media (Surette, 2013). Individuals who use violent media to inform their own aggressive tendencies also tend to engage in criminal offences which have already been committed, using such media to ascribe motivation and inspiration to replicate their own real-life experience of violence and aggression. Access to real-world models of aggression by way of interpersonal relationships with significant others, age and gender are also seen to be catalytic factors not only in the development of aggressive behaviour, but also in the development of maladaptive attitudes

towards aggression. These attitudes play the role of maintaining and justifying one's projection of aggressive behaviour along a criminogenic pathway, at times (Surette, 2013).

The catalyst model is considered superior when compared to the General Aggression Model (GAM) due to its' provision of a valid framework from which to understand aggression and subsequent engagement in criminal behaviour. However, it too has its own limitations. Although the catalyst model may allow a deeper understanding of the how and why behind an individual's engagement in multiple forms of aggression, the specific focus on more serious manifestations of violence and aggression such as assault, murder, and violent sexual violations, may consequentially limit its consideration of low-level aggressive behaviours such as self-defence (Ferguson, 2023). Despite this, the catalyst model is built on the premise that aggression is influenced by genetic, environmental, and social factors which is considered to be a suitable and robust way to view aggression (Ferguson & Dyck, 2012). The General Aggression Model (GAM) on the other hand is rooted in social learning theory with an overarching premise built on notions of aggressive 'schemas' which are maintained and developed through environmental observations across the lifespan (Ferguson & Dyck, 2012). The all-encompassing perspective promoted by the GAM is one of aggression as *always* being a maladaptive behaviour in today's society. Thus, supporters of the GAM tend to also adopt a perspective of aggression as a behaviour that always produces negative and harmful consequences to others which lacks empirical value in the field of aggression (Ferguson & Dyck, 2012).

Measures of Aggression

There are a number of tools available used in the measurement of aggression, such as the Cohen-Mansfield Agitation Inventory (CMAI); a widely used observation-based method of assessing the incidence of aggressive behaviours among older adults (Ravyts et al., 2021). The Aggressive Behaviour Scale (ABS) is another observation-based method used to assess four forms of aggressive behaviour in a population of older adults (Perlman & Hirdes, 2008). The ABS is also able to assess the presence of psychiatric illness and cognitive impairments and the relationship such factors have with aggressive behaviour. The Aggression-ES-A is another tool which has been developed as a brief experience sampling (ES) method and is used in the measurement of aggression in real-time and in real-life contexts (Murray et al., 2022). The 'tool' is self-administered by participants and measures four key types of aggressive behaviour which relate to one's temper, insults, outward verbal aggression and

relational aggression. Negative emotionality, provocation and substance use are also assessed as a means to create a tool which distinctly encapsulates the concept of aggression whilst also acknowledging the diverse nature of aggression (Murray et al., 2022). Hostility has also been identified as a dimension of aggressive behaviour which has been measured with widespread use of the Buss-Durkee Hostility Inventory (BDHI). Positive correlations have been observed between total scores on the BDHI, trait anger on the State Anger Scale (STAS) and general hostility on the Hostility and Direction of Hostility Questionnaire (HDHQ), which provides support for the role of hostile attributions within a sample of individuals with depressive tendencies (Moreno et al., 1993).The BDHI is an inventory based on self-reports across 75 items which are designed to measure key factors of anger, hostility and aggression (Fernandez et al., 2015). Social desirability bias is considered to influence the accurate measurement of aggressive behaviour with this scale and although this can be controlled for, there are other shortfalls noted regarding this tool. Primarily, critiques of the BDHI reference a lack of support for the subscale structure based on results from factor analysis studies, which prompted the development of the BPAQ in aggression research (Fernandez et al., 2015).

To measure trait aggression and related aggressive behaviour requires a particular measurement tool and one which has been widely used and accepted within this field of research is that developed by Buss and Perry, known as the Buss-Perry Aggression Questionnaire (BPAQ) (Gallagher & Ashford, 2016). This questionnaire consists of a number of questions designed to measure four factors of aggression; that being physical and verbal aggression, anger and hostility. Facets of personality such as neuroticism, as outlined in the Five Factor Model (FFM), have also been found to correlate with high scores on the BPAQ, indicating that high levels of neuroticism are positively associated with subscales of trait aggression in the BPAQ (Dam et al., 2019). Such findings were reported from a population of violent offenders currently incarcerated in Denmark and suggest a greater incidence of aggressive tendencies among highly neurotic individuals. Furthermore, findings of strong correlations between FFM factors and the BPAQ provides a more robust overview and understanding of aggression in the community, which in turn provides support for the use of the BPAQ in our study. The 29-item four factor structure of the BPAQ has been found to be adequately replicated in a Hungarian sample of adults which was representative of the national population based on age, gender, education, and density (Gerevich et al., 2007).

Furthermore, the removal of two reverse-coded questionnaire items (“I can think of no good reason for ever hitting a person” and “I am an even-tempered person”) from the full

BPAQ, which measure physical aggression and anger respectively, increased model fit. This suggests such factors may not accurately reflect an aggressive individual. Additional analyses conducted in the study revealed the shortened 12-item version of the BPAQ had improved model fit when compared to the full BPAQ, indicating such a version may be superior in measuring aggression (Gerevich, Bacskai & Czobor, 2007). Similar findings were reiterated by Bryant and Smith (2001) who utilised five different samples of undergraduate students to ascertain the validity of the full and shortened versions of the BPAQ. Interestingly, Bryant and Smith (2001) adapted the questionnaire from a five-point to a six-point likert scale for the purpose of being able to categorise participants as aggressive or non-aggressive. Results concluded the reproduction of the original 29-item BPAQ into a 12-item version of the BPAQ (BPAQ-SF) maintained the conceptual content of the original factors whilst producing an aggression questionnaire which is considered superior when comparing the construct validity of each version. Bryant and Smith (2001) also suggested predictions regarding aggressive behaviour and perceptions can be enhanced by utilising the shortened version of the BPAQ. These findings solidify the use of the BPAQ-SF in our research.

The short form version of the Buss-Perry Aggression Questionnaire (BPAQ-SF) has been used in a wide range of settings and is designed to measure physical and verbal aggression, anger, and hostility (Diamond & Magaletta, 2006). An even briefer version of the BPAQ and BPAQ-SF has also been conceptualised by drawing aspects from the BPAQ, BPAQ-SF and the BDHI and conceptualising these factors into one brief self-report measure labelled the Brief Aggression Questionnaire (BAQ) (Webster et al., 2014). This brief measure has been shown to have good convergent and discriminant validity with the BPAQ and the BDHI, which suggests the BAQ is able to produce valid and reliable scores as a brief aggression measure.

Results from use of the BAQ indicated males were more aggressive both verbally and physically in comparison to their female counterparts, whilst gender did not appear to mediate a significant difference in scores on anger and hostility subscales. The BPAQ-SF however is assessed as possessing greater “internal consistency reliability” (Webster et al., 2014, pp. 136) in comparison to the BAQ. The BPAQ-SF has also been found to have “good model fit” (Cunha et al., 2022, pp. 17) when compared to the original version, and also offers the ability to generalise findings across different samples, as evidenced by the strong measurement invariance found between community subjects and prisoner perpetrators of intimate partner violence (Cunha et al., 2022).

The BPAQ-SF has been found to have good comparable factorial validity across genders, adequate reliability and has a high correlation with subscales on the Personality Assessment Inventory (PAI) (Diamond & Magaletta, 2006). Use of the BPAQ-SF has also been validated in a number of different cultures, including Hungarian (Zimonyi et al., 2021), Portuguese (Pechorro et al., 2016), Romanian (Sabareanu, n.d.) and Dutch (Hornsveld et al., 2009). The results from these studies also provide support for the use of the BPAQ-SF among different samples with success, as evidenced by their validation on samples of college students, adolescent offenders, prisoners, detainees in a forensic psychiatric facility for serious violent offending, and secondary school students, respectively. Criticisms of the BPAQ-SF and its predecessor, the BPAQ, are often built on a perceived inability of such scales to apply the results to the general population given that testing populations are often college students or female-saturated samples (Gerevich et al., 2007). However, results from this study indicated the BPAQ was able to be replicated successfully on a representative sample, thus providing support for the generalisability of an aggression measure.

However, there is still a lack of research in the existing literature which describes the effectiveness of the BP-AQ SF in measuring aggression within the general population and outside of a correctional environment. In addition to this, both the BPAQ and BPAQ-SF are based on self-reported answers and, as a result, are heavily reliant on truthful reporting by participants to produce an accurate understanding of aggression as a whole. This is perceived as a major criticism – not of the aggression measures themselves, but of the difficulty in obtaining accurate views and behaviours regarding sensitive matters (Rasinski et al., 1999). Honest answers regarding one's criminal conviction history for example was not found to be significantly affected by social desirability bias, nor did a financial incentive seem to improve the quality of participants' responses (Preisendorfer & Wolter, 2014). Furthermore, participants' responses to the sensitive question regarding their criminal convictions did not appear to differ in validity, indicating that all participants provided a truthful answer. Despite this, the way participants perceive a survey seems to influence the likelihood of participants providing an honest response to survey items. Participants with a more positive attitude were more inclined to answer truthfully, which became more evident when one's experience with surveys was factored out (Preisendorder & Wolter, 2014). Given the reliance on self-reported answers, and that the answers provided are an honest indication of a participant's views of aggression, the desire to generate truthful responses is a focus of our study.

As has been found, it can be difficult to obtain truthful answers to measures reliant on self-reports, especially when the subject matter is of a sensitive nature. A lot of the literature

has considered the impact of truth-telling in self-reports however this tends to be within the context of financial and economical decision-making and less within the field of aggression. For example, all participants were found to practice dishonesty within the context of determining compliance with financial regulations, simply to gain a small financial incentive (Friesen & Gangadharan, 2013). Individual beliefs and preferences also seem to influence how honest participants are regarding economic decisions (Danz et al.,2020). However, if participants were not provided with any information regarding the financial incentive, they could expect to receive simply by participating, participants tended to provide more honest responses. Even so, when participants are aware of the financial incentives to respond truthfully, false reports are still evident (Danz et al., 2020). This suggests that even if participants are advised of the possible financial benefit available to them based on the truthfulness of their responses, they are not always deterred from being dishonest.

Previous research would also suggest that participants' perceptions of themselves and how they are perceived in social contexts contributes to an increased incidence of truth-telling (Abeler et al., 2019). Thus, if being perceived as a liar is considered incongruent with one's perception of themselves, and is generally disapproved of amongst the wider society, an individual is considered more likely to respond truthfully. Self-preservation of one's status and reputation may therefore outweigh the cost of being dishonest in self-reports (Abeler et al., 2019). The perceptions individuals hold about themselves and others also appears to influence people's behaviour (Ickes et al., 1997) and how they present themselves to others (Vasire & Gosling, 2004).

Self-perceptions have also been found to influence engagement in direct and indirect aggression, as evidenced by the finding that relational aggression is more common among girls who are considered popular based on social perspectives (Mayeux & Cillessen, 2008). In comparison, boys who considered themselves to be popular tended to exhibit more direct aggression. However, both boys and girls who believed themselves to be well-liked but were actually disliked by others tended to demonstrate behaviour consistent with relational aggression more so in comparison to their peers without the same perceptions (Mayeux & Cillessen, 2008). Self-perceptions have also been found to influence engagement in delinquent behaviours among young juveniles (Smith et al., 2015). Negative views of the self, "perceptual bias" (Smith et al., 2015, pp. 616) and narcissistic tendencies were all found to be contributory factors to aggression among young male offenders. This finding provides support for the idea that aggressive behaviour is exhibited in different ways depending on the person's individual perception of themselves. Smith et al. (2015) suggested a "positive

perceptual bias" (Smith et al., 2015, pp. 616) may result in a greater degree of tolerance for aggressive behaviours before an escalation occurs. It could therefore be postulated that having a negative 'perceptual bias' may encourage the maintenance of and engagement in negative aggressive behaviours across time; a view consistent with the catalyst model of aggression. Support for this idea can be found in the study completed by Sherman et al. (2015) who found that both individual personality characteristics and real-life experiences contributed to differences in behaviour such as aggression, and emotional responses to a situation.

The second aggression measure we used in this study is the Perception of Aggression Scale, also known as the POAS (Jansen et al., 2013). Predominant use of the POAS has occurred in a psychiatric nursing setting; the goal of which is to determine a nurse's perceptions toward aggression and violence displayed by patients within a clinical setting (Jansen et al., 2013). It was initially developed as a 32-item scale and used in a sample of Dutch nurses employed in a psychiatric hospital however it has also been used successfully on sample of Turkish nursing students, which provides further evidence to support the use of the POAS as a reliable and appropriate measure of attitudes towards aggression (Bilgin et al., 2011).

The POAS has also been applied successfully to a sample of Swiss nurses with a background in psychiatry, with results demonstrating nurses from this population tend to perceive aggression as an undesirable phenomenon (Abderhalden et al., 2002). Use of the POAS in this context was found to produce comparable results to that of the original study completed by Jansen et al. (1997), who also found the general perception of aggression to be 'normal', threatening, or functional, depending upon the setting in which such behaviour manifested (Jansen et al., 1997). Nurses who worked in settings which utilised measures such as isolation tended to have a more unfavourable perception of aggression when compared to nurses working in settings where such measures were not applied.

Differences have also been observed between nurses in different countries, with European nurses more likely to perceive aggression as a dysfunctional behaviour than their Swiss and German counterparts (Abderhalden et al., 2002). However, some items measured on the full version of the POAS appeared to translate poorly into a 12-item version of the scale, as evidenced on a sample of Swiss nurses in a psychiatric facility (Needham et al., 2004). Despite this finding, the brief version of the POAS has been found to demonstrate good internal consistency (Cronbach's alpha = 0.76-0.83) and test-retest reliability (Pearson's r = 0.87) in a sample of Chinese clinical staff (Wong & Chien, 2017). Use of the short-form

POAS within a French psychiatric facility has also been supported and is considered to be a stable measure of attitudes towards aggressive behaviour (De Benedictis et al., 2012). Interestingly, older participants were also found less likely to endorse the use of aggression as a protective measure in comparison to their younger counterparts, which suggests that age plays a role in the perception of aggression. Gender also appears to play a role in the perception of aggression, with more males found to perceive aggression as negative when compared to their female counterparts in a sample of Swedish nurses (Palmstierna & Barredal, 2006). Of note, the original German version of the POAS was successfully translated in the study by Palmstierna and Barredal (2006) and applied to a Swedish population, demonstrating the applicability of the scale in a range of countries and clinical settings.

As discussed, the POAS, BPAQ and its predecessor, the BPAQ-SF, are widely regarded as popular and valid measures within the field of aggression, and among a wide range of sample populations. However, a criticism of these measures is due to the heavy reliance on respondents being honest and truthful in the responses they provide. This has proven to be an aspect of self-report measures which is difficult to obtain, especially in surveys regarding sensitive topics such as finances, criminality, and compliance with economic regulatory frameworks. Individual characteristics, self-perceptions and perceptions of others also appear to predict one's engagement in aggressive behaviours at a greater rate than those without the same predilections. In this study, we attempt to improve the truthfulness of responses on the POAS and BPAQ-SF by offering a financial incentive and informing participants of a 'truth-telling algorithm' which was applied to their responses. Our overarching goal of this study was to determine how individuals perceive aggression within themselves, and to what degree they believe others in the general population to share similar perceptions. More so, do individuals become increasingly honest regarding their attitudes and views if they *believe* their responses are scanned for inaccuracies?

Another focus was on the responses obtained from our adapted version of the BPAQ-SF. Overall, both of the measures employed in our study rely on capturing the true views of the sample population given they are both self-report measures.

Truthfulness in online surveys

The ability to generate truthful and accurate responses in surveys can be difficult as individuals are often influenced by social desirability bias (SDB); a bias which is inherently bound by one's desire to appear prosocial and better than they really are (Brenner &

DeLamater, 2016). This theory proposes individuals' own internalised belief and value systems are reflective of the environment and society of socialisation, which differs across populations. This has an impact on the types of behaviours which are considered normative and subsequently reported (Brenner & DeLamater, 2016). As a result, individuals may respond in a way which is considered socially desirable rather than responding based on what a certain factor intends to measure (Anguiano-Carrasco et al., 2013).

Bayesian Truth Serum (BTS) has been applied in the field of environmental behaviour where participants are asked to provide self-reported information regarding their energy consumption (Gamberini et al., 2014). Energy consumption is considered to be a behaviour which is susceptible to judgement by social norms and is postulated to share connections with values of respect, efficiency and wealth. Gamberini et al. (2014) applied both implicit and explicit measures in their study to determine participants' engagement in energy-saving or wasteful behaviours as they relate to the use of electronic devices. Implicit measures in this context involved the use of an autobiographical form of the Implicit Association Test (IAT) which is designed to investigate how often one *remembers* engaging in a particular behaviour – energy consumption – without directly measuring the target behaviour. The explicit measure utilised by Gamberini et al. (2014) was a direct questionnaire about how often one engages in energy-saving or wasteful energy consumption practices. The combination of both implicit and explicit measures employed in this study demonstrated the influence social desirability bias (SDB) has on how participants respond to questions regarding socially desirable behaviours such as energy consumption. Results revealed participants' responses differed between the implicit and explicit measures, indicating that participants attempted to conceal behaviours considered to be unfavourable whilst declaring their engagement in energy-conserving practices (Gamberini et al., 2014).

Behaviour considered to challenge societal norms which is often accompanied by negative characteristics (such as drunk driving, drug use, and violence and aggression) is frequently under-reported in surveys (Brenner & DeLamater, 2014). It is expected that such under-reporting is more evident in interviewer-administered surveys as opposed to self-administered or online questionnaires, however this isn't always the case. While social desirability bias has commonly been used to explain over-reporting of some normative and positive behaviours such as physical activity, it has at times been insufficient in fully accounting for or consistently confirming the relationship between social desirability and over-reporting (Brenner & DeLamater, 2014).

Physical aggression in particular, is not only seen more often among children, but it is also considered one of the most socially undesirable behaviours based on the results of an experiment utilising self-report methods in the measurement of physical, verbal and indirect aggression (Anguiano-Carrasco et al., 2013). Higher physiological responses have also been found on measures of indirect aggression, suggesting that questions pertaining to the use of indirect aggression are liable to the influence of social desirability bias more so than physical aggression, anger, and proactive aggression (Poltavski et al., 2018). Social desirability also appears to influence how often females make reports of partner abuse. Women with high scores on social desirability measures were found less likely to report incidences of "perpetrating psychological aggression, physical assault, and sexual coercion" (Bell & Naugle, 2007, p. 25). Being perceived as a victim of physical aggression by a partner also seemed to be considered another socially undesirable behaviour. As such, results indicated a lower likelihood of reporting abuse as women, whereas males appeared less influenced by social desirability (Bell & Naugle, 2007). In general, individuals tend to be less willing to report engagement in negative and socially undesirable behaviours such as physical aggression. However, men especially have been found less likely to report both their own and their partner's verbal and physical aggression (Riggs et al., 1989). Likewise, individuals are equally unwilling to report their own engagement in negative behaviours, especially if such behaviours are considered socially undesirable or more severe in comparison (Riggs et al., 1989). The results from this study provide support for the notion that in order to appear better than we truly are, we must minimise our own aggressive behaviours whilst highlighting the same behaviours in others.

Although, identifying negative behaviours in others doesn't always equate to making another feel superior. Individuals with a DSM-5 diagnosis of Intermittent Explosive Disorder (IED) seem to be influenced by social desirability to a lesser extent than their non-diagnosed counterparts and are also less likely to minimise the socially undesirable responses by others (Steakley-Freeman et al., 2018). Interestingly, participants with an IED diagnosis were posited to have a greater awareness of their own negative behaviours which pre-empted such a diagnosis, and as such, were found to exhibit higher motivations to change. These findings suggest that if an individual has the ability to acknowledge their own negative behaviours whilst also emphasising undesirable phenomena, they may be less impacted by social desirability bias. Denying or avoiding the provision of truthful responses is thus considered incompatible with the acceptance, acknowledgement, and eventual treatment of aggressive behaviours (Steakley-Freeman et al., 2018).

Intimate partner violence (IPV) is another area of research associated with negative consequences and harm to another (Freeman et al., 2015). IPV is however another sensitive topic affected by social desirability bias with differing perspectives as to whether or not IPV has occurred within a relationship. Males who engaged in more impression management (IM) i.e., an increasingly deceptive form of social desirability tended to decrease their admittance of using psychological aggression against a partner. However, males were more likely to use negotiating tactics (as identified by their partners) when impression management tactics were also in play (Freeman et al., 2015). In comparison to females, males were also found more likely to under-report the incidence of IPV, and thus over-report engagement in behaviours which are considered prosocial (i.e., an IPV-free relationship) to portray their relationship as stable and positive in social contexts (Freeman et al., 2015). The portrayal of more socially desirable behaviours also appears to extend so far as an offender population, with IM found to predict recidivism among a correctional sample (Mills et al., 2003). Offenders high in IM tended to exhibit greater interpersonal abilities and therefore engage in less criminal behaviours in an interpersonal context. However, the reverse was found for offender's low in IM scores, with greater incidences of violence observed among this population of recidivist offenders (Mills et al., 2003). A positive finding resulting from this study however is the evidence found which supports self-report methodologies as an effective method to predict both general and violent recidivism among samples of offenders. A relationship has been found to exist between antisocial attitudes and recidivism rates, indicating the more antisocial an individual is assessed as being, the greater their risk of engaging in criminal behaviour again (Mills & Kroner, 2005).

However, when IM is removed from the equation, the relationship between antisocial attitudes and rates of recidivism is weakened. More interestingly however is the result that higher levels of antisocial tendencies tended to produce more honest responses on self-report measures among a correctional sample (Mills & Kroner, 2005). Perhaps this finding could be explained by the idea that an individual already involved in the criminal justice system as a result of their engagement in antisocial behaviours has little to lose, and therefore has little interest in appeasing the views of others. As a result, an increase in honest responses is observed, as social desirability may be considered contradictory to achieving desired outcomes.

As has been demonstrated, impression management and social desirability biases can influence the reporting (or not) of negative behaviours such as IPV and recidivism rates of both general and violent offending (Mills & Kroner, 2005). However, survey results can

equally be influenced by the desire of participants to engage in impression management which manifests in the over-reporting of socially acceptable and normative behaviours. Each of these key factors form the basis of identity theory; a theory which proposes one's identity can be conceptualised by two factors: prominence and salience (Brenner & DeLamater, 2016). Prominence refers to the idealistic feeling an individual has towards their identity with no real knowledge as to the likelihood of this manifesting in their behaviour. Salience refers to the likelihood of one enacting their idealistic identity. There is also thought to be a positive causal relationship between the value one places upon their idealistic identity and the enactment of specific behaviours. For example, individuals who place a high and idealistic value on exhibiting a particular identity are more likely to inaccurately identify one's engagement in certain behaviours (i.e., exercise) in an online survey, as a means to maintain their ideal self. In essence, if an individual's idealistic identity is one who engages in regular exercise to appear athletically fit and able, then that individual may be more likely to respond in a manner that maintains their view of self; responses which may not reflect one's true engagement in regular exercise. In comparison, there is a negative causal relationship found between over-reporters and the actual occurrence of particular behaviours (Brenner & DeLamater, 2016). In these cases, over-reporters have a tendency to over-estimate how often they engage in desirable behaviours (i.e., exercise) in order to appear prosocial. Furthermore, the higher an individual is in prominence, the more likely they become to engage in over-reporting whereas the reverse effect is evident for participants with low prominence.

Consequently, both prominence and salience are likely to contribute towards the expression of bias and impression management in surveys as participants battle with the desire to represent oneself in a manner that is congruent with their idealistic identity. Brenner and DeLamater's (2014) study provided further verification of the positive relationship between the value attributed to one's identity and the regularity of over-reporting which is consistent with the findings presented by Brenner and DeLamater (2016). Such over-reporting is explained in both studies as a means to maintaining one's idealistic identity as per the factors of prominence and salience outlined in identity theory.

Reporting information in general can also have serious implications depending on the nature of what information is reported. Rape is one example where the information reported may not be entirely accurate given the sensitive nature of such a claim. Rape is defined as a lack of consent to engage in sexually autonomous behaviours and may include the use of force and intimidation to create a sense of fear and submissive behaviours in the victim (Dowds, 2020). In the case of rape reports, the psychological, financial, social, and legal

ramifications of making such a report often dissuade individuals from reporting such crimes in the first place which makes it difficult to obtain accurate information regarding the frequency of such violent crime (Allen, 2007). Reasons for not reporting the occurrence of a rape was most often attributed to the perception that such an attack was related to a personal matter or the view there would be insufficient evidence to attain a conviction. The result is a distinct under-reporting of rape as the costs associated with such reports are perceived to outweigh the incentive to report (Allen, 2007). If incentives to report are improved by way of providing additional social supports, for example, an increased likelihood of reporting on sensitive topics may be achieved. Differences in both self-under and over-reporting rates of arrest during adolescence and young adulthood have also been found to exist when compared to official reports (Emmert et al., 2017). Females were found less likely to report the frequency of being arrested as indicated by higher rates of both under- and over-reports. This discrepancy could be the result of females responding in a way which maintains their idealistic self while also demonstrating the influence of social desirability bias on measures of criminal behaviour.

Just as reports of rape and one's history of criminal activity are influenced by cost, incentives, and social desirability, so too is reported information on aggression. Fear, differences in perceptions and the ability and knowledge to report the incidence of aggressive behaviour has been found to influence how *often* nurses report becoming a victim of perpetuated aggression by a patient (Christensen & Wilson, 2022). Exposure to physical and verbal aggression within the workplace has been found to contribute to a higher rate of workplace injuries, as well as an under-reporting of near misses and accidents (Jiang et al., 2018). The common perception that reporting incidents of verbal aggression will result in retribution to the victim may encourage under-reporting behaviours, while exposure to physical aggression increases the likelihood of negative consequences for the victim (Jiang et al., 2018). Thus, if individuals are provided with the incentive to not only respond truthfully on measures of aggression but are also given assurances that such responses will remain confidential to create a sense of security, under and over-reports of aggressive behaviour may be kept to a minimum.

Questionnaire measures which rely on self-reported information and answers are often characterised by limited validity due to over-reporting of socially desirable behaviours such as church attendance and the frequency of engagement in physical activity by participants (Brenner & DeLamater, 2014). Older individuals, families and individuals with tertiary-level education have been found to indicate a higher rate of attendance at religious activities

however once social desirability was factored into the equation, older individuals were found to attend church at a similar rate to other age groups and family configurations (Larson, 2019). Individuals with an annual income of more than $120,000 (USD) were also found to consider environmental impacts at a higher rate than low-income individuals however when social desirability (SD) was considered, such results were no longer significant (Larson, 2019). These results indicate the effect SD has on portraying oneself as better than one really is – whether by over-reporting how often one attends church, how often one considers the impact of their behaviour on the environment, or by minimising how aggressive they are.

Self-report methods are however often used to assess one's engagement in risky health behaviours such as nicotine, alcohol, and substance use (Crutzen & Goritz, 2010). Results from this study indicated social desirability only appeared to have a small influence on self-reported substance use, as did lower educational levels. Individuals with less educational experience demonstrated higher levels of social desirability and increased substance use in recent times. Interestingly, individuals with greater educational prowess also exhibited higher levels of social desirability as indicated by lower self-reports of alcohol and nicotine use (Crutzen & Goritz, 2010). The results from these studies indicate social desirability may have limited effects on an online-based survey given the private nature of the online environment, which may promote rather than discourage the reporting of socially undesirable behaviours.

Asking direct survey questions specific to a particular behaviour are also anticipated to increase levels of over-reporting, regardless of whether the survey is self- or interviewer-administered. The use of online surveys is posited to generate responses by participants which are less likely to be impacted by social desirability bias, given the anonymity of the online environment (Gnambs & Kaspar, 2016). In contrast, results indicated both paper- and web-based surveys generated similar responses on items pertaining to positive personality traits and symptoms of mental health which suggests social desirability bias did not affect one administrative mode more than the other (Gnambs & Kaspar, 2016). Similar findings were reported by Dodou and Winter (2014) who found no difference between the suitability of web-based of paper-based surveys when it comes to the effect of social desirability. However, surveys which ask more sensitive questions were found to be less influenced by social desirability when the anonymity of participants is maintained (Dodou & Winter, 2014). The same has also been found for web-based surveys when compared to face-to-face interviews, which appear to be less impacted by social desirability (Heerwegh, 2009). Responses to web-based surveys have also been found to be better, which may be attributable

to the preference younger populations have for internet-based methods of survey administration. However, the quality of responses obtained has been found to be slightly lower for web-based surveys in comparison to paper-based or administrator led interviews (Heerwegh, 2009).

The findings above provide support for the hypothesis that there is more than social desirability bias at play when it comes to the measurement of normative behaviours. However, maintaining participants' anonymity in web-based surveys has been shown to lessen the impact of social desirability basis while also reducing levels of social anxiety and improving self-esteem among participants (Johnson, 1999). In comparison, paper-based methods of survey administration produced higher levels of social desirability and social anxiety (Johnson, 1999). As previous research has shown, surveys regarding prosocial behaviours are susceptible to social desirability bias, however, is this also the case in surveys measuring aggressive behaviours; behaviours that are perceived as less socially desirable? Research has found the BPAQ specifically is susceptible to social desirability bias, with the highest loadings found on statements pertaining to physical aggression (Vigil-Colet et al.,2012). Similar findings were reported for the anger subscale which had greater loadings in comparison to hostility. Interestingly, a higher use of indirect aggression tends to be displayed by people who are more affected by social desirability bias. These findings support the idea that social desirability can cause variations in how people respond to questions which measure socially undesirable behaviours such as aggression (Vigil-Colet et al.,2012).

Although online surveys can elicit more honest responses from participants and are thought to be less influenced by social desirability bias when compared to interviewer-administered surveys, the issue of maintaining a participant's attention and engagement level in the survey remains (Liu & Wronski, 2018). The use of "trap questions" (Liu & Wronski, 2018, pp. 32) is one method typically used in online surveys to identify participants who are lacking in attention in an effort to increase the accuracy of responses. The results of this study suggest using only one trap question which is easy to respond to; placement of the trap question near the end of the survey once all necessary questions or measures have been completed; and remit the use of an announcement prior to the trap question. Liu and Wronski (2018) also suggested that asking participants a subsequent question regarding whether or not they noticed the initial trap question, was insufficient in accurately identifying the attentive responders from the inattentive. The 'Captcha' trap question was assessed as being the most effective in identifying inattentive participants' (Liu & Wronski, 2018). Informing participants their answers would only be accepted if they demonstrated sound knowledge of

the survey has appeared to be the most effective method for improving attention, reducing non-responses, and increasing the amount of time a participant spent on a survey, which would thereby indicate they had a good knowledge of the survey content (Clifford & Jerit, 2015). Maintaining anonymity of participants' responses also serves a similar yet slightly lower influence on socially desirable survey responses (Clifford & Jerit, 2015). In essence, utilising the most effective method to maintain participant anonymity and limit the over-reporting of socially desirable behaviours and attitudes may also reduce the impact of SD on survey outcomes. Replicating these suggestions in a similar manner in our study may increase the quality of the data obtained from our survey whilst maximising the quality of responses received.

This study aims to address the factors of prominence and salience through the chosen questionnaires so to accurately measure the impact such factors may have on survey responses and outcomes. Ensuring anonymity of participants' responses to survey questions, regardless of whether they are directly or indirectly related to a specific behaviour, may also increase the likelihood of honest responses. When coupled with the BTS method, it is expected the outcomes of this study will accurately reflect individuals' perspectives towards, and the undertaking of aggressive behaviour in the general population.

Bayesian Truth Serum (BTS)

Bayesian Truth Serum (BTS) is expected to induce truth-telling in surveys on sensitive topics by advising participants a 'truth-telling algorithm' would be applied to their responses. In effect, a study utilising BTS practices would advise participants that higher scores would be applied to more honest responses, and thus, the expected payout post-survey is postulated to be larger in comparison to those who provide dishonest responses (Weaver & Prelec, 2013). BTS functions with the assumption that people will over-represent the majority of others' who share their perspective which is based on their own internal biases and assumptions, formed from their own experiences. In general, people believe themselves to be better than they are. They often downplay undesirable biases whilst promoting positive ones, all the while holding beliefs and biases which are consistent with those of their own in-group (Pronin, 2006). People's beliefs about themselves necessarily impacts on their subsequent judgements of others and the proportion of the population who hold views consistent with their own. Applying the BTS method can incentivise and generate more truthful responses on surveys by assigning an incentive to truth-telling and a penalty to dishonest answers (Weaver

& Prelec, 2013). However, when it comes to the measurement of behaviour that is considered immoral or denounced by societal standards, offering significant incentives may be insufficient in maximising honesty due to identification via survey payment schemes (Turner et al., 1998). By maintaining anonymity of survey participants and applying BTS methodology, participants may be more inclined to respond with greater care and honesty (Turner et al., 1998).

BTS operates on the premise that higher 'information scores' are applied to the most common response which is derived from the same sample population who provides the answers (Frank et al., 2017). The most 'common' response within a sample is thus considered the most truthful answer if they fit within the majority of overall responses (Georgieva, 2016). However, rare responses which do not necessarily 'fit' with the majority response are also privy to high scores, if the responses offered by participants match the predictions made by the same group of participants (Georgieva, 2016). Thus, if participants respond in a manner which is considered to be outside of the 'norm', the resulting indication is one of a dishonest answer. BTS utilises both individual responses and prediction responses to determine the truthfulness of one's perspective. For example, a coin is flipped ten times and participants are asked to report whether each toss is heads or tails. For each 'heads' that is reported, participants are rewarded with an additional monetary bonus. The same participant then predicts the number of coin flips which were reported to be 'heads' by all the other participants in the survey. For those participants who reported 'tails' (and thus received no additional monetary bonus) could demonstrate a higher degree of honesty. When BTS is applied to such an experiment, the difference in information scores between groups (experimental and control) is expected to be overt, thus illustrating an improvement in honest answers in the group which the BTS methodology is applied to (Frank et al., 2017).

On multiple-choice surveys based on self-reported answers, participants are asked to select both the option that aligns with their perspective on a particular matter (information score), and to make a prediction about how other participants will respond to the same matter (prediction score). Honesty of participants' responses is thereby determined by the sum of their information score in comparison to the accuracy of predictions made across each survey item (Frank et al., 2017). People want to believe that others in the general population share the same opinion as them, especially when it comes to the measurement of sensitive matters such as politics (Frank et al., 2017), scientific misconduct (John et al., 2012), or engagement in criminal or illegal activities (Nagin & Pogarsky, 2003).

BTS is considered an appropriate method to use for the purpose of improving survey responses (Frank et al., 2017). BTS utilises an incentive-based mechanism which offers a financial reward to elicit sensitive information and beliefs regarding a certain event (Offerman et al., 2008). BTS methodology is regarded as an improved foundation from which to measure subjective and concealed beliefs about a particular topic, however, it is not without its criticisms. In a study completed by Offermann et al. (2008), participants provided a probability estimate for an uncertain event in the stock market before being presented with a score dependent on whether each probability estimate was true or incorrect. Bayesian methodology was applied to participants' responses. While BTS is posited as the most successful method in eliciting individual attitudes and beliefs, it did not completely remove the influence of subjective opinions and ambiguous attitudes regarding the context of the stock market (Offerman et al., 2008).

In the context of research, the initial results of a study completed by van de Schoot et al. (2021) indicated the views academic leaders had of their research students were inconsistent in comparison to the actual views of the students regarding their practice of fraudulent research practices (van de Schoot et al., 2021). The method involved in this study involved the distribution of an online-based survey regarding the fabrication of data, removing extraneous outliers to achieve significant results, and the publication of three similar articles as opposed to one which uses the same data set across each publication. All of these factors were considered to be questionable research practices. BTS was then applied to participants' responses to determine whether or not an honest response was provided, and to establish whether a difference could be observed regarding individuals' attitudes towards fraudulent research practices. When BTS methodology was applied, the results indicated that an increasing number of research candidates admitted to the acknowledgement of an individual who did not contribute to research in a significant manner; a practice known as "gift authorship" (van de Schoot et al., 2021). More participants also indicated a belief that their peers would partake in the same practices however would not admit doing so and expressed intentions to share in questionable research practices when exposed to a high-pressure environment. Furthermore, the moral and normative behaviours of a research supervisor were found to contribute to fraudulent practices, as demonstrated with the use of BTS (van de Schoot et al., 2021).

BTS has been applied in the field of economics, in which a financial incentive was offered to participants playing a simple two-player game. Results indicated that offering a financial reward encourages participants to provide more accurate responses about their own

behaviour while discouraging the falsification of their true beliefs (Traumann & van de Kuilen, 2015). BTS has also been cited as a strong mechanism capable of eliciting honest and truthful responses regarding product reviews and ratings, which are crucial to the success of quality control practices and online marketplaces (Kamble et al., 2018). An incentive-based framework in this context is regarded as fundamental to the operation and reputation of online platforms which rely heavily on unbiased feedback and reviews.

Contingent valuation (CV) is another area Bayesian truth serum (BTS) has been applied in; an area in which estimates are made as to the value individuals place on environmental products (Barrage & Lee, 2010). In this context, "hypothetical bias" (Barrage & Lee, 2010, p. 140) is a real concern, which is the result of participants responding in a different manner depending on whether or not they are financially bound to respond one way or another. However, the application of BTS in this context did not appear to mitigate the influence of such bias whereas the use of consequential 'expectations' was more effective in eliminating bias (Barrage & Lee, 2010). In effect, if participants believe their responses will invoke a particular change due to set expectations, they are more likely to respond in a truthful and accurate manner.

A list experiment design in which participants are asked to indicate the number of statements they agree with; agreements which remain hidden from the researcher, has also been successfully employed to measure sensitive matters such as unprotected sexual behaviours and the prevalence of violence among intimate partners (Lepine et al., 2020). Although the study did not specifically utilise BTS methodologies, the selected methodology which was applied included similar aspects to those observed in BTS. The survey administrator was not privy to the personal information of participants, thus maintaining their anonymity, which is a key aspect of BTS. Bias was reduced and stigmatisation of sensitive health behaviours appeared to be limited, allowing for a more accurate estimate of the prevalence of unsafe sexual behaviours, sexually transmitted diseases, and the occurrence of intimate partner violence (IPV) (Lepine et al., 2020). BTS has also been successfully applied against other socially undesirable behaviours including self-deception, impression management, lying, over-estimation of one's knowledge about a subject, and the choice to promote oneself over others (Simunovic & Zezelj, 2023). While the use of the BTS method could not eliminate the effect of social desirability completely, the success of the method in mitigating the effect of socially desirable responding was proven.

In the context of economics, it is suggested a person's willingness to accept (WTA) in exchange for a good or product is not significantly different from the maximum price they are

willing to pay (WTP) to purchase goods or products (Georgieva, 2016). Bayesian methods were applied in this context with the goal of obtaining honest responses about WTA and WTP practices, despite the real truth being unknown. They utilised an incentive-based framework under Bayesian methodologies to determine users' WTA and WTP for a new product created by Uber; a taxi-based service for drivers and passengers. Results did not support the hypothesis, indicating the application of BTS methodology did not highlight a difference between groups (Georgieva, 2016). Despite this, BTS was still able to generate more accurate data in comparison to non-incentivised approaches. The financial incentive-based framework which BTS operates under *encourages* participants to provide more honest responses, albeit not always significantly, yet is posited as a more reliable method in contexts where the truth remains to be known (Georgieva, 2016).

Within this study, predictions about aggressive behaviour are collated from the entire sample population to determine the "most common" and "most uncommon" responses. This allowed the survey administrator to determine whether or not participants were responding truthfully based on their personal circumstances, perspectives and behaviour, without the administrator ever having to know such personal information. Assigning higher information scores to participants' responses (information report) in comparison to others' perspectives (prediction report), responses of which are all drawn from the same testing population, will enable predictions to be made about the truthfulness of participants' responses. Based on this BTS premise, it is expected that in this study, participants in the experimental condition will demonstrate higher information scores in comparison to the control group (based on the most common response provided across the testing population). This outcome is expected due to the disclosure of financial compensation to participants in the experimental condition; thus, creating an incentive which is expected to elicit a larger degree of truthfulness from participants assigned to this condition. This study is expected to demonstrate that when BTS is applied correctly, information about highly personal and sensitive topics from people in the general population can be accurately understood which may not otherwise be possible through other, and more intrusive means (i.e., face to face interview). Data obtained from the experimental condition (when compared to data obtained from the control condition) will demonstrate that when BTS is applied to a survey, more truthful and accurate responses can be obtained. BTS will therefore improve the quality of survey responses which we anticipate when applied correctly, will subsequently increase predictive validity in psychological research.

Materials & Methods

Participants

The sample population consisted of 289 participants who were registered survey participants of Prolific; an online platform specialising in the facilitation of reliable participants for scientific research. No identifying variables were collected for the purposes of this research to maintain the anonymity of participants. The only prerequisites for participating in the research were that participants were a registered member on Prolific and that their first language was English. The participants demographics are displayed in Table 1.

Table 1.
Demographics of participants

Characteristic	Female	Male	Total
Age 18-24	92 (31.83%)	47 (16.26%)	139 (48%)
Age 25-40	90 (31.14%)	36 (12.46%)	126 (43%)
Age 41-60	15 (5.19%)	9 (3.11%)	24 (9%)
Age Total	197 (68.17%)	92 (31.83%)	289 (100%)

Data was also collected across three geographical regions (Americas, Europe and Africa). The ‘Americas’ included participants from both the US and South America. Geographical data is displayed in Table 2. There were participants from another 21 countries, all of whom had a total number of participants of less than 2% of the overall sample group. When this data was computed to regions of the world, as defined by the World Health Organisation (WHO), the highest number of participants resided in the Americas, Europe and Africa.

Table 2.
Geographical information of participants

Region	Participants
Americas	121 (41.86%)
Europe	87 (30.10%)
Africa	7 (2.42%)
Other	4 (1.38%)

Design

This study employed one independent variable (application of BTS) as a repeated-measures variable. A second independent variable (informing participants of a “truth-telling

algorithm") was utilised as a between-subjects variable. The dependent variables were the aggression questionnaires (POAS and adapted BPAQ-SF).

For the between subject variable, participants were randomly assigned by Qualtrics to one of two conditions – a control group (CNTR) or an experimental group (EXP). The control group (CNTR) included 148 participants (51%) and 143 participants (49%) were assigned to the experimental group (EXP). By assigning participants to one of two groups, we aimed to determine whether there was a difference in how people respond to questions about aggressive behaviour and traits, if some individuals were provided with further information upon completion of the first survey which advised them a "truth-telling algorithm" (BTS) would be applied to their next set of answers, would they respond differently, and better yet, would they be more truthful than they may have previously been?

The survey was constructed in Qualtrics before it was distributed via the online platform Prolific.

Materials

The Perception of Aggression Scale (POAS) was displayed to participants first, which assesses individuals' views on whether aggressive behaviour is considered a dysfunctional and incomprehensible phenomenon, or whether it is viewed as a functional and desirable phenomenon. Statements posed to participants can be found in Appendix A. The second questionnaire displayed to participants was an adapted short-form version of the Buss-Perry Aggression Questionnaire (BPAQ-SF) which measures aggressive behaviour in four different domains – physical aggression, verbal aggression, anger and hostility. Statements posed to participants can be found in Appendix B.

The Perception of Aggression Scale (POAS) is a tool developed for the purpose of understanding the perspective frontline staff hold towards aggression displayed by psychiatric patients at an inpatient facility (Benedictis et al., 2012). While it was initially developed as a 32-item scale and applied in a psychiatric inpatient facility in Sweden, it has also been adapted to a 12-item version and used successfully in the same setting. The use of the 12-item version amongst this environment yielded results which supported the POAS as a stable measure of underlying, and subjective attitudes towards violence and aggression. Items in the POAS correspond to whether an individual views aggressive behaviour as a dysfunctional and incomprehensible phenomenon or as a functional and comprehensible phenomenon. For example, 'aggression is an unpleasant and repulsive behaviour' is one item included in the subscale of 'aggression as a dysfunctional/undesirable phenomenon' – a 'yes' response to this

item would indicate the individual has a negative attitude towards aggressive behaviour. In contrast, 'aggression is a healthy reaction to feelings of anger' is an item included in the subscale for 'aggression as a functional/comprehensible phenomenon' – a 'yes' response to this item would indicate the individual has a positive attitude towards aggressive behaviour.

Psychometric properties of the POAS also reveal evidence of the tool as an effective and useful tool in the measurement of perceptions towards aggression. The full version of the POAS was initially developed with a total of 60 items, and demonstrated significant internal consistency with alphas of .87, .82 and .50 across three key dimensions (Jansen et al., 2005). A brief version of the POAS was later developed, yet the shortened version also demonstrated positive internal consistency, with alphas of .67 found for 'aggression as a functional and comprehensible phenomenon' and .69 for 'aggression as a dysfunctional and undesirable behaviour'. Equal support has been found for the brief version in a Swiss sample of psychiatric nurses, which demonstrated "Cronbach's alpha of 0.69 for factor 1 and 0.67 for factor 2" (Needham et al., 2004, p. 39-40). Test-retest reliability of the brief POAS also demonstrated significant results, as indicated by $r = 0.76$ for subscale 1 ('aggression is a functional and comprehensible phenomenon' and $r = 0.77$ for subscale 2 ('aggression is a dysfunctional and incomprehensible phenomenon' (Needham et al., 2004).

For the POAS, two of the 12 items were adapted slightly to provide a more generalised overview of aggression outside of a medical context. The original items which were adapted for the purposes of this study were included in the subscale which measured aggression as a functional/comprehensible phenomenon – 'aggression is the start of a positive nurse-patient relationship' and 'aggression is an opportunity to get a better understanding of the patient's situation'. These items were adapted to 'aggression is the start of a positive person to person relationship' and 'aggression is an opportunity to get a better understanding of the person's situation', respectively.

The second survey displayed to participants was an adapted version of both the Buss Perry Aggression Questionnaire Short Form (BPAQ-SF), which included 12 items, and the original BPAQ, which includes 29 items. As a result, a 17-item survey was developed (Appendix 'A'). There were two reverse-coded questions in the full BPAQ which were not included in this study – 'I can think of no good reason for ever hitting a person' and 'I am an even-tempered person'. These questions were removed for the purpose of simplifying the project.

The short-form version of the BPAQ ordinarily includes 12 items across four subscales (physical aggression, verbal aggression, anger and hostility). Internal consistency

of the BPAQ has also been regarded as significant, as evidenced by use of the alpha coefficient on a sample of 1253 participants with results detailing alphas of "Physical Aggression, .85; Verbal Aggression, .72; Anger, .83; and Hostility, .77 (total score = .89)" (Buss & Perry, 1992, p. 455). Similar results were found by Gerevich et al. (2007), which demonstrated a Cronbach alpha of 0.82 for 'physical aggression', 0.68 for 'verbal aggression', 0.70 for 'anger' and 0.75 for 'hostility'. Test-retest reliability figures provide further evidence to support the use of the BPAQ across time, with a significant mean score of 0.81 observed across each subscale (Webster et al., 2015).

Four of the 12 items in the BPAQ-SF measure physical aggression. To obtain a more robust measure of aggression, and specifically physical aggression, we included additional questions from the 'physical aggression' subscale of the full BPAQ, which includes nine total items to measure physical aggression; a factor of aggression we anticipated would be more closely associated with aggressive behaviour than other facets of aggression measured by the BPAQ. We therefore included each of these nine 'physical aggression' items in the survey we completed.

Procedure

Participants were asked to provide a response to questions pertaining to the measurement of aggression after being assigned to one of two conditions. The Perception of Aggression Scale (POAS) was displayed to all participants first. The second questionnaire displayed to participants was an adapted short-form version of the Buss-Perry Aggression Questionnaire (BPAQ-SF). The experimental group received information that a truth telling algorithm had been applied to their answers in the POAS before completing the BPAQ-AF. No detail on the algorithm was provided. The control group was asked to complete the second questionnaire without further instruction.

Both groups were asked to report on their own behaviour and beliefs (information report), and to also indicate what proportion of the population they believed would endorse the same view (prediction report). BTS methodology was applied to *all* of the participants' responses. All participants provided their informed consent prior to participating in the research. As a reward for their participation in the research, all participants received financial compensation. Prolific implement their own ethical payment principles, with a minimum hourly rate being £6.00 and the recommended hourly rate being £9.00. As we only required participants to complete an easy-use online survey with minimal effort, we set compensation

at £7.50 per hour, which was regarded as a "good" level of remuneration based on Prolific's ethical payment principles.

All participants were advised the survey was expected to take no longer than 15 minutes to complete. Participants were thanked for their time following completion of the questionnaire, after which they received their financial compensation. The amount of financial compensation received by the participants was the same, regardless of which group they were assigned to. Participants' responses were collated and stored securely on Open Science Framework (https://osf.io/).

Results

Data Preparation

The data obtained from the completed survey was analysed using SPSS (version 28) and 'R' (R Core Team, 2020).

One data set was removed prior to data analysis due to incomplete responses, while four additional data sets were removed due to a failure on the 'test' question in which participants responded 'yes' to owning a three-headed dog at some point in their life. The response times for these data sets were also incredibly quick in comparison to the remaining participants (370 sec average vs 511 sec average respectively), which suggests the responses provided may have been inaccurate. The data set for the two participants who identified themselves as being in the 'non-binary' gender category were also removed from analysis as they were both assigned to the experimental condition. As such, their responses may have acted as a confounding variable once data analysis was completed. As there were only a few participants in the oldest age group (Table 1), the data set for these participants were combined with the middle age group to create one older age group (24-60 years of age) and one younger aged group (18 – 24 years). The final sample population included 289 participants after exclusions were made.

BTS scores were calculated for both questionnaires according to Schoenegger (2023). BTS methodology assigns higher scores to more frequently endorsed responses than those which are collectively predicted, and rewards participants based on such responses. An information score is then calculated based on the most 'common' response. To obtain an information score, the frequency of the most 'common' response is calculated alongside the mean of that answer's predicted frequency, which is derived from the n^{th} root of the product of *all* predicted answers. The information score (i-score) is calculated by the relative

frequency of an answer, divided by the geometric means of an answer's predicted frequency $\left(\frac{\text{Relative Frequency of Answer}}{\text{Geometric Mean of Answer's Predicted Frequency}}\right)$ which then rewards participants directly based on their i-score and their accuracy of predictions made about others' responses' (Schoenegger, 2023).

All the participants' responses were coded into different categories as they pertained to each of the two questionnaires which were completed. For the first part of the survey (POAS), participants' responses were coded into 'desire' and 'undesire' as they pertained to the moral alignment of the questions posed ('aggression is a desirable/functional phenomenon' and 'aggression is an undesirable/dysfunctional phenomenon'). For the second part of the survey (BPAQ-SF), participants' responses were coded into four separate categories to reflect the subgroup of the assessed factor – physical aggression ('physagr'), anger ('anger'), hostility ('hostile') and verbal aggression ('verbagr').

Based on the inclusion criteria described in the methods section, coupled with the data sets which were removed prior to analysis (7), the final data set for analysis was 289 participants.

Primary Analysis

Pre-manipulation checks

All participants completed the Perception of Aggression Scale (POAS) first, prior to the manipulation (the application of BTS). Participants' responses were combined into one overall score and two separate categories – 'undesire' and 'desire 'as described above. 'Undesire' combined the responses for questions 1-6, and the 'desire' category combined responses for questions 7-12. The labelling of these categories was derived to reflect the moral alignment posed by the corresponding questions, in that questions 1-6 relate to aggression as an undesirable behaviour, and questions 7-12 relate to aggression as a desirable behaviour.

We completed an independent samples t-test to determine if group responses differed between the experimental and control group overall and in their responses to questions relating to aggression as a desirable trait and an undesirable trait. Results are displayed in Table 3. As predicted, there was no difference in BTS scores between the experimental group and the control group for the POAS overall, nor if the scores were analysed according to aggression as an undesirable phenomenon. However, the aggression as a desirable

phenomenon returned a significance level at 0.053 indicating that the groups may differ in their views on aggression as an acceptable trait or how they report on these.

Table 3.
Descriptive statistics for the POAS.

Category	Condition	N	Mean	SD	df	t	p
POAS all	Cntrl	148	-8.82	1.27			
	Exp	141	-8.86	1.24	287	.231	.661
POAS undesirable	Cntrl	148	-16.68	2.43			
	Exp	141	-16.79	2.43	287	.403	.894
POAS desirable	Cntrl	148	-0.97	0.39			
	Exp	141	-0.92	0.49	287	-.893	.053

Note: POAS = Perception of Aggression Scale, Exp = experimental group, Cntrl = control group

Post-manipulation analysis

An independent Samples T-Test was completed to determine whether there were any differences in BTS scores between groups after the experimental group had been told that a truth telling algorithm had been applied. The results demonstrated a non-significant difference between the control group and the experimental group which indicates the application of BTS methodology had little impact on participants' responses (Table 3). To get a more nuanced understanding we compared each category separately. No significant difference was found for the physical aggression, verbal aggression, hostility or anger category (Table 4).

Table 4.
Descriptive statistics for the BPAQ-SF

Category	Condition	N	Mean	SD	df	t	p
BPAQ-SF all	Cntrl	148	-1.85	0.31			
	Exp	141	-1.85	0.46	287	-.04	.198
Phys Agr.	Cntrl	148	-1365	0.32			
	Exp	141	-1.64	0.49	287	-.26	.161
Verb. Agr.	Cntrl	148	-2.13	0.46			
	Exp	141	-2.14	0.53	287	.25	.732
Anger	Cntrl	148	-0.80	0.37			

	Exp	141	-0.86	0.31	287	1.36	.245
Hostility	Cntrl	148	-2.87	0.65		-.31	
	Exp	141	-2.84	0.70	287		.329

Note: BPAQ-SF = Buss Perry Aggression Questionnaire – Short Form; Cntrl = control group; Exp – experimental group; Phys Agr = Physical Aggression; Verb. Agr = Verbal Aggression

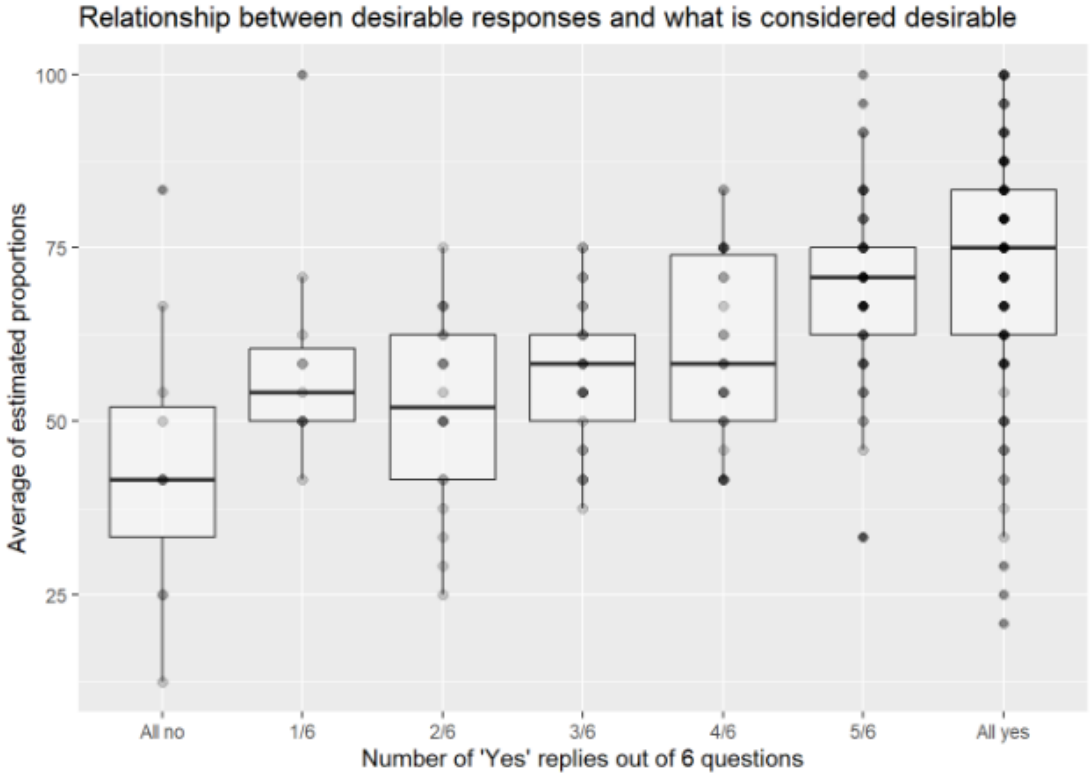
Exploratory Analysis

Part One: POAS

Given the differences in responses to the POAS for the aggression as a desirable phenomenon that almost reached significance, we conducted an exploratory analysis of the data set to gain a more in depth understanding of our results.

When it comes to analysing the relationship between desirable responses and what is considered desirable (in the general population) for the first survey (POAS), we can see several things. When comparing participants' responses to questions, those who answered 'no' to all questions (40%) differed to those who answered 'yes' to all questions (75%). This would suggest the majority of people perceive aggression to be an undesirable or dysfunctional phenomenon, and they also believe the majority of the general population share the same perspective. We also see the greatest amount of variance in 'all yes' responses which indicates participants hold quite different opinions about aggression as an undesirable behaviour. Results are displayed in Figure 1.

Figure 1. Desirable responses by question

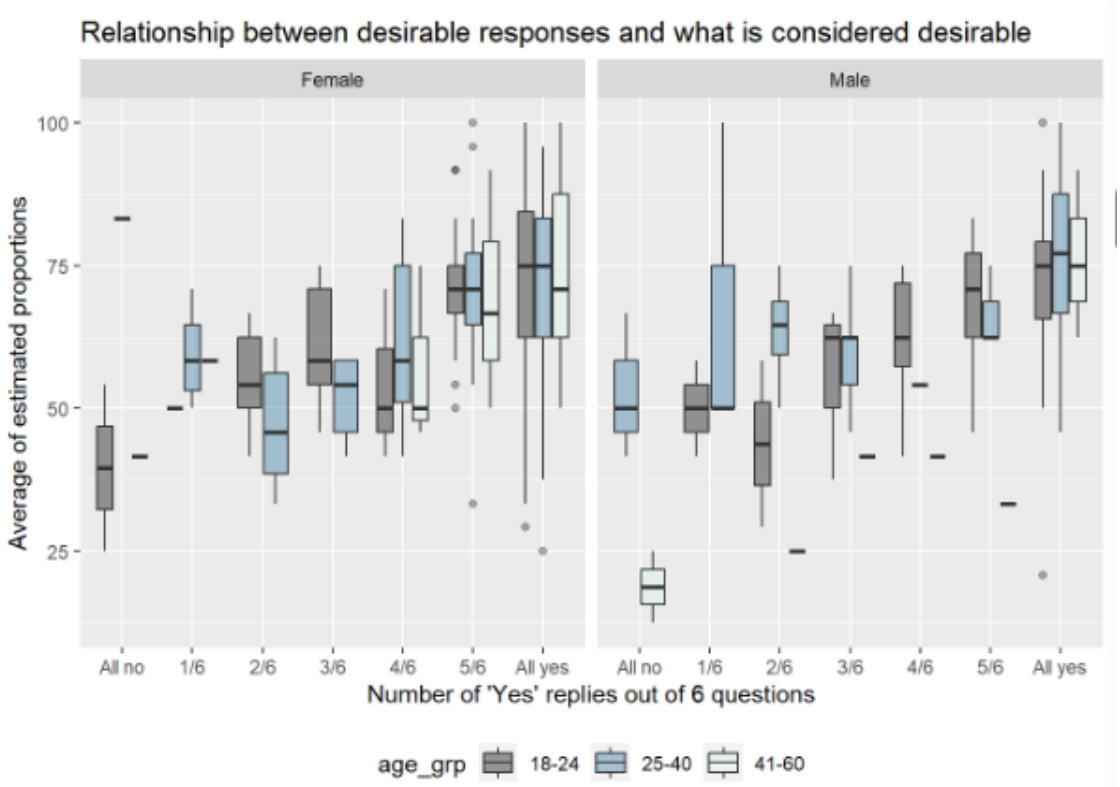


We started by comparing the relationship between desirable responses and what is considered desirable based on gender and age (Figure 2). For females, the only participants who answered ‘no’ to each question, as in, they did not view aggression as a desirable behaviour, were in the youngest age category (18-24). For the participants who answered ‘yes’ to every question, in that they viewed aggression as a functional and comprehensible phenomenon, participants in both the youngest and middle age category had the same median. Participants in both of these age groups viewed aggressive behaviour as desirable, and also believed that 75% of the general population shared the same view. We also see the greatest amount of variance in all ‘yes’ responses in comparison to all ‘no’ responses for female participants.

For the male counterparts, participants aged 41-60 featured prominently in all ‘no’ and all ‘yes’ responses which indicates a variety of views on aggression. Interestingly, participants in this age group had a tendency to answer all ‘yes’ or all ‘no’ to each statement, which is where we see the greatest amount of variance in responses. This also indicates that participants in this age group share perspectives from one extreme to the other, in that they either see aggression as an undesirable phenomenon or they see aggression as a desirable behaviour. For all ‘no’ responses, the majority of responses fell below the average response for the other two age groups; with the ‘top’ response being 25% and the lowest being 0%,

This indicates that while some participants viewed aggression as a desirable response to a situation, they did not hold high estimates that others in the general population would share the same perspective. For all ‘yes’ responses, the average estimate of the same view in the general population was the same for both the youngest and the oldest age groups, with both of these groups estimating 75% of the population shared the view that aggression is a desirable and comprehensible phenomenon. The first statement posed to participants for the POAS was “aggression is the start of a positive relationship”. As evidenced by the responses given to this statement, we can see that for males aged 25-40, the majority of responses were in the upper half of the data set in that the majority of participants answered ‘yes’ and believed that 50% of the population shared the same perspective.

Figure 2. Desirable responses by Age and Gender

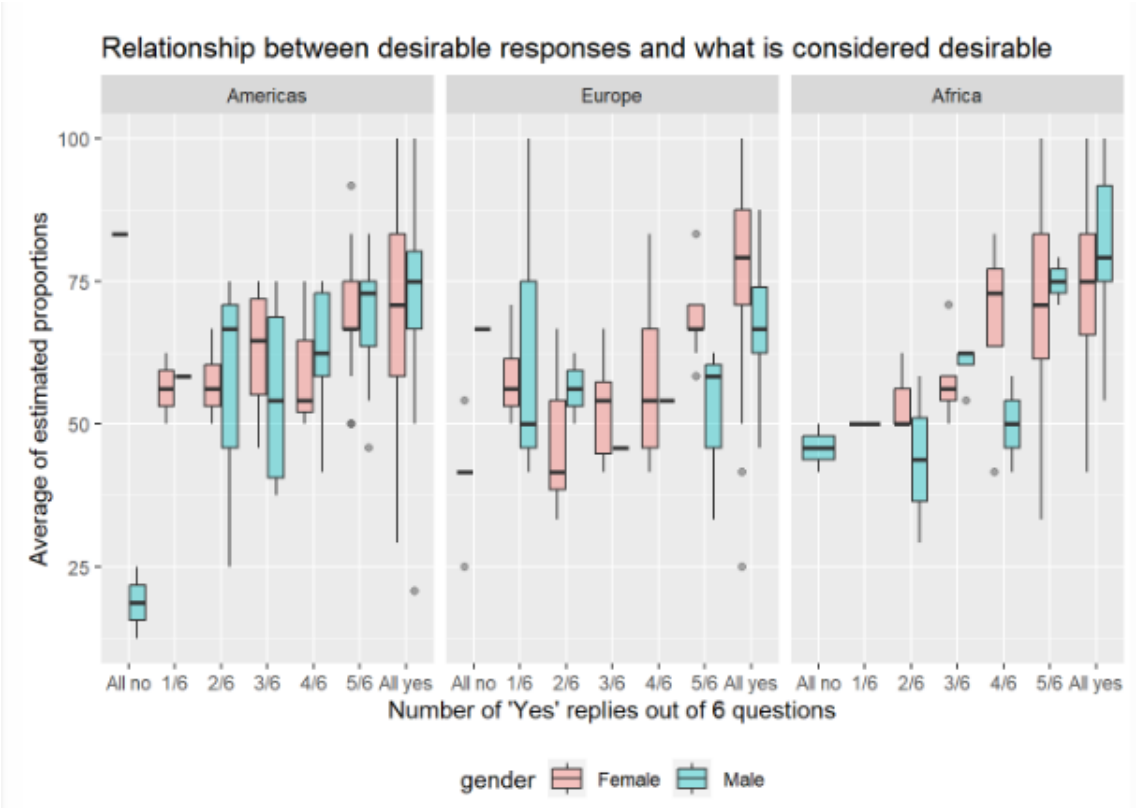


We then analysed the relationship between desirable responses and what is considered desirable, as they related to gender and region (Figure 3). Males in the Americas appear to have the most diverse responses as males in this region for all ‘no’ responses were much lower than their female counterparts, which also indicates some males view aggression as a desirable phenomenon however are not as confident that others share the same view. The average responses for females in the Americas was highest for all ‘yes’ responses, which is

also where we see more varied responses, with the majority of females perceiving aggression as desirable, and estimating almost 75% of the population concur with those perceptions.

In Europe, more females view aggression as desirable in comparison to their male counterparts, whereas more males viewed aggression as a healthy reaction to feelings of anger than their female counterparts. Males also had a higher average than females for question one, indicating that males tend to view aggression as the start of a positive relationship more so than females. Overall, it seemed females in Europe had more positive views of aggression as a desirable behaviour than their male counterparts, with the average number of responses for females being higher on questions three to six. In Africa, males were the only gender to answer 'no' to each statement, in that they did not view aggressive behaviour as a desirable trait. However, the average male responses for all 'yes' were higher than their female counterparts, indicating while some males view aggression as undesirable, a lot of males also view aggression as desirable. Question four asked participants whether or not they agree with the statement that "aggression is a form of communication and as such not destructive". Responses to this statement indicate the majority of males in Africa agree with this statement and believe 50% of the population also agree. Females on the other hand had a much more varied response, with the average response indicating females agree with this statement and perceive almost 75% of the population to hold the same view. Overall, males in the Americas and Africa tend to view aggression as a more desirable behaviour than their female counterparts, whereas females in Europe view aggression as more desirable than their male counterparts.

Figure 3. Desirable Responses by Gender and Region



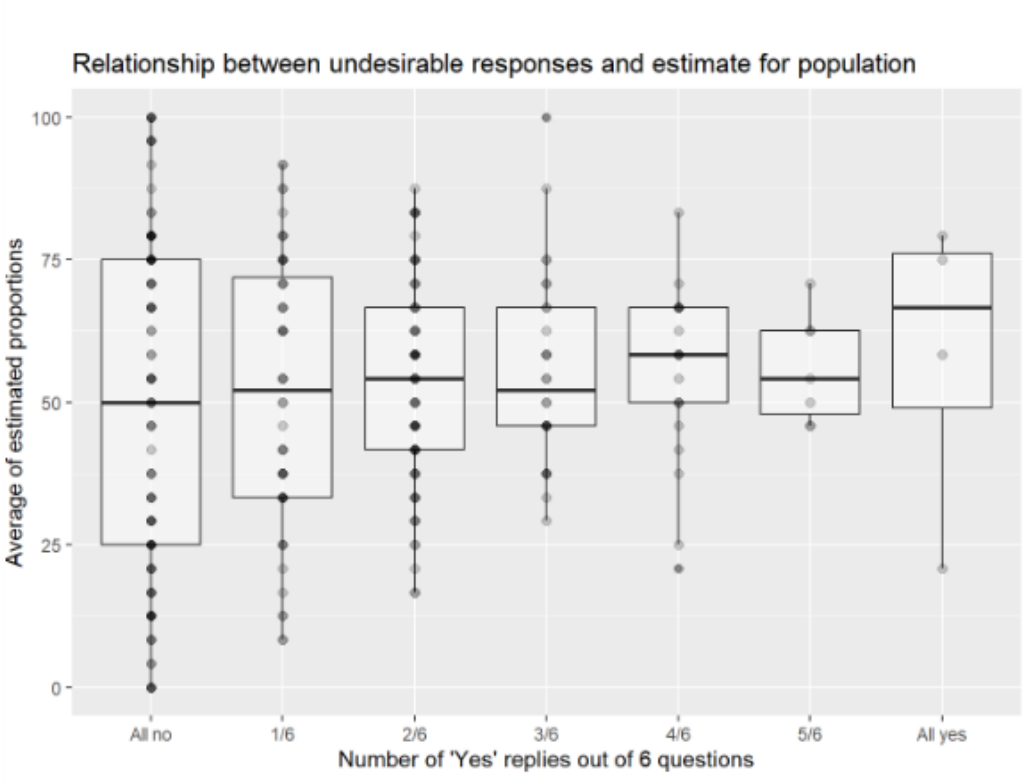
We then move on to consider the relationship between ‘undesirable’ responses and estimates for the general population (Figure 4). In essence, we analysed whether or not individuals perceived aggression as an undesirable or incomprehensible phenomenon and what percentage of the population they believed also held the same views. The majority of responses indicated most participants viewed aggression as an undesirable phenomenon in that we see the most ‘no’ responses to all questions posed in comparison to other questions on their own. We also see a greater variety of responses for all ‘no’ answers as some individuals saw aggressive behaviour as undesirable yet estimated that 0% of the general population shared the same view. At the other end of the scale, i.e., some participants agreed that every aspect of aggressive behaviour canvassed in the survey were undesirable and estimated 100% of the general population would also share their perspective. The majority of responses in this category (‘no’ to all questions) indicated most participants viewed aggression as undesirable and estimated 50% of the population to hold similar views.

In comparison to the above, a smaller number of participants viewed aggression as a desirable phenomenon in that they answered ‘yes’ to all questions posed, thus indicating they did not see aggression as an undesirable phenomenon. However, most responses in this category sit at a higher value when it comes to estimating how much of the general

population, they believe hold the same perspective towards aggression. Most participants estimated approximately 70% of the general population share the same perspective which suggests most individuals in this category view aggression as something to be desired and also believe a higher proportion of the general population are supportive of those views.

When we consider each question posed on an individual basis, we can see the majority of participants in these categories viewed aggression as an undesirable phenomenon and estimated 50-60% of the population corroborated those views. For example, question three asked participants whether or not they agree with the statement that 'aggression is hurting others mentally and physically'. The majority of participants' responses sit just above 50%, indicating that most participants agree with this statement and also believe just over half of the general population share the same view towards aggressive behaviour.

Figure 4. Undesirable Responses and Population Estimates

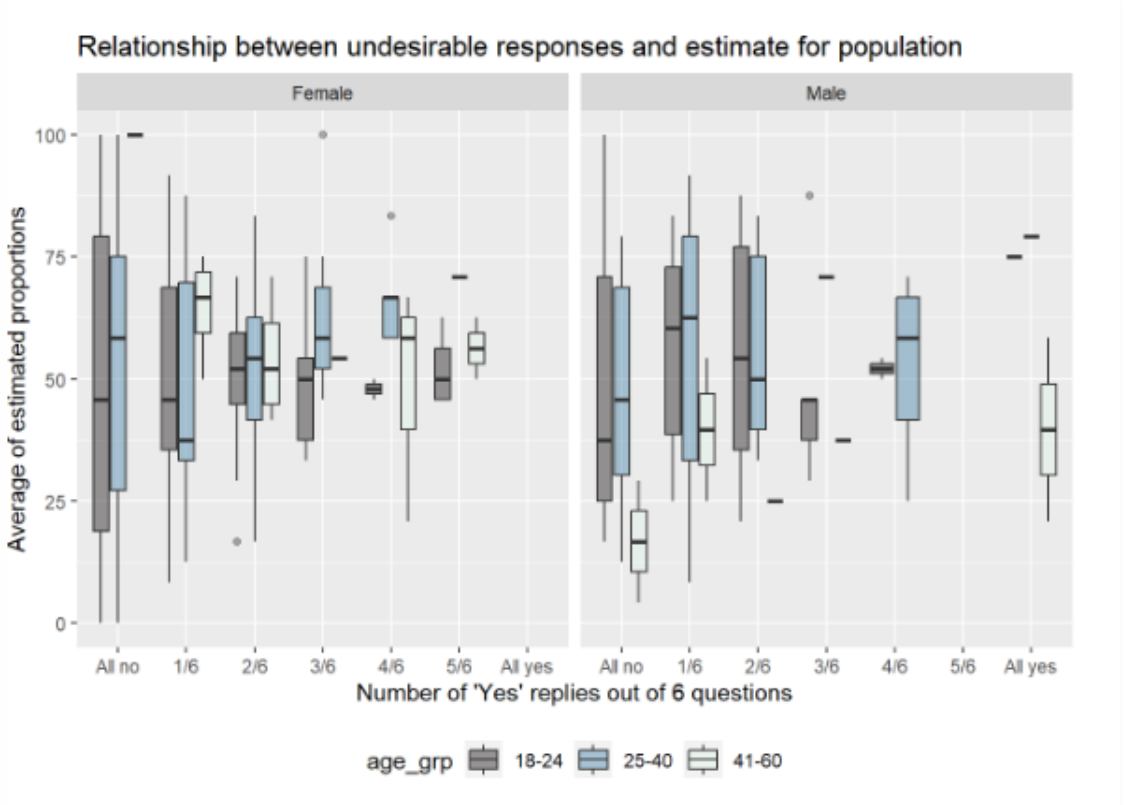


The same analyses were completed for gender and age group (Table 7). Females had a higher median of responses in comparison to their male counterparts for all 'no' responses, indicating that females were more likely to view aggressive behaviour as undesirable than males. This pattern was observed across each age group which suggests that females in the youngest, middle *and* oldest age groups were all more likely to view aggression as

undesirable in comparison to their male counterparts. Interestingly, no females in any age group answered ‘yes’ to all questions however some males in each age group did answer ‘yes’ to each question. This indicates some males hold very strong perspectives towards aggressive behaviour as a desirable phenomenon, some of whom also estimate a high proportion of the general population share those same views.

For question one, participants were asked whether or not they agree with the statement ‘aggression is an unpleasant and repulsive behaviour’ and were then asked to estimate the percentage of the population who they believe would also agree with their views. The results indicated the median number of responses for female and male participants in the youngest (18-24) and middle (25-40) age groups were higher for males than females, indicating that males had a more favourable view towards aggressive behaviour than their female counterparts. These results also suggest that younger participants held different views to their older-aged peers, with the median number of participants in the oldest age group for both genders being remarkably higher (females) or lower (males) than their counterparts. Females in the oldest age group had a higher median of responses when compared to females in the younger age groups for question one and estimated almost 75% of those in the general population shared the same perspective. Males in the oldest age category however had a lower median than their younger counterparts, thus indicating the majority of participants in this group did not feel as strongly towards aggression as an unpleasant and repulsive behaviour as their younger counterparts. When comparing these particular age groups (41-60) between males and females we can see that females appear to hold very contrasting views of aggression as an unpleasant and repulsive behaviour in comparison to their male counterparts. Females also estimate a higher proportion of the general population also agree with their view on this question.

Figure 5. Undesirable Responses and Population Estimates by Gender

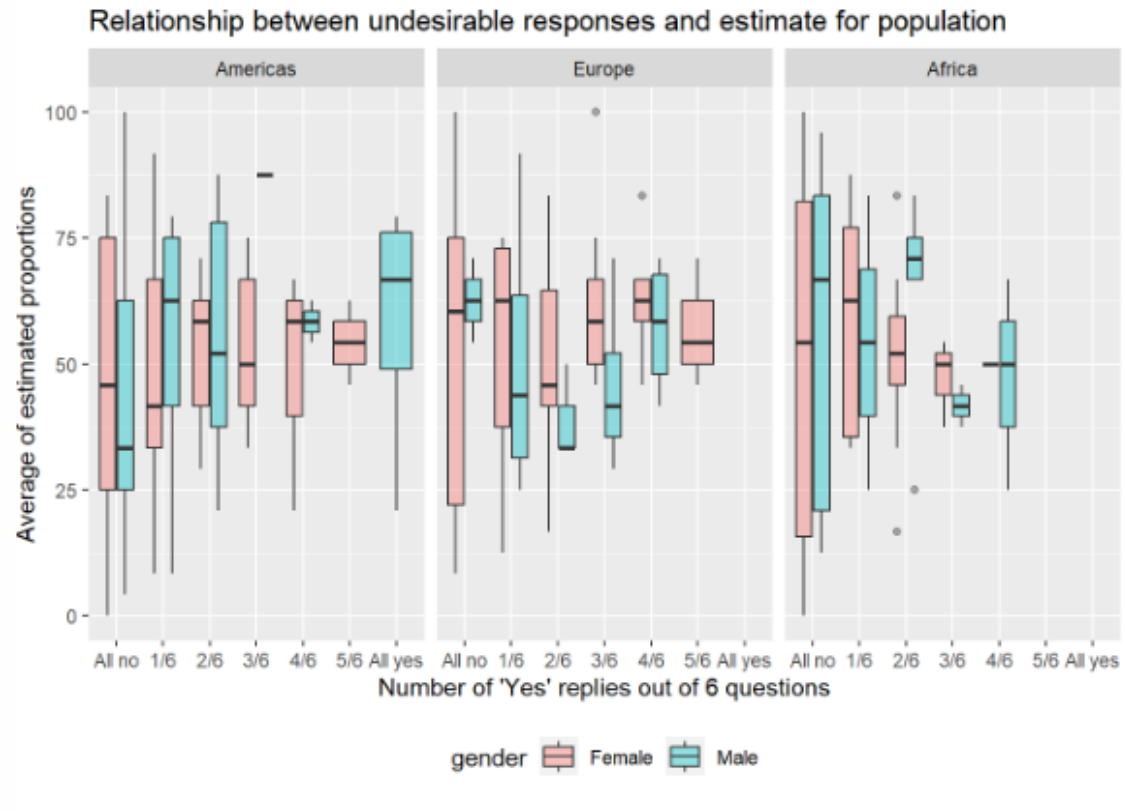


We then analysed the relationship between ‘undesirable’ responses and estimates for the population as they related to region and gender (Figure 6).

The majority of females in the Americas had a lower median in comparison to females in Europe and Africa for all ‘no ‘answers, thus indicating females in the Americas were more likely to view aggression as a desirable phenomenon in comparison to their counterparts in other regions of the world. When it came to estimating the percentage of the general population who would also share the same perspective, their responses indicated a lower estimate (less than 50%) than their female counterparts in Europe (60%) and Africa (55%). Males in the Americas also had a far lower median for all ‘no’ responses in comparison to their counterparts in Europe and Africa, indicating males in the Americas had similar views towards aggression as their female counterparts. However, their estimates for the views of the general population were far lower in comparison to their female counterparts, with the majority of respondents indicating approximately 30% of the population also shared similar views. In comparison, the majority of males in Africa estimated approximately 70% of the general population shared similar views towards aggression; that being aggression is an undesirable and incomprehensible phenomenon. Interestingly, the median number of responses for males in Africa was higher than their female counterparts for all ‘no’ answers.

For statement five – “aggression is always negative and unacceptable; feelings should be expressed in another way” – females in the Americas and Europe were the only respondents to answer ‘yes’ to this statement, indicating they agreed with such a statement. The majority of respondents also appeared to make very similar estimates of the general population also agreeing with that statement; that being that approximately 55% of the population also shared the same view. At the other end of the scale i.e., all ‘yes’ responses, males in the Americas were the only group to answer ‘yes; to each statement, thus indicating they viewed aggression as an undesirable and dysfunctional means by which to act. No males in Europe or Africa responded in the same way, nor did females in any region.

Figure 6. Undesirable Responses and Population Estimates by Region and Gender



Part Two: BPAQ-SF

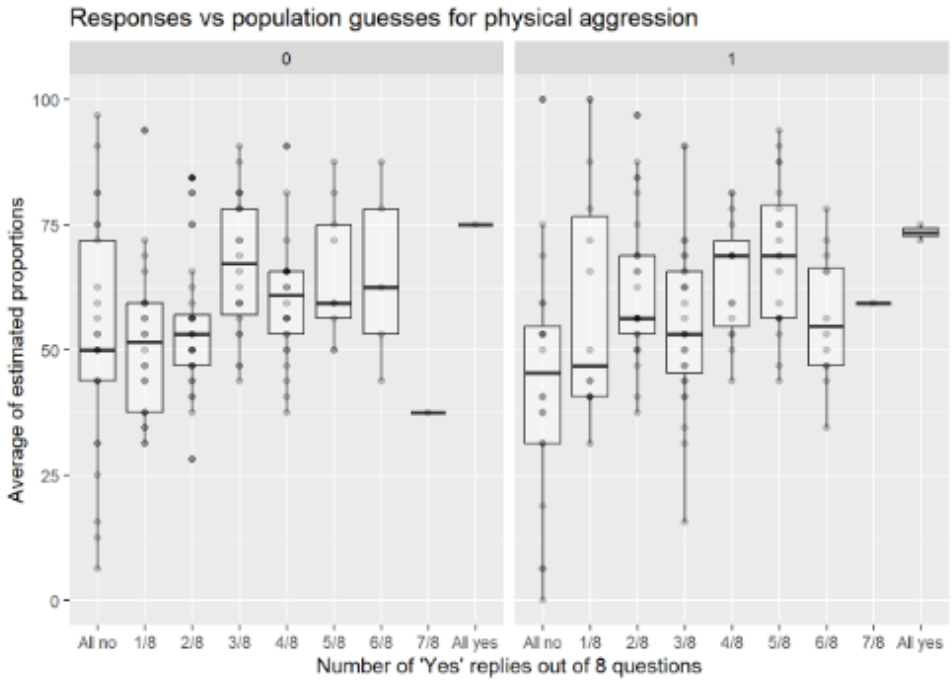
We continued our exploratory analyses of each individual domain as they related to each condition (control vs. experimental), age and gender.

Physical Aggression

For physical aggression, the average number of all ‘yes’ responses by participants in the control group sat at 50%, indicating the majority of participants in this group did not view aggression as a positive attribute, and they believed approximately 50% of the wider

population shared the same view. In comparison, responses made by participants in the experimental condition were slightly lower, with the majority indicating approximately 45% of the wider population shared the same view. Question seven invited participants to indicate whether or not they agree with the statement ‘if I have to resort to violence to protect my rights, I will’. There was no spread of data for this statement in either group, thus indicating only a few participants answered ‘yes’ to this question. For those that did, responses indicated most participants in the control group believed approximately 38% of the wider population shared the same view, in comparison to approximately 60% in the experimental condition. We see similar results for all ‘yes’ responses, in that only a few participants answered ‘yes’ to each statement, thus indicating they viewed aggressive behaviour, specifically physical aggression, as a positive attribute. For the few participants in the control group who responded ‘yes’ to all questions posed, they indicated approximately 75% of the general population would also share the same views. The estimated proportion in the experimental group was only slightly lower at approximately 73%. Results are displayed in Figure 7.

Figure 7. Responses and Population Estimates for Physical Aggression



Physical Aggression by Region

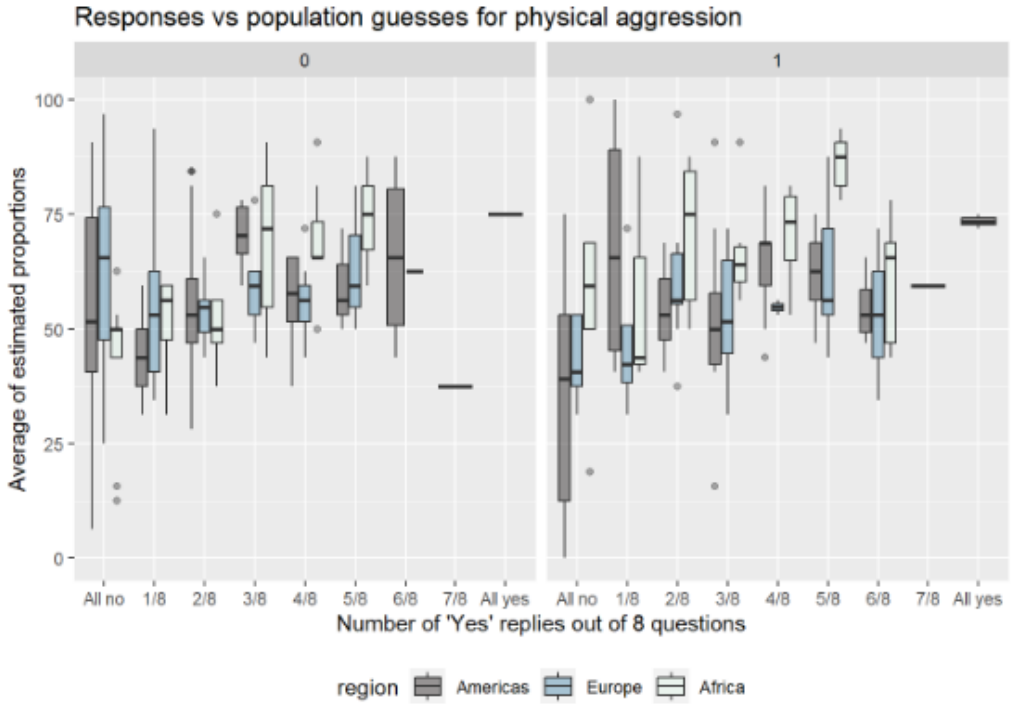
We then analysed the responses and estimated proportions for physical aggression as they pertained to regions in the world (Americas, Europe and Africa). In the control group, participants who were identified as residing within Europe and Africa had higher averages in

comparison to their American counterparts. For example, question one asked participants whether or not they agreed with the statement 'given enough provocation, I may hit another person'. On average, participants in Europe and Africa who answered 'yes' to this statement also indicated 55% and 60% (respectively) of the general population shared the same view. On the other hand, the average response for participants in the Americas was approximately 45% which indicates Americans do not feel as strongly about the use of physical aggression as their European and African counterparts. However, in the experimental group for question one, Americans had a far higher average response when compared to their counterparts, with those that answered 'yes', also indicating approximately 70% of the general population shared the same view.

Other than question one, participants who were identified as residing within Africa had the highest average for the remaining physical aggression questions in comparison to their European and American counterparts. For question five – 'once in a while, I can't control the urge to strike another person' – the highest average across each physical aggression question was seen for this statement by participants in Africa. The majority of participants who were identified as being from Africa, indicated they agreed with this statement, and estimated approximately 88% of the general population also agreed with their views. These findings indicate that Africans have a far more favourable perspective about the use of physical aggression in comparison to participants in the Americas and Europe. Results are displayed in Figure 8.

Figure 8.

Figure 8. Physical Aggression Responses by Region

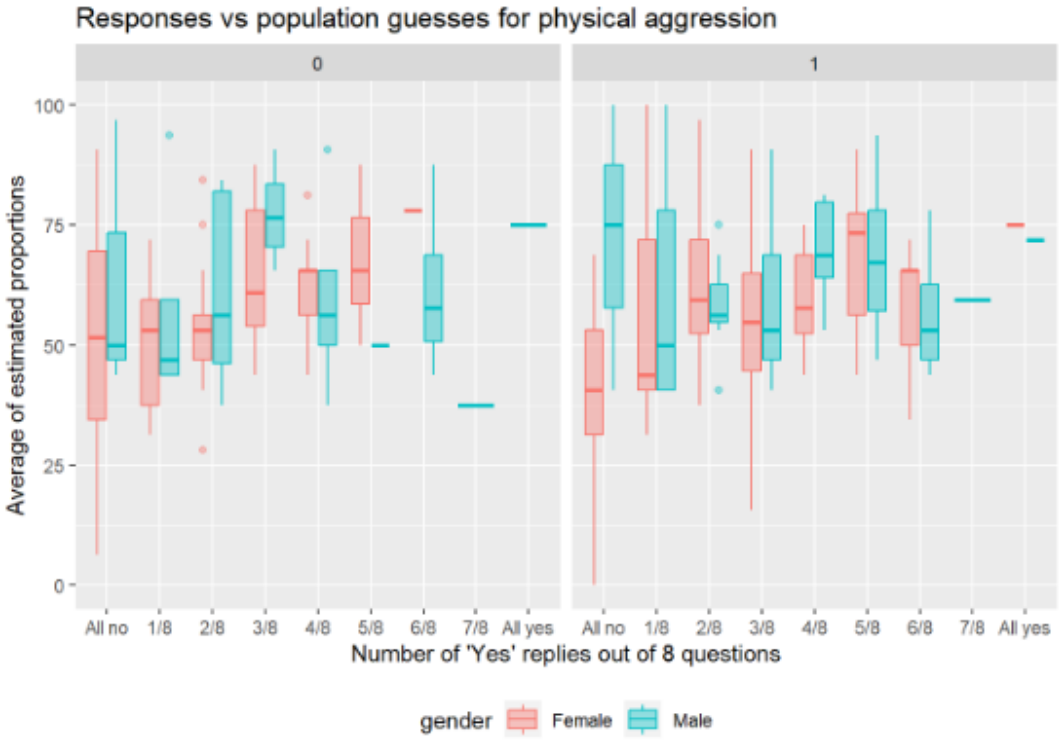


Physical Aggression by Gender

We then compared responses between genders (Figure 9). On average, more females than males in the control group answered ‘no’ to all questions, indicating they did not view physical aggression as a positive behaviour. However, the reverse can be seen in the experimental condition, with more males than females answering ‘no’ to all questions, thus indicating that females were more likely to agree with the use of physical aggression in comparison to their male counterparts. Females did however appear less confident with their estimates for the views of the general population, with estimates of approximately 40% made in the experimental group in comparison to approximately 51% in the control group.

Interestingly, only males in the control group answered ‘yes’ to the statement ‘if I have to resort to violence to protect my rights, I will’ and ‘yes’ to all statements, indicating physical aggression is a necessary act at times. Similar results can be seen in the experimental group, with only male participants answering ‘yes’ to statement seven. However, there was also a small number of female participants who answered ‘yes’ to each statement in the experimental condition, who also estimated a higher proportion (75%) of the general population to share the same view in comparison to their male counterparts’ estimation (73%).

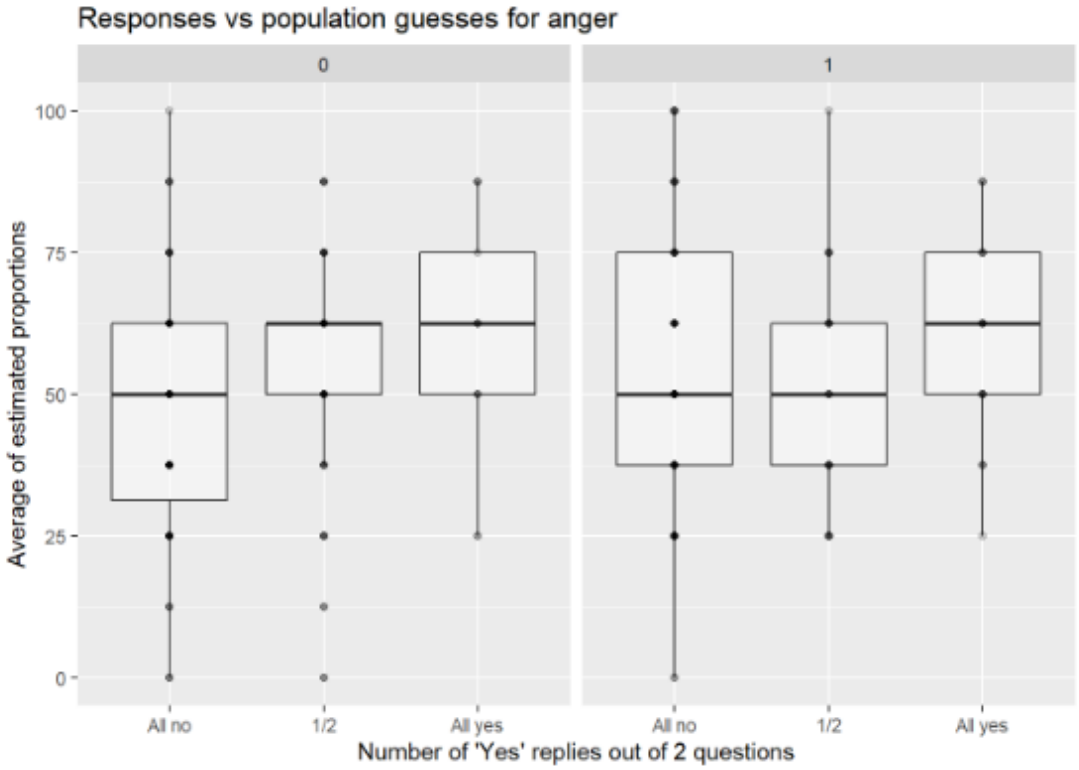
Figure 9. Physical Aggression Responses by Gender



Anger

We then analysed responses pertaining to the 'anger' subscale of the BPAQ-SF (Figure 10). In total, there were two questions which encouraged participants to assess their view of anger as an aggressive trait, as well as that of the general population. The two statements were 'sometimes I fly off the handle for no good reason' and 'I sometimes feel like a powder keg ready to explode'. The average response for participants in the control and experimental condition were the same for 'all no' and 'all yes' responses, with both groups estimating 50% and 65% (respectively) of the general population to share the same view about anger as themselves. We do see a difference however in responses to statement one – 'sometimes I fly off the handle for no good reason'. Participants' responses in the control group indicated a higher average of 65% in comparison to participants' responses in the experimental group, who estimated 50% of the general population to agree with their view. This indicates participants in the control group were more likely to agree with the statement made and were more confident others in the population would also be in agreeance in comparison to their experimental counterparts.

Figure 10. Responses and Population Estimates for Anger



Anger by Gender

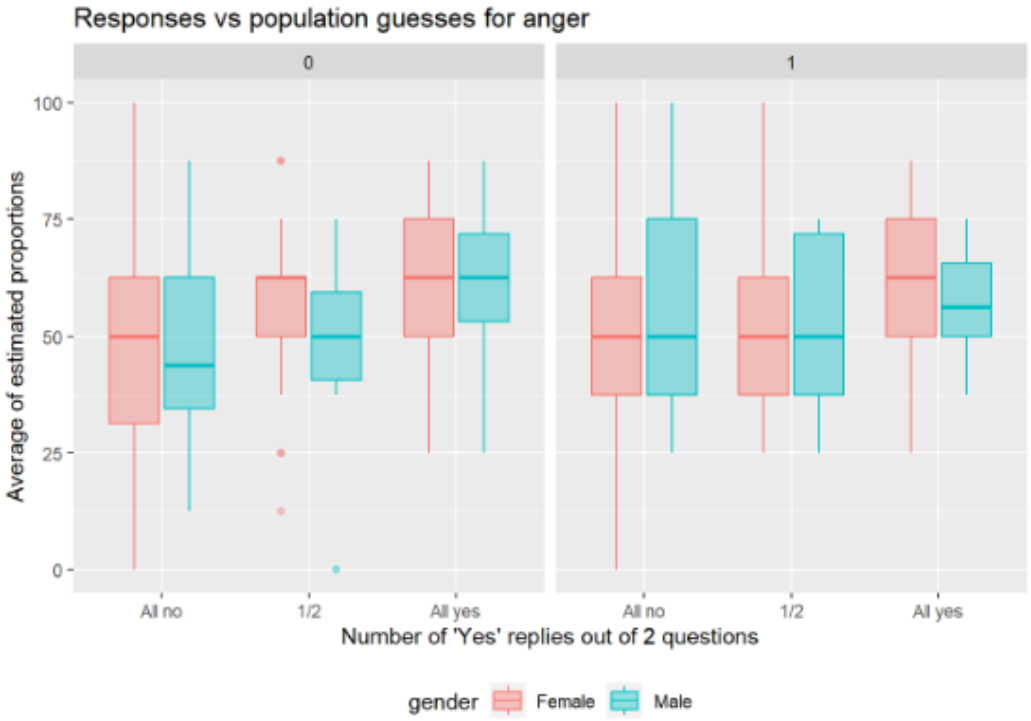
In regards to gender, females in the control group were more likely than males to answer ‘no’ to all statements whereas females and males in the experimental condition responded the same, on average. Females in both groups also responded the same on average, with the majority of female participants in each group estimating 50% of the population to also agree with their perspective.

For ‘all yes’ responses, both males and females in the control group answered the same on average, with an estimated proportion of 65% for agreed perspectives. In comparison with the experimental group, the average response for females was higher than it was for males, indicating females were more likely to agree with the use of anger in certain situations. The average response for males in the experimental condition were lower than their female counterparts, and when also compared to their male counterparts in the control group. This may indicate males were wearier of how their responses could be perceived, given they were made aware a ‘truth-telling algorithm’ would be applied. As such, they may have responded in a more reserved manner for the purpose of maintaining desired perceptions.

For statement one, females in the control group responded higher than males on average and estimated approximately 65% of the wider population to share the same view. In comparison, males in the control group estimated approximately 50% of the population to agree with this statement. There were also a few outliers for females in this group, with some participants estimating approximately 85% of the population to share their same perspective, and the lowest being approximately 15%. The difference between plots for this group further suggests a difference in opinions between genders on this item.

Interestingly, both genders in the experimental condition responded the same on average however the distribution of data for females suggests that while both genders seem to agree with one another on this item, amongst the females themselves is a large difference in views. Results are displayed in Figure 11 (pp. 53).

Figure 11. Anger by Gender

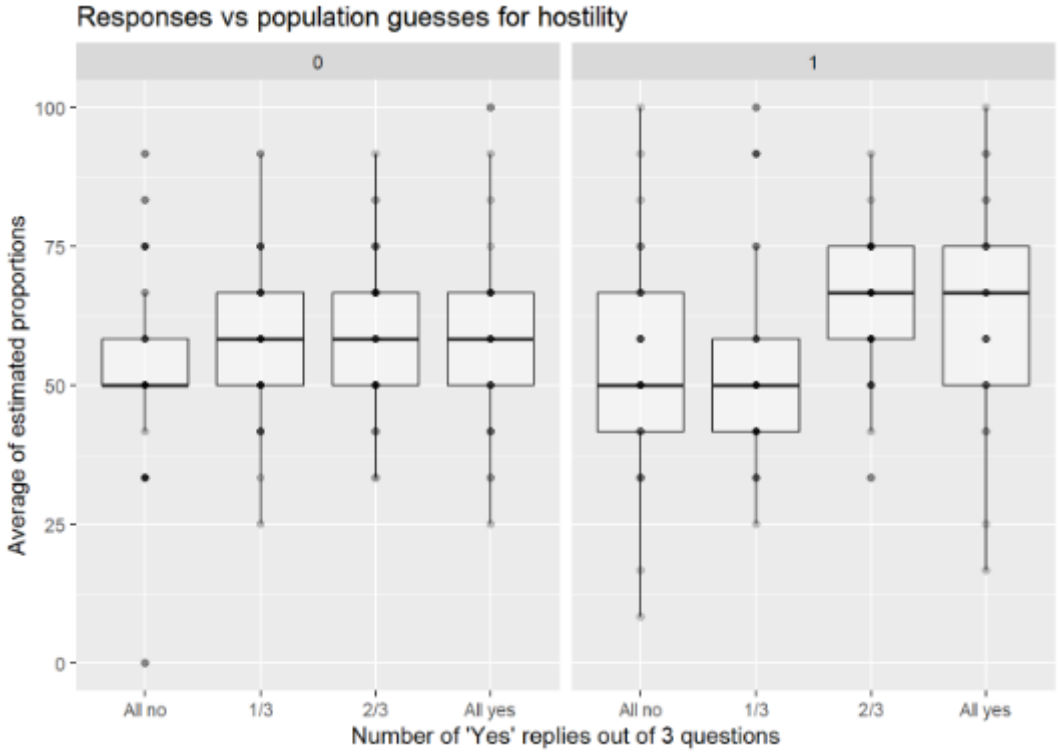


Hostility

Overall, there were three statements pertaining to the measurement of hostile behaviour – ‘at times I feel I have gotten a raw deal out of life’, ‘other people always seem to get the breaks’ and ‘I wonder why sometimes I feel so bitter about things’. In the control group, the average response by participants was the same for each statement, with all three statements resulting in a ‘yes’ and all of whom estimated approximately 60% of the general population would agree with such statements. A lower average can be seen for ‘all no’ responses in this group, with the majority of participants who answered in this manner estimating approximately 50% of the population to share the same view. We do however see a difference in responses when we compare the two groups. For participants in the experimental condition, the average response for ‘all no’ answers and a ‘yes’ to the first statement was the same, with both groups estimating 50% of the population to be in agreeance with their own perception of hostility. Interestingly, participants in the experimental group who answered yes to the second statement or to all questions posed both estimated approximately 65% of the population shared the same view. This is higher than those in the control group with an estimated proportion of 60%. Although the difference is

not significant, it does suggest knowledge of a ‘truth-telling algorithm’ may have had a slight impact on how participants chose to respond. Results are displayed in Figure 12.

Figure 12. Responses and Population Estimates for Hostility



Hostility by Gender

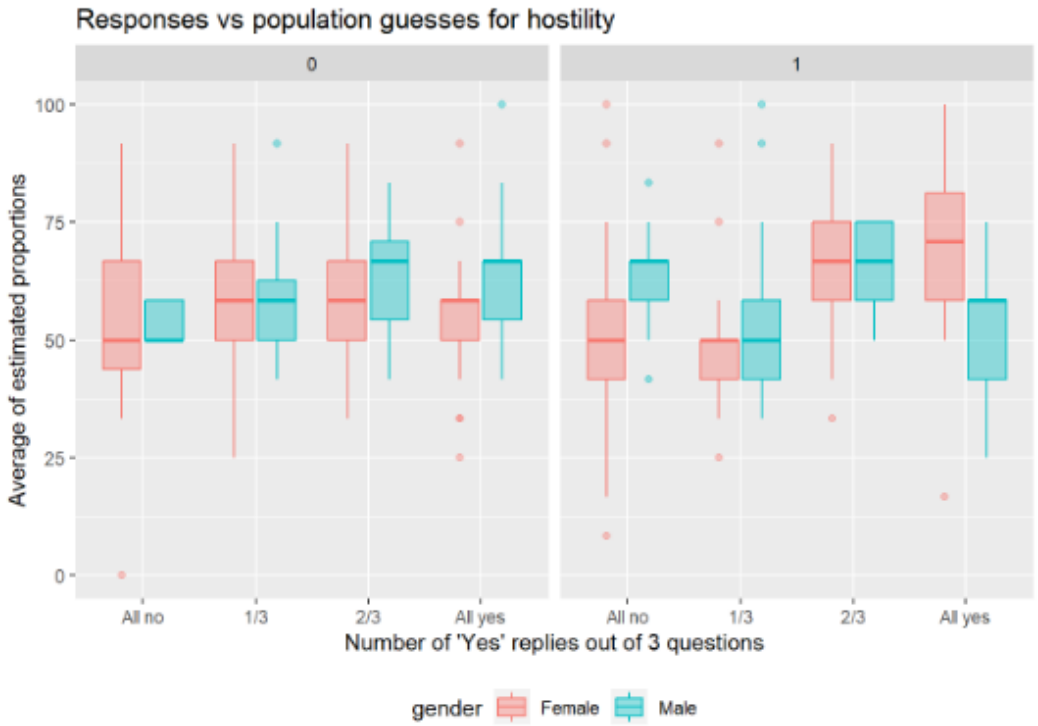
We then analysed the difference in responses between genders. For ‘all no’ responses, both genders respond the same on average in the control group, with both groups also estimating 50% of the population to share the same view towards hostility. The same can be seen for females in the experimental condition, who also estimated 50% of the population to share their view. In comparison, males in the experimental condition attributed higher estimates to the views of the general population than their female counterparts, and more so than both genders in the control group. The size of the plot for male participants in both groups also suggests males tended to agree with one another more than female participants, who tended to hold different opinions towards hostility within their own group.

Males in the control group were more likely to agree with each statement made, thus indicating a favourable opinion of hostility. However, we see the reverse in the experimental condition, with females more likely to agree with each statement in comparison to their male counterparts. Females in the experimental group also had the highest estimate in comparison

to their male and control group counterparts for ‘all yes’ responses, with the average estimate being approximately 70%. In comparison, males in the experimental condition made estimates of approximately 58% whereas females and males in the control group had estimates of 58% and 62%, respectively.

Overall, male participants tended to respond to each statement in the same manner as female participants, as indicated by the same average across groups (and gender) for the majority of questions. However, for some statements we see that male participants attributed higher estimates to the views of the general population in comparison to their female counterparts which suggests that while males may have responded the same as females on some statements, they tended to believe more of the population would share their perspective on hostility. Females also tended to be more varied in their responses which indicates a lower level of agreeance on hostility within their own gender group in comparison to that of males, responses of which indicated a higher level of agreeance on the hostility subscale. This may suggest males in general are more likely to share the same or similar opinions on hostility than females. Such results could be attributed to gender ‘norms’ within society, which often portrays males as more violent and aggressive when compared to females. Results are displayed in Figure 13.

Figure 13. Hostility by Gender



Discussion

The present study builds on previous research in a few ways, albeit unexpectedly. Firstly, we showed that while the application of Bayesian methodologies has proven success in other fields, the influence of such methodology in our own aggression research was limited. This indicates more robust methods may need to be employed when attempting to measure a controversial topic such as aggression, given the additional influence of social desirability bias. Secondly, we investigated differences in aggression across cultures and regions which lends further evidence of support to cross-cultural aggression norms. Finally, our study demonstrated interesting differences between genders which seem to defy societal norms and stereotypical behaviour.

Our study found a non-significant difference between groups, indicating the application of Bayesian methodologies (Bayesian Truth Serum or 'BTS') had little impact on the way in which participants responded to statements about aggression. In effect, our hypothesis was not supported based on the results of this study, which is also inconsistent with previous research. In philosophical research for example, BTS was found to influence how participants responded to questions pertaining to one's moral alignment, virtue, and freedom (Schoenegger, 2023). Utilising Bayesian methods to measure inconsistency in data sets has also demonstrated a difference in responses regarding 'at-risk' behaviour amongst youth. The results corroborated the influence of Bayesian methods on adolescents' self-reported engagement in risky behaviours such as substance use, sexual behaviour, and criminality (Rosenbaum, 2009). Research also suggests that self-reported substance use is under-reported due to stigma and the prospect of negative consequences from the criminal justice system if one were to fully describe their use of alcohol and drugs (McMillan et al., 2009). The use of a Bayesian model demonstrated a larger effect size when compared to a model which does not consider the likelihood of under-reporting. In effect, the application of Bayesian methodology illustrated the effects and prevalence of drug use amongst juvenile offenders (McMillan et al., 2009).

In regards to one's engagement in illegal activities such as driving whilst intoxicated or cheating on an exam, the proportion of people who admitted to engaging in such behaviours was found to be nearly twice as high in the BTS condition when compared to the control condition (Loughran et al., 2014). Participants in the BTS condition also acknowledged a greater willingness to participate in illegal behaviour with lower estimates of perceived risk when compared to their control counterparts. This research provides evidence

of how BTS methodology has been applied to measure people's perspectives and 'honest' views regarding a number of immoral and stigmatised topics. While we may not have found sufficient evidence to support our hypothesis, previous research would provide support for the idea that people's willingness to disclose information on sensitive topics differs significantly when they are encouraged to be exact and deliberate in their responses. It may also be reasonable to assume that individuals are more inclined to minimise their engagement in behaviours which are considered incompatible with societal norms as a means to 'save face' amongst social groups.

The application of Bayesian Truth Serum (BTS) methodologies is also expected to maximise truthful responses by offering a financial incentive. In this study, participants were promised a financial incentive prior to participating, while half the participants were also advised a 'truth-telling algorithm' would be applied to their responses. The financial incentive offered did not appear to influence a significant difference between how each group responded to the proposed statements. In the study completed by Schoenegger (2023), participants in the experimental group who had BTS methodology applied to their responses received a lower pay per hour on average in comparison to the control condition. The difference in average pay did not appear to limit the effects of BTS, with differences in distributions still observed between groups, thus indicating the applicability of BTS on questions pertaining to morality (Schoenegger, 2023). Similarly, offering a financial incentive to respond to a survey about one's criminal convictions was found to improve participants' willingness to engage and the rate at which participants responded (Preisendorfer & Wolder, 2014). However, the overall quality of participants' responses remained consistent, regardless of participants receiving a financial incentive or not.

Interestingly, individuals also appear less willing to respond to surveys regarding polarising topics such as spending habits (Mulder & de Bruijne, 2019). This also seems to be the case for surveys canvassing increasingly sensitive topics such as abortion and drink-driving, with respondents more likely to respond truthfully if perceived risks and losses were mitigated (Rasinski et al., 1999). Disapprovement by an interviewer and being embarrassed were considered the two greatest risks and losses, respectively, perceived to result from the completion of a survey with unaddressed anonymity concerns. Canvassing participants' anonymity concerns prior to answering questions regarding sensitive topics was also found to maximise truthfulness in survey responses (Rasinski et al., 1999). When it comes to predicting participants' privacy-protecting behaviours, both monetary and social rewards have been shown to positively correlate with a participants' willingness to disclose sensitive

and personal information about themselves (Wang et al., 2017). However, social benefits such as improved relationships with significant others, greater inclusivity amongst different social groups and increased emotional support were perceived as more likely to encourage self-disclosure and honest responses than the benefits offered by monetary rewards alone. Offering a financial incentive to participants in our research appeared to contribute to a good response however future research may wish to consider the application of both financial and social incentives to enhance participants' honest self-disclosure on matters regarded as sensitive or private.

Aggression is one topic that is often perceived as a behaviour which is undesirable or negative, as evidenced by the results of our study. In the context of intimate relationships for example, aggression against partners can be perceived as 'normal' or commonplace, dependent on individual standards and personal circumstances (Arriaga et al., 2016). Social norms and past aggression are some factors identified as influencing one's perceptions and, at times, tolerance of aggressive behaviour. Perhaps the nature of our survey topic was considered too 'sensitive', and as such, discouraged certain individuals from participating. It also wasn't possible to obtain full anonymity of participants given the survey was completed online and as a result, participants' IP addresses were made available to the researchers. However, an IP address only links to a geographical location as opposed to a specific address. It could be hypothesised that the combination of only partial anonymity coupled with a sensitive survey topic discouraged some individuals from responding. If this were the case, that may explain why the majority of participants were identified as being from US, African and European populations as they felt the most comfortable to respond to questions regarding aggression. Maintaining confidentiality of participants' responses may however encourage an improved response to surveys, especially on topics regarding controversial and sensitive matters. Offering a survey which promotes anonymity of participants', and their responses is thought to improve the rate at which participants are recruited for online surveys (Van Selm & Jankowski, 2006). Furthermore, utilising research methods which guarantee an individual's full confidentiality has been shown to improve self-reports of intimate partner violence (Moffitt et al., 1997); a relational and physical form of aggression.

Gender

In our study, we observed a number of differences between genders. For the POAS, males tended to view aggression as more desirable than their female counterparts however

men appeared less confident in their assessments of the view of the general population in comparison to females. There has long been debate surrounding gender differences on many topics of contention, including violence and aggression. Within gender research, men have consistently been shown to project and engage in physically aggressive behaviour more so than women, with women noted as being more likely to engage in indirect or displaced aggression (McDermott, 2015). Men and women also appear to hold different motivations for engaging in aggressive behaviour, with men more likely to utilise aggression as a means to exude greater reproductive status and a high-quality reputation. Women on the other hand tend to be incentivised by other competing demands such as childbearing and motherhood as a reason to avoid engaging in physical aggression against others (McDermott, 2015). Women are also more likely to aggress against a female counterpart they perceive to be of lower status in comparison, whereas males tend to behave aggressively towards high-status men who pose a greater threat to their own status (Terrell et al., 2008). Gender differences are also found amongst males and females high in thrill-seeking and narcissistic personality traits. Males high in these factors were more likely to engage in direct physical aggression towards a competitor, whereas aggression is not used as a function upon meeting for females (Terrell et al., 2008). Men who display more physical aggression have also been found to present with lower levels of agreeableness, tend to be less extroverted and exhibit characteristics consistent with neuroticism (Burton et al., 2007). However, there is evidence to suggest physical aggression in women is associated with more depressive behaviours, increased adaptability, and higher conscientiousness (Burton et al., 2007).

Interestingly, when country of origin was taken into consideration, we found that females in America perceived aggression to be more desirable than their male counterparts and were confident others in the general population would share the same perspective (Figure 3). The same result was found amongst participants who were identified as being from Africa, with females more likely to view aggressive behaviour as a desirable phenomenon in comparison to their male counterparts. However, the difference between these groups was minimal, which does suggest males and females in Africa do share some similarities when it comes to aggressive behaviour. This is consistent with previous research which found fewer differences between genders in regards to the measurement of relational aggression (Belgrave et al., 2011). This suggests the androgynous belief system African American individuals use to inform their attitudes, beliefs and behaviours creates a shared perspective on engagement in aggressive behaviour (Belgrave et al., 2011). Both male and female adolescents with poor adjustment scores were found to exhibit less engagement in prosocial activities and increased

participation in displays of overt and relational aggression. A poor learning and physical environment, low social class and inadequate environmental conditions has been found to contribute to anger, bullying, oppositional and defiant behaviours and fighting were found to contribute to displays of verbal and physical aggression against peers in a sample of adolescents in South Africa (Myburgh et al., 2015). This provides support to our own research which found that participants from Africa had the highest average for the majority of statements in the physical aggression subscale. This suggests that individuals in Africa are more likely to endorse the use of physical aggression in comparison to participants in the Americas and Europe, However, the reverse was found in the anger subscale, with participants from Africa *less* likely to agree with anger being used against others.

In our study of the POAS, females were found to be more likely to view aggressive behaviour as undesirable in comparison to males, and this was consistent across each age group canvassed. Female participants also did not respond in a way which indicated any favouritism towards aggressive behaviour, with no female participant answering 'yes' to all statements posed, thus indicating not one aspect of aggressive behaviour was perceived as desirable to women. This finding may support previous research which found that female participants tended to perceive aggressive vignettes as more aggressive than did males (Stewart-Williams, 2002). Further gender differences regarding the perception of aggression were observed, with women found to perceive more acts as aggressive, while males did not share the same perspective. These findings indicate the display of aggressive behaviour is more acceptable between men which then gives rise to a higher incidence of aggression amongst males (Stewart-Williams, 2002).

Within a nursing environment, females were again found to perceive aggression as unacceptable behaviour when compared to their male counterparts (Bilgin et al., 2016). Exposure to aggression in an earlier setting was also found to contribute to differing views on aggressive behaviour, with those who had been exposed to such behaviour in the past found to be less understanding towards aggression, and more likely to perceive the like as unacceptable and abnormal. Overall, however, females viewed aggression as a negative phenomenon more so than their male counterparts, as indicated by their responses on the POAS (Bilgin et al., 2016). The same was found in our research, with females in America and Europe found to be the only participants who indicated their agreeance with the statement 'aggression is always negative and unacceptable; feelings should be expressed in another way' on the POAS. Male participants in America on the other hand were the only participants to disagree with each statement posed, thus indicating an unfavourable perspective towards

aggression. This lies in contrast to popular belief, which posits that men are more likely to engage in aggressive behaviour, are more likely to be the perpetrators of violent sexual offences and are less likely to fall victim to another when compared to females (Padgett & Tremblay, 2020). However, given differences already exist between men and women and the way they perceive aggression, perhaps American men perceive aggression in an entirely different way. In a study of US university students, males were found to associate 'manhood' or their status and masculinity to take action (Weaver at al., 2010). In turn, men were found to engage in more physically aggressive behaviours in comparison to women as a means to establish status and maintain their manhood. Males have also been found to minimise the seriousness of behaviours typically associated with bullying, be endorsing of relational aggression and less likely to exhibit empathetic behaviours for victims when compared to females (Johnson et al., 2013). Males also tend to be more accepting of physically aggressive behaviour, which comes to be perceived as increasingly unacceptable and harmful to others when the perpetrator of such aggression is also male (Basow et al., 2007). However, perceptions of relational or indirect aggression were not found to differ based on gender, with both male and females perceiving victims of relational aggression to share in equal amounts of harm, regardless of their gender (Basow et al., 2007).

In the control group of our experiment for the BPAQ, more females than males perceived aggression as a negative phenomenon which is consistent with previous research. However, the reverse was found in the experimental condition, with females found to be more likely to endorse the use of physical aggression in comparison to their male counterparts. This may suggest that females are more likely to respond truthfully if it is made known that their responses will be analysed further to determine 'truth-telling'. Women have been found to admit to their criminal behaviour more infrequently when compared to men, and especially among an older population (Preisendorfer & Wolter, 2014). Females did however appear less confident in their responses, as indicated by lower estimates for the views of the general population when compared to females in the control group. This may suggest that while females appeared more likely to agree with the use of physical aggression, they were less confident such views would be consistent with the views of those in the general population, thus indicating that such views may be characterised as 'abnormal' in the context of societal norms and stereotypes. Typically, females are found to be more likely to utilise aggressive tactics of an indirect nature and do so on a more frequent basis in comparison to males (Cote et al., 2007). Women have also been found to engage in the use of aggressive tactics at a similar frequency to men, which is based on self-reports and research

within interpersonal relationships (Cross & Campbell, 2011). Societal norms also appear to contribute to differences among genders in relation to aggressive behaviour, with male perpetrators of violence and aggression against a female considered to be more serious and to result in harsher punitive consequences in comparison to female perpetrators of aggression against male victims (Cross & Campbell, 2011). Consequently, depending upon the societal norms which are upheld in a given society, perspectives on aggression and one's engagement in such behaviour is likely to vary across genders.

Males, however, are more likely to engage in physical aggression (Cote et al., 2007). Differences between genders regarding their perspectives on aggression may be partially explained by social, biological, and evolutionary perspectives. For example, exposure to aggressive media during childhood and adolescence is thought to contribute to aggressive behaviour as individuals transition into adulthood. Additionally, the encouragement of gender-specific behaviours depending upon whether one is a male or female is also thought to contribute to differences in perceptions of and engagement in aggressive behaviours (Cote et al., 2007). Such findings provide support for our own results, which indicated that males in the control group of the BPAQ survey were the only participants who agreed with the statement 'if I have to resort to violence to protect my rights, I will', thus indicating a positive view of the use of physical aggression. Similarly, only males in the experimental condition were found to agree with the statement 'other people always seem to get the breaks', which suggests a view that endorses the use of hostility.

Within the measurement of anger as a subscale of the BPAQ, female participants in the experimental condition were found to be more endorsing of the use of anger in certain situations when compared to their male counterparts. This finding appears to be consistent with previous research which has suggested females are more likely to engage in indirect aggressive behaviours, such as anger. In a sample of female and male prisoners, female prisoners were found to score higher on measures of anger, and on several different subscales, which indicated women tend to be angrier in nature (Suter et al., 2002). They also found women tend to display thoughts, feelings, and behaviours consistent with a high level of anger, which is often provoked by perceived experiences of injustice. Women also tend to internalise their anger rather than overtly demonstrate anger towards others in a physical manner as do males (Suter et al., 2002). This is consistent with further research which found females to be higher in expressions of anger in comparison to their male counterparts (Burt, 2014). Women also lacked in their ability to control their anger, however, following the

administration of rehabilitative interventions, scores on measures of anger control increased significantly for women, and more so than men (Burt, 2014).

In regards to hostility, both male and female participants in our study were found to respond in similar manners, thus indicating a shared perspective on hostile behaviour. Males did however appear to be more confident in their responses, with males found to attribute a higher estimate to the views of the general population in comparison to their female counterparts. Males therefore tended to believe more individuals would corroborate their views towards hostility. Limited research exists which details any possible gender differences when it comes to hostility however women with low self-esteem have been found to engage in more hostile behaviour towards other women; a behaviour which is not generally exhibited by women high in self-esteem (Cowan et al., 1998). Women who were found to have an emotional reliance on men, exhibit negative emotions, and who reported feeling dissatisfied with their current lifestyles also had a higher tendency to act hostile towards other women. Women were also found to be more hostile in comparison to men in a sample of individuals pursuing treatment for drug abuse disorders (Robinson et al., 2001). This remained consistent even once gender norms were controlled for. However, gender was not identified as the only factor at play when comparing levels of hostility, depression, or anxiety. While differences existed between genders when it came to the measurement of hostility, such differences were not attributed to gender but rather differences in negative consequences resulting from social and individual circumstances, such as the absence of social supports (Robinson et al., 2001).

Previous research also suggests individuals' experiences of abuse can contribute to a greater level of aggression, which again provides evidence of social, economic, and environmental risk factors in the development of aggression. Individuals who have previously experienced childhood abuse tend to exhibit higher levels of aggression, anger, and hostility in adulthood, which is also compounded by a predisposition to narcissistic traits (Ford et al., 2009). Men in particular, are more likely to endorse the use of physical and emotional violence and abuse against an intimate partner if they have also experienced violence within their family unit as a child (O'Hearn & Margolin, 2000). Such exposure also appears to influence male's perspectives towards aggression and is used as a means to justify and condone the use of aggression within male-to-female relationships. Substance abuse among adolescents also contributes to a higher rate of aggressive behaviour, anger and negative emotionality, especially with use of heroin and morphine (Fauziah et al., 2012). Exposure to physical and emotional abuse at the hands of a parent, sibling, intimate partner or within a

family dynamic has also been found to share a positive relationship with the incidence of sexual aggression in adulthood (King et al., 2019). Observing aggression and violence within domestic relationships also appears to be a contributing factor. However, becoming a victim of sexual abuse during childhood shares the most significant relationship with sexual aggression among adults (King et al., 2019). Each of these studies solidifies the importance of considering the role abuse plays in the development of aggressive tendencies and perspectives.

There does however remain one unaddressed but important factor to consider with the use of BTS methodologies – BTS operates on the premise that predictions can be made regarding how truthful someone is compared to pre-existing criteria. However, in the field of aggression, it is difficult to explicitly determine a pre-existing 'aggressive' criteria as it is not entirely known how aggressive people are. Research suggests that aggression is a complex behaviour which manifests itself differently depending on a number of factors, such as one's educational level, gender, age, life experiences and cultural context. Thus, it is possible to determine the *factors* which contribute to aggressive behaviours, but not so much *how* aggressive people truly are.

Among a population of youth, anger and physical aggression was found to be especially prevalent among young males, more so than their female counterparts, and increasingly more when compared to their older peers (Sharma & Marimuthu, 2014). Substance use, emotional unwellness, isolation, poor grades, and influences from both the family and social environments were all found to contribute to aggression among youth (Sharma & Marimuthu, 2014). Such findings also seem to provide support for the catalyst model of aggression based on the contributory factors outlined as being associated with aggressive behaviour, especially among youth. However, the question remains – *how* do you determine how aggressive people truly are? In the context of sport betting, predictions about the market are found to improve more with a larger sample size, as is the consistency of estimations made about the same market when employing Bayesian truth serum methodologies (Dai et al., 2021). However, the use of BTS in this context demonstrated only a small number of participants needed to know the 'correct' or 'true' answer to a specific question in order to provide a consistent estimator of a particular topic or behaviour. Given it is difficult to obtain a 'true' answer to how aggressive individuals are, obtaining the same consistent estimators or predictions in the context of aggression is therefore difficult, and perhaps, impossible to achieve. The sample size of our own study could be considered small (n = 289) in comparison to the population size of Dai et al's (2021) study (n = 400), which

would suggest that due to the small sample size, predictive power is limited. Dai et al. (2021) also posits that only a few participants need know the 'correct' answer to obtain a consistent estimator under BTS conditions which may explain why our results indicated little difference between the control and BTS groups. If the 'true' answer to *how* aggressive people are cannot be obtained, such findings could be expected.

A detection-type warning was employed in our survey, as indicated by the statement that a 'truth-telling algorithm' would be applied to participants' responses. Such a warning is considered to deter participants from having the motivation to provide dishonest answers (Pace & Borman, 2006). However, the combination of both a detection-type warning and a warning that dishonest responses would result in negative consequences to the participant tend to produce the lowest levels of fictious responses. Similarly, informing participants they would not receive a financial incentive if they provided dishonest responses did not appear to influence how participants perceived a particular topic (Pace & Borman, 2006).

Age

Females across all age groups in our study were found to be more likely to view aggression as undesirable in comparison to their male counterparts. Females in the youngest age group (18-24 years old) were also found more likely to indicate a response consistent with the view of aggression as an undesirable behaviour when compared to all other age groups, and across genders. This finding is interesting given that physical aggression among young partners tends to be the highest in comparison to couples in late adolescence (Moffitt & Caspi, 1999). Both men and women in a young adolescent relationship have also been found to aggress equally towards one another across time, both physically and psychologically (Capaldi et al., 2005). As couples age, physical aggression is also found to decrease, indicating the display of antisocial behaviours such as aggression becomes a phenomenon to detest as an individual ages (Capaldi et al., 2005).

Young women have also been found to exhibit a tendency to perceive acts of aggression as more harmful and as an inability to control oneself (Pugh, 2009). This may support our own research, which found that female participants in the youngest age group estimated almost 50% of the general population to agree with their perspective of aggression as an unpleasant and repulsive behaviour. While this was not as high as their male counterparts, it does provide evidence to support the idea that young women do tend to perceive aggression as increasingly harmful when compared to others, such as middle-aged

females in our study. Perhaps this could be explained by previous research, which suggests that the use of violence amongst young adolescent females was justifiable depending on the context (Cummings & Leschied, 2009). In effect, if the young females in our study also held similar views towards aggression as a repulsive and unpleasant behaviour but gave credence to the use of such aggression in certain situations, such findings in our study would be expected.

In comparison, males in the youngest and middle age categories were found to 'score' higher on statements of aggression as 'an unpleasant and repulsive behaviour', which indicated adolescent and middle-aged males tended to have a more favourable view towards aggression than their female counterparts. This is consistent with previous research which found that young males were more aggressive in comparison to their female counterparts and were also more likely to instigate fear and injury towards their peers (Hilton et al., 2000). However, as in previous research, both females and males were found to self-report an equal engagement in violence and nonphysical aggression. Similarly, amongst a population of adolescents and young adults, males were more likely than females to self-report a higher incidence of aggression during adolescence and early adulthood (Liu & Kaplan, 2004). They were also found to be lower in levels of self-control and attachment but higher in the display of antisocial attitudes when compared to their female counterparts. Increased amounts of stress also appear to influence one's engagement in aggressive behaviour, with those identified as engaging in aggression as an adolescent found to be more likely to engage in aggression as an adult (Liu & Kaplan, 2004). This may lend support to our own findings which showed that younger males have a higher tendency to agree with and/or engage in aggression than their female peers. This was also seen to be the case when middle-aged males' responses were compared to those of their female counterparts. Thus, if the same young male participants completed the same survey in 10 years' time when they reach the age of their middle-aged counterparts, we could expect to see similar responses based on the findings from previous research.

Younger males were also found to hold greatly contrasting views to their eldest peers, with males in the eldest age category (41-60) scoring significantly lower on measures of aggression as an undesirable behaviour. This indicates older males have a more tolerant perspective on aggressive behaviour when compared to their younger counterparts, which may suggest aggression within this population is seen as more acceptable or appropriate. Responses by males in this age group would however indicate a tendency to perceive aggression as an undesirable phenomenon, which is consistent with previous research

indicating engagement in aggressive behaviour dissipates as individuals age (Capaldi et al., 2005). In a sample of older female and male participants (aged between 20 and 55), feelings of anger were more likely to result in general and physical aggression among males rather than females (Chen et al., 2012). Likewise, increased feelings of social humiliation or embarrassment were associated with an increased incidence of relational aggression, but only among males. Females on the other hand had an increased tendency to respond to a specific social situation angrily yet they were less likely to react aggressively in times of social embarrassment (Chen et al., 2012). Thus, differences in emotional resources, ability to regulate internalised emotions and the incidence of social upset appear to be contributing factors to the display of general, physical, and relational aggression among older adults.

Previous research also suggests aggression remains stable over time, particularly for males, however research also indicates a pattern consistent with the desistance of aggressive behaviour in adulthood (Piquero et al., 2012). The same pattern has also been observed among an offender population, who were found to engage in less aggressive acts during adulthood in comparison to their behaviour as a child and young adolescent. Unconsciously, individuals may also perceive aggression in different ways based on their own "time perspective" (Stolarski et al., 2016, pp. 1) which is developed from personal experiences and events which in turn contributes towards one's engagement in aggression (Stolarski et al., 2016).

It has been demonstrated that past, present, and future factors of an individuals' circumstances also have an impact on a number of aggressive behaviours, such as anger, hostility, impulsivity and displays of verbal and physical aggression. A negative perspective of the past, hedonistic states, and a learned helplessness outlook were factors of 'time perspective' associated with higher levels of anger and hostility (Stolarski et al., 2016). Interestingly, hedonism is also identified as a contributing factor to one's engagement in offending-related behaviour. Thus, the personal circumstances of each individual participant may have contributed towards differing perspectives on aggression among genders and ages, as evidenced by our findings. Males in general were found to have a more favourable perspective on aggression in our study when compared to females, which may indicate males manifest and display factors of aggression as outlined above more so than their female counterparts and as such, are more likely to score higher on measures of aggression.

Limitations

A limitation could be due to the fact two different surveys were presented to participants. By displaying two different surveys to participants, each with their own set of specific questions, we were unable to draw comparisons between each participants' response on each survey. Participants in the experimental condition in particular were incentivised to respond truthfully, upon being informed a 'truth-telling algorithm' would be applied to their answers. However, as the second survey (BPAQ) had different statements in comparison to the first survey (POAS), participants in this group may have indeed responded in a truthful manner yet this could not be determined given the difference in statements. As a result, a non-significant difference was found between groups which could be attributed to the fact that two different surveys were displayed to participants. A future study in this area could display the same survey i.e., the BPAQ to participants in order to allow more in-depth comparisons to be made.

Another limitation of this study may be found in our adapted version of Bayesian methodology. Ordinarily, when Bayesian Truth Serum (BTS) methodology is applied, one group of participants are incentivised financially to respond truthfully, whereas the other group is incentivised financially *and* advised a higher 'score' is applied to truthful responses, thus indicating a higher financial payout is possible (Weaver & Prelec, 2013). In our study, all participants had financial incentive to respond to the questions posed, however, for the experimental group, the 'additional' incentive was simply in the form of a statement, which advised participants a 'truth-telling algorithm' would be applied to their responses. Perhaps this statement was insufficient in creating further incentive for one to respond truthfully given that regardless of the response given or the group that participants were randomly assigned to, participants were rewarded the same financially.

On a similar wavelength is the difference in *when* participants received their financial incentive to respond to the survey. As our survey was completed via Prolific, the participants who were recruited were provided with a financial assurance prior to their completion of the survey. This assurance is guaranteed due to Prolific's own ethical payment principles. Following completion of the survey, participants received their financial payout, as previously assured. The point at which participants received their financial incentive for participating, may have impacted on participants' willingness to respond to the survey initially. However, in a web-based survey completed by Bosnjak and Tuten (2003), the provision of a financial incentive pre- or post-survey did not appear to have any significant

influence on a participants' willingness to engage in the survey (Bosnjak & Tuten, 2003). Therefore, it could be postulated that 'promised' incentives prior to the survey and 'actual' incentives granted upon completion of the survey had little to no influence on participants' willingness to engage.

Another limitation of our study was that we expected participants who responded to the survey to be predominantly from the US and UK populations however that was not the case. In our study, the majority of participants were identified as being from the US, Africa and Europe which suggests that perhaps culture plays a part in participants' desire to complete an online survey regarding aggression. When it comes to surveys, younger participants tend to prefer web- and app-based methods for survey distribution; a preference which is less present for older participants (45-54 years old) (Mulder & de Bruijne, 2019). Therefore, the use of an online platform to administer the survey (via Prolific) may have inadvertently limited the participant pool due to preferred survey administration methods. This resulted in a broader, yet predominantly younger sample group, which does suggest the way in which participants were recruited for our survey had an impact on the final sample population. A combination of both web-based and paper-based surveys has been found to yield more positive results in research (Koivula et al.,2019). Providing the opportunity to complete a survey in a participant's preferred method may encourage increased response rates and in turn, a more representative sample from which to draw robust conclusions.

Another limitation could be found in our exclusion criteria - specifically, a lack of consideration for the influence of one's educational level. In our study, education level was not considered as part of the exclusion criteria. However, there appears to be a relationship between levels of educational prowess and participant recruitment, with more educated individuals found to be more likely to engage in research surveys (Mulder et al., 2019). Furthermore, individuals with greater educational achievements were found to be less likely to admit to their engagement in criminal behaviour when compared to those with a lower educational level (Preisendorfer & Wolter, 2014).

While not excluded, the number of participants (24) in the oldest age group (41-60) only made up 9% of the total sample population. As a result, conclusions able to be drawn from this population were limited and should be treated with caution. The general consensus based on previous research is that aggression dissipates as individuals age (Capaldi et al., 2005), therefore we could have expected to see lower 'scores' on measures of aggression, indicating a less favourable view towards aggression from this population of participants, given current trends in aggression research.

Future Directions

Future research may wish to consider achieving full anonymity to encourage more truthful responding in surveys about sensitive and personal topics. It may also be beneficial to utilise popular survey methods i.e., online, or web-based depending on the intended sample group, given that different methods are enticing in different ways to different people.

Further consideration of screening criteria which identifies mental health disorders or illnesses, disabilities or underlying health conditions should also be given. Individuals with learning disabilities have been found to engage in an increased amount of physical aggression which can, at times, result in serious harm and injury to others (Tyrer et al., 2006). Such aggression is more often than not seen amongst young adults and males with intellectual disabilities, and amongst individuals identified as having a psychological disorder such as anxiety or depression (Tyrer et al., 2006). Incidences of aggression and aggressive behaviour do appear to be more common among mentally ill patients, with higher levels of substance misuse, previous criminal convictions for violence and an increased likelihood of being a victim of violence were all reported as contributing to a higher rate of aggression amongst this population (Hodgins et al., 2007). Individuals within this population also demonstrated lower educational levels, and some who were homeless, which provides further support for the role that environmental and social factors appear to play in the development of aggression.

Interestingly, individuals from a UK-based sample were also found to be less aggressive than their US-based counterparts. Mental health disorders such as depression and schizophrenia also tend to invoke different perceptions within the general population, with most people appearing to have seemingly mixed opinions about the cause and preferred treatment methods for such conditions (Angermeyer & Dietrich, 2006). People with mental illness also tend to be socially excluded, are perceived as being more violent and dangerous now more so than they were 70 years ago and are also considered to have an unpredictable nature. Differences in perceptions of mental illness also differ across regions and cultures, with individuals in the US more likely to support the idea that mental health disorders are caused by personal factors whereas amongst a German population, people are less willing to consider individual factors as the cause of mental illness (Angermeyer & Dietrich, 2006). African Americans tend to endorse the view that genetic and environmental factors bear limited significance in the cause of mental illness (Angermeyer & Dietrich, 2006). Consequently, engagement in aggressive behaviour and perspectives on such behaviour may

be different depending upon the circumstances, society, culture, and individual characteristics at play when aggression arises. Aggression may be perceived as acceptable and justifiable more so than, and despite societal or cultural norms.

As indicated earlier, exposure to child, sexual, psychological and substance abuse also has an impact on the incidence and prevalence of aggressive behaviour, which begs the question to be asked – should we be asking about sensitive topics such as abuse and aggression? Although the measurement of abuse was not a focus of this particular study, the influence abuse has on aggressive behaviour renders it an important factor to consider in psychological research. Becker-Blease and Freyd (2006) indicate that a failure to ask about abuse may create a sense of distrust between participants and researchers, which also relays evidence that abuse is an insignificant factor to address in research (Becker-Blease & Freyd, 2006). By considering more uncomfortable yet necessary participant factors, future research may be able to increase benefits to participants so to encourage an improved response rate on surveys. Acknowledging important aspects of a participants' circumstances are likely to do more good than harm, which renders the importance of considering abuse histories.

Future research may also wish to include a larger sample size to improve the predictive power of BTS methodologies in the field of aggression. Maccann et al. (2011) also suggest utilising both self-reports and observer-reports in research as a means to determine truth-telling of respondents against known criteria. People with a tendency to aggress may also to differ in their perspectives of aggression when compared to individuals without the same predilections. Consequently, if an individual is able to justify the use of aggression in certain contexts, and also has aggressive tendencies, their perspective on aggression is likely to influence their responses and how they process 'aggressive' situations (MacCann et al.,2011). Aggression also tends to be a subjective topic; views of which differ depending on a number of factors. Thus, people tend to show support for statements which corroborate the views of the society they prescribe to and may not necessarily be an accurate reflection of their own views. Social desirability bias, cultural norms and *how* individuals are informed about 'truth-telling' detection methods are also noted as being influential on procuring honest responses (MacCann et al., 2011).

Conclusion

In this study, we attempted to determine whether the application of Bayesian Truth Serum (BTS) methodologies encouraged participants to reveal their honest perceptions

regarding aggressive behaviour. Findings indicated that the application of a financial-based incentive such as that employed with the use of BTS did not produce a significant difference between groups regarding the measurement of aggressive behaviour and perceptions. However, the results obtained indicated that males tend to be more aggressive than females and are more likely to endorse the use of physical aggression and violence against others which is consistent with previous research. While the results from our study may not have been significant, the results obtained provide confirmation that aggression is generally regarded as an undesirable behaviour which generates negative consequences for oneself and others.

Future research may wish to include additional screening criteria to limit the influence of educational prowess, psychological instability, and anxiety on measures of aggression. It may also be worthwhile to consider the influence of abuse on perspectives towards aggression, and to canvas one's experience of abuse to foster a comfortable and accepting research environment for participants. Consideration should also be given to the influence of social desirability bias on the truthfulness of participants' responses however rather than attempting to control such bias, it may instead be more beneficial to implement tools which incentivise and motivate individuals to respond truthfully, as in the research completed by Mooney and Daffern (2015). Research also suggests obtaining a larger sample size, utilising self-reports and observer-reports and applying combination-type warning systems may also serve to increase predictive and respondent validity.

Reference List

Abderhalden, C., Needham, I., Friedli, T. K., Poelmans, J., & Dassen, T. (2002). Perception of aggression among psychiatric nurses in Switzerland. *Acta Psychiatrica Scandinavica*, *106*, 110-117.

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, *87*(4), 1115-1153.

Mason, R. L. (2011). *Learning at work: A model of learning & development for younger workers* [Doctoral dissertation, Massey University]. Massey Research Online. http://mro.massey.ac.nz/handle/10179/2862

Allen, J. J., & Anderson, C. A. (2017). Aggression and violence: Definitions and distinctions. *The Wiley handbook of violence and aggression*, 1-14.

Allen, J. J., Anderson, C. A., & Bushman, B. J. (2018). The general aggression model. *Current opinion in psychology*, *19*, 75-80.

Allen, W. D. (2007). The reporting and underreporting of rape. *Southern Economic Journal*, *73*(3), 623-641.

Anderson, C. A., & Dill, K. E. (2000). Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. *Journal of personality and social psychology*, *78*(4), 772.

Anderson, C. A., & Bushman, B. J. (2002). Human aggression. *Annual review of psychology*, *53*(1), 27-51.

Angermeyer, M. C., & Dietrich, S. (2006). Public beliefs about and attitudes towards people with mental illness: a review of population studies. *Acta Psychiatrica Scandinavica*, *113*(3), 163-179.

Anguiano Carrasco, C., Vigil Colet, A., & Ferrando Piera, P. J. (2013). Controlling social desirability may attenuate faking effects: A study with aggression measures. *Psicothema*.

Barchia, K., & Bussey, K. (2011). Individual and collective social cognitive influences on peer aggression: Exploring the contribution of aggression efficacy, moral disengagement, and collective efficacy. *Aggressive behavior*, *37*(2), 107-120.

Arriaga, X. B., Capezza, N. M., & Daly, C. A. (2016). Personal standards for judging aggression by a relationship partner: How much aggression is too much? *Journal of Personality and Social Psychology*, *110*(1), 36-54.

Barchia, K., & Bussey, K. (2011). Predictors of student defenders of peer aggression victims: Empathy and social cognitive factors. *International Journal of Behavioral Development*, *35*(4), 289-297.

Barnes, S. E., Howell, K. H., Thurston, I. B., & Cohen, R. (2017). Children's attitudes toward aggression: Associations with depression, aggression, and perceived maternal/peer responses to anger. *Journal of child and family studies*, *26*, 748-758.

Barrage, L., & Lee, M. S. (2010). A penny for your thoughts: Inducing truth-telling in stated preference elicitation. *Economics letters*, *106*(2), 140-142.

Basow, S. A., Cahill, K. F., Phelan, J. E., Longshore, K., & McGillicuddy-DeLisi, A. (2007). Perceptions of relational and physical aggression among college students: Effects of gender of perpetrator, target, and perceiver. *Psychology of Women Quarterly*, *31*(1), 85-95.

Becker-Blease, K. A., & Freyd, J. J. (2006). Research participants telling the truth about their lives: the ethics of asking and not asking about abuse. *American psychologist*, *61*(3), 218.

Belgrave, F. Z., Nguyen, A. B., Johnson, J. L., & Hood, K. (2011). Who is likely to help and hurt? Profiles of African American adolescents with prosocial and aggressive behavior. *Journal of youth and adolescence*, *40*, 1012-1024.

Bell, K. M., & Naugle, A. E. (2007). Effects of social desirability on students' self-reporting of partner abuse perpetration and victimization. *Violence and victims*, *22*(2), 243-256.

Bernaldo-De-Quirós, M., Piccini, A. T., Gómez, M. M., & Cerdeira, J. C. (2015). Psychological consequences of aggression in pre-hospital emergency care: Cross sectional survey. *International journal of nursing studies*, *52*(1), 260-270.

Bilgin, H., Keser Ozcan, N., Tulek, Z., Kaya, F., Boyacioglu, N. E., Erol, O., ... & Gumus, K. (2016). Student nurses' perceptions of aggression: An exploratory study of defensive styles, aggression experiences, and demographic factors. *Nursing & health sciences*, *18*(2), 216-222.

Bilgin, H., Tulek, Z., & Ozcan, N. (2011). Psychometric properties of the Turkish version of the Perception of Aggression Scale. *Journal of psychiatric and mental health nursing*, *18*(10), 878-883.

Björkqvist, K. (2018). Gender differences in aggression. *Current opinion in psychology*, *19*, 39-42.

Bonta, J., & Andrews, D. A. (2016). *The psychology of criminal conduct*. Taylor & Francis.

Bosnjak, M., & Tuten, T. L. (2003). Prepaid and promised incentives in web surveys: An experiment. *Social Science Computer Review*, *21*(2), 208-217.

Brenner, P. S., & DeLamater, J. (2016). Lies, damned lies, and survey self-reports? Identity as a cause of measurement bias. *Social psychology quarterly*, *79*(4), 333-354.

Bryant, F. B., & Smith, B. D. (2001). Refining the architecture of aggression: A measurement model for the Buss–Perry Aggression Questionnaire. *Journal of Research in Personality*, *35*(2), 138-167.

Buchmann, A., Hohmann, S., Brandeis, D., Banaschewski, T., & Poustka, L. (2014). Aggression in children and adolescents. *Neuroscience of aggression*, 421-442.

Burt, I. (2014). Identifying gender differences in male and female anger among an adolescent population. *The Professional Counselor*, *4*(5), 531.

Burton, L. A., Hafetz, J., & Henninger, D. (2007). Gender differences in relational and physical aggression. *Social Behavior and Personality: an international journal*, *35*(1), 41-50.

Bushman, B. J., & Huesmann, L. R. (2010). Aggression. *Handbook of social psychology*.

Bushman, B. J., Giancola, P. R., Parrott, D. J., & Roth, R. M. (2012). Failure to consider future consequences increases the effects of alcohol on aggression. *Journal of experimental social psychology*, *48*(2), 591-595.

Buss, A. H., & Durkee, A. (1957). An inventory for assessing different kinds of hostility. *Journal of consulting psychology*, *21*(4), 343.

Buss, A. H., & Perry, M. (1992). The aggression questionnaire. *Journal of personality and social psychology*, *63*(3), 452.

Capaldi, D. M., Shortt, J. W., & Kim, H. K. (2005). A life span developmental systems perspective on aggression toward a partner. *Family psychology: The art of the science*, 141-167.

Catalano, S., Smith, E., Snyder, H., & Rand, M. (2009). Female victims of violence. *U.S. Department of Justice Publications and Materials (7)* *https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1006&context=usjusticematls*

Chen, P., Coccaro, E. F., & Jacobsen, K. C. (2012). Hostile attributional bias, negative emotional responding, and aggression in adults: Moderating effects of gender and impulsivity. *Aggressive Behavior*, *38*(1), 47-63.

Christensen, S. S., & Wilson, B. L. (2022). Why nurses do not report patient aggression: A review and appraisal of the literature. *Journal of Nursing Management*, *30*(6), 1759-1767.

Clifford, S., & Jerit, J. (2015). Do attempts to improve respondent attention increase social desirability bias?. *Public Opinion Quarterly*, *79*(3), 790-802.

Côté, S. M., Vaillancourt, T., Barker, E. D., Nagin, D., & Tremblay, R. E. (2007). The joint development of physical and indirect aggression: Predictors of continuity and change during childhood. *Development and psychopathology*, *19*(1), 37-55.

Cowan, G., Neighbors, C., DeLaMoreaux, J., & Behnke, C. (1998). Women's hostility toward women. *Psychology of Women Quarterly*, *22*(2), 267-284.

Craig, I. W., & Halton, K. E. (2009). Genetics of human aggressive behaviour. *Human genetics*, *126*, 101-113.

Cross, C. P., & Campbell, A. (2011). Women's aggression. *Aggression and Violent Behavior*, *16*(5), 390-398.

Crutzen, R., & Göritz, A. S. (2010). Social desirability and self-reported health risk behaviors in web-based research: three longitudinal studies. *BMC public health*, *10*(1), 1-10.

Cummings, A. L., & Leschied, A. W. (2009). Understanding aggression with adolescent girls: Implications for policy and practice. *Canadian Journal of Community Mental Health*, *20*(2), 43-57.

Cunha, O., Peixoto, M., Cruz, A. R., & Gonçalves, R. A. (2022). Buss-Perry Aggression Questionnaire: Factor structure and measurement invariance among Portuguese male perpetrators of intimate partner violence. *Criminal Justice and Behavior*, *49*(3), 451-467.

Dai, M., Jia, Y., & Kou, S. (2021). The wisdom of the crowd and prediction markets. *Journal of Econometrics*, *222*(1), 561-578.

Dam, V. H., Hjordt, L. V., da Cunha-Bang, S., Sestoft, D., Knudsen, G. M., & Stenbæk, D. S. (2021). Trait aggression is associated with five-factor personality traits in males. *Brain and behavior*, *11*(7), e02175.

Danz, D., Vesterlund, L., & Wilson, A. J. (2020). *Belief elicitation: Limiting truth telling with information on incentives* (No. w27327). National Bureau of Economic Research.

De Benedictis, L., Dumais, A., Stafford, M. C., Côté, G., & Lesage, A. (2012). Factor analysis of the French version of the shorter 12-item Perception of Aggression Scale (POAS) and of a new modified version of the Overt Aggression Scale (MOAS). *Journal of psychiatric and mental health nursing*, *19*(10), 875-880.

DeWall, C. N., Anderson, C. A., & Bushman, B. J. (2011). The general aggression model: Theoretical extensions to violence. *Psychology of violence*, *1*(3), 245.

Diamond, P. M., & Magaletta, P. R. (2006). The short-form Buss-Perry Aggression questionnaire (BPAQ-SF) a validation study with federal offenders. *Assessment*, *13*(3), 227-240.

Dodou, D., & de Winter, J. C. (2014). Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior*, *36*, 487-495.

Dowds, E. (2020). Towards a contextual definition of rape: Consent, coercion and constructive force. *The Modern Law Review*, *83*(1), 35-63.

Emmert, A. D., Carlock, A. L., Lizotte, A. J., & Krohn, M. D. (2017). Predicting adult under- and over-reporting of self-reported arrests from discrepancies in adolescent self-reports of arrests: A research note. *Crime & Delinquency*, *63*(4), 412-428.

Fauziah, I., Mohamad, M. S., Chong, S. T., & Abd Manaf, A. (2012). Substance abuse and aggressive behavior among adolescents. *Asian Social Science*, *8*(9), 92.

Ferguson, C. J., Cruz, A. M., Martinez, D., Rueda, S. M., Ferguson, D. E., & Negy, C. (2008). Personality, parental, and media influences on aggressive personality and violent crime in young adults. *Journal of Aggression, Maltreatment & Trauma*, *17*(4), 395-414.

Ferguson, C. J., & Dyck, D. (2012). Paradigm change in aggression research: The time has come to retire the General Aggression Model. *Aggression and violent behavior*, *17*(3), 220-228.

Ferguson, C. J. (2023). An evolutionary model for aggression in youth: Rethinking aggression in terms of the Catalyst Model. *New Ideas in Psychology*, *70*, 101029.

Fernandez, E., Day, A., & Boyle, G. J. (2015). Measures of anger and hostility in adults. In *Measures of personality and social psychological constructs* (pp. 74-100). Academic Press.

Ford, J. D., Fraleigh, L. A., & Connor, D. F. (2009). Child abuse and aggression among seriously emotionally disturbed children. *Journal of Clinical Child & Adolescent Psychology*, *39*(1), 25-34.

Forrest, S., Eatough, V., & Shevlin, M. (2005). Measuring adult indirect aggression: The development and psychometric assessment of the indirect aggression scales. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression*, *31*(1), 84-97.

Frank, M. R., Cebrian, M., Pickard, G., & Rahwan, I. (2017). Validating Bayesian truth serum in large-scale online human experiments. *PloS one*, *12*(5), e0177385.

Friesen, L., & Gangadharan, L. (2013). Designing self-reporting regimes to encourage truth telling: An experimental study. *Journal of Economic Behavior & Organization*, *94*, 90-102.

Freeman, A. J., Schumacher, J. A., & Coffey, S. F. (2015). Social desirability and partner agreement of men's reporting of intimate partner violence in substance abuse treatment settings. *Journal of interpersonal violence*, *30*(4), 565-579.

Gallagher, J. M., & Ashford, J. B. (2016). Buss–Perry Aggression Questionnaire: Testing alternative measurement models with assaultive misdemeanor offenders. *Criminal Justice and Behavior*, *43*(11), 1639-1652.

Gamberini, L., Spagnolli, A.,Corradi, N., Sartori, G., Ghirardi, V., & Jacucci, G. (2014). Combining implicit and explicit techniques to reveal social desirability bias in electricity conservation self-reports. *Energy Efficiency, 7,* 923-935.

Garofalo, C., & Velotti, P. (2017). Negative emotionality and aggression in violent offenders: The moderating role of emotion dysregulation. *Journal of Criminal Justice*, *51*, 9-16.

Georgieva, M. (2016). *Exploring the discrepancy between willingness-to-pay and willingness-to-accept with the Bayesian Truth Serum* [Master's Thesis, Erasmus School of Economics].

Gerevich, J., Bácskai, E., & Czobor, P. Á. L. (2007). The generalizability of the buss–perry aggression questionnaire. *International Journal of Methods in Psychiatric Research*, *16*(3), 124-136.

Gnambs, T., & Kaspar, K. (2017). Socially desirable responding in web-based questionnaires: A meta-analytic review of the candor hypothesis. *Assessment*, *24*(6), 746-762.

Heerwegh, D. (2009). Mode differences between face-to-face and web surveys: an experimental investigation of data quality and social desirability effects. *International Journal of Public Opinion Research*, *21*(1), 111-121.

Hilton, N. Z., Harris, G. T., & Rice, M. E. (2000). The functions of aggression by male teenagers. *Journal of Personality and Social Psychology, 79*(6), 988-994.

Hodgins, S., Cree, A., Alderton, J., & Mak, T. (2007). From conduct disorder to severe mental illness: associations with aggressive behaviour, crime and victimization. *Psychological medicine*, *38*(7), 975-987.

Hornsveld, R. H., Muris, P., Kraaimaat, F. W., & Meesters, C. (2009). Psychometric properties of the aggression questionnaire in Dutch violent forensic psychiatric patients and secondary vocational students. *Assessment, 16*(2), 181-192.

Huesmann, L. R., Dubow, E. F., & Boxer, P. (2009). Continuity of aggression from childhood to early adulthood as a predictor of life outcomes: Implications for the adolescent-limited and life-course-persistent models. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression*, *35*(2), 136-149.

Huesmann, L. R., Eron, L. D., Lefkowitz, M. M., & Walder, L. O. (1984). Stability of aggression over time and generations. *Developmental psychology*, *20*(6), 1120.

Ickes, W., Snyder, M., & Garcia, S. (1997). Personality influences on the choice of situations. In *Handbook of personality psychology* (pp. 165-195). Academic Press.

Jansen, G. J., Middel, B., & Dassen, T. W. (2005). An international comparative study on the reliability and validity of the attitudes towards aggression scale. *International Journal of Nursing Studies*, *42*(4), 467-477.

Jansen, G., Dassen, T., & Moorer, P. (1997). The perception of aggression. *Scandinavian Journal of Caring Sciences, 11*(1), 51-55.

Jiang, L., Probst, T. M., Benson, W., & Byrd, J. (2018). Voices carry: effects of verbal and physical aggression on injuries and accident reporting. *Accident Analysis & Prevention*, *118*, 190-199.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532.

Johnson, C., Heath, M. A., Bailey, B. M., Coyne, S. M., Yamawaki, N., & Eggett, D. L. (2013). Adolescents' perceptions of male involvement in relational aggression: Age and gender differences. *Journal of School Violence, 12*(4), 357-377.

Joinson, A. (1999). Social desirability, anonymity, and internet-based questionnaires. *Behavior Research Methods, Instruments & Computers, 31*(3), 433-438.

Kamble, V., Marn, D., Shah, N., Parekh, A., & Ramchandran, K. (2018). *A truth serum for large-scale evaluations*. Working paper.

Kellermann, A. L., & Mercy, J. A. (1992). Men, women, and murder: Gender-specific differences in rates of fatal violence and victimization. *Journal of Trauma and Acute Care Surgery*, *33*(1), 1-5.

King, A. R., Kuhn, S. K., Strege, C., Russell, T. D., & Kolander, T. (2019). Revisiting the link between childhood sexual abuse and adult sexual aggression. *Child Abuse & Neglect*, *94*, 104022.

Kirk, D. S., & Hardy, M. (2014). The acute and enduring consequences of exposure to violence on youth mental health and aggression. *Justice Quarterly, 31*(3), 539-567.

Koivula, A., Räsänen, P., & Sarpila, O. (2019). Examining social desirability bias in online and offline surveys. In *Human-Computer Interaction. Perspectives on Design: Thematic Area, HCI 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part I 21* (pp. 145-158). Springer International Publishing.

Larson, R. B. (2019). Controlling social desirability bias. *International Journal of Market Research, 61*(5), 534-547.

Lépine, A., Treibich, C., & d'Exelle, B. (2020). Nothing but the truth: Consistency and efficiency of the list experiment method for the measurement of sensitive health behaviours. *Social Science & Medicine*, *266*, 113326.

Liu, R. X., & Kaplan, H. B. (2004). Role stress and aggression among young adults: The moderating influences of gender and adolescent aggression. *Social Psychology Quarterly*, *67*(1), 88-102.

Liu, M., & Wronski, L. (2018). Examining completion rates in web surveys via over 25,000 real-world surveys. *Social Science Computer Review*, *36*(1), 116-124.

Lochman, J. E., Barry, T., Powell, N., & Young, L. (2010). Anger and aggression. *Practitioner's guide to empirically based measures of social skills*, 155-166.

Longino, H. E. (2001). What do we measure when we measure aggression?. *Studies in History and Philosophy of Science Part A*, *32*(4), 685-704.

Loughran, T. A., Paternoster, R., & Thomas, K. J. (2014). Incentivizing responses to self-report questions in perceptual deterrence studies: An investigation of the validity of deterrence theory using Bayesian truth serum. *Journal of Quantitative Criminology*, *30*, 677-707.

MacCann, C., Ziegler, M., & Roberts, R. D. (2011). Faking in personality assessment: Reflections and recommendations. *New perspectives on faking in personality assessment*, 309-329.

Mayeux, L., & Cillessen, A. H. (2008). It's not just being popular, it's knowing it, too: The role of self-perceptions of status in the associations between peer status and aggression. *Social Development*, *17*(4), 871-888.

McDermott, R. (2015). Sex and death: Gender differences in aggression and motivations for violence. *International Organization*, *69*(3), 753-775.

McMillan, G. P., Bedrick, E., & C'deBaca, J. (2009). A Bayesian model for estimating the effects of drug use when drug use may be under-reported. *Addiction*, *104*, 1820-1826.

Mills, J. F., Loza, W., & Kroner, D. G. (2003). Predictive validity despite social desirability: Evidence for the robustness of self-report among offenders. *Criminal Behaviour and Mental Health*, *13*(2), 140-150.

Mills, J. F., & Kroner, D. G. (2005). An Investigation Into the Relationship Between Socially Desirable Responding and Offender Self-Report. *Psychological Services*, *2*(1), 70.

Mitrofan, O., Paul, M., Weich, S., & Spencer, N. (2014). Aggression in children with behavioural/emotional difficulties: seeing aggression on television and video games. *BMC psychiatry*, *14*, 1-10.

Moffitt, T. E., & Caspi, A. (1999). Findings about partner violence from the Dunedin Multidiscplinary Health and Development Study (NCJ 170018). Washington, DC: Department of Justice, Office of Justice Programs, National Institute of Justice.

Moffitt, T. E., Caspi, A., Krueger, R. F., Magdol, L., Margolin, G., Silva, P. A., & Sydney, R. (1997). Do partners agree about abuse in their relationship?: A psychometric evaluation of interpartner agreement. *Psychological Assessment, 9*(1), 47–56. https://doi.org/10.1037/1040-3590.9.1.47

Mooney, J. L., & Daffern, M. (2015). The relationship between aggressive behaviour in prison and violent offending following release. *Psychology, Crime & Law*, *21*(4), 314-329.

Moreno, J. K., Fuhriman, A., & Selby, M. J. (1993). Measurement of Hostility, Anger, and Depression in Depressed and Nondepressed Subjects. *Journal of Personality Assessment, 61*(3), 511-523.

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs*. [R package]. Retrieved from https://cran.r-project.org/package=BayesFactor.

Mulder, J., & de Bruijne, M. (2019). Willingness of online respondents to participate in alternative modes of data collection. *Survey Practice, 12*(1), 1-11.

Muñoz-Rivas, M. J., Graña, J. L., O'Leary, K. D., & González, M. P. (2007). Aggression in adolescent dating relationships: Prevalence, justification, and health consequences. *Journal of Adolescent Health*, *40*(4), 298-304.

Murray, A. L., Eisner, M., Ribeaud, D., & Booth, T. (2022). Validation of a brief measure of aggression for ecological momentary assessment research: The Aggression-ES-A. *Assessment, 29*(2), 296-308.

Myburgh, C., Poggenpoel, M., & Nhlapo, L. (2015). Patterns of a culture of aggression amongst Grade 10 learners in a secondary school in the Sedibeng District, South Africa. *curationis*, *38*(1), 1-8.

Nagin, D. S., & Pogarsky, G. (2003). An experimental investigation of deterrence: Cheating, self-serving bias, and impulsivity. *Criminology*, *41*(1), 167-194.

Needham, I., Abderhalden, C., Dassen, T., Haug, H. J., & Fischer, J. E. (2004). The perception of aggression by nurses: psychometric scale testing and derivation of a short instrument. *Journal of psychiatric and mental health nursing*, *11*(1), 36-42.

O'Connor, D. B., Archer, J., & Wu, F. W. (2001). Measuring aggression: Self-reports, partner reports, and responses to provoking scenarios. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression*, *27*(2), 79-101.

Offerman, T., Sonnemans, J., Van de Kuilen, G., & Wakker, P. P. (2009). A truth serum for non-bayesians: Correcting proper scoring rules for risk attitudes. *The Review of Economic Studies*, *76*(4), 1461-1489.

O'Hearn, H. G., & Margolin, G. (2000). Men's attitudes condoning marital aggression: A moderator between family of origin abuse and aggression against female partners. *Cognitive Therapy and Research*, *24*, 159-174.

Pace, V. L., & Borman, W. C. (2006). The use of warnings to discourage faking on noncognitive inventories. *A closer examination of applicant faking behavior*, 283-304.

Padgett, J. K., & Tremblay, P. F. (2020). Gender differences in aggression. *The Wiley Encyclopedia of Personality and Individual Differences: Personality Processes and Individual Differences*, 173-177.

Palmstierna, T., & Barredal, E. (2006). Evaluation of the Perception of Aggression Scale (POAS) in Swedish nurses. *Nordic journal of psychiatry*, *60*(6), 447-451.

Pechorro, P., Barroso, R., Poiares, C., Oliveira, J. P., & Torrealday, O. (2016). Validation of the Buss–Perry aggression questionnaire-short form among Portuguese juvenile delinquents. *International journal of law and psychiatry*, *44*, 75-80.

Perlman, C. M., & Hirdes, J. P. (2008). The aggressive behavior scale: a new scale to measure aggression based on the minimum data set. *Journal of the American Geriatrics Society*, *56*(12), 2298-2303.

Piquero, A. R., Carriaga, M. L., Diamond, B., Kazemian, L., & Farrington, D. P. (2012). Stability in aggression revisited. *Aggression and Violent Behavior*, *17*, 365-372.

Poltavski, D., Van Eck, R., Winger, A. T., & Honts, C. (2018). Using a polygraph system for evaluation of the social desirability response bias in self-report measures of aggression. *Applied psychophysiology and biofeedback*, *43*, 309-318.

Preisendörfer, P., & Wolter, F. (2014). Who is telling the truth? A validation study on determinants of response behavior in surveys. *Public Opinion Quarterly*, *78*(1), 126-146.

Pronin, E. (2007). Perception and misperception of bias in human judgment. *Trends in cognitive sciences*, *11*(1), 37-43.

Pugh, S. (2009). *A phenomenological study of aggression and young adult females* [Doctoral dissertation]. Texas Woman's University.

R Core Team (2020). *R: A Language and environment for statistical computing*. (Version 4.0) [Computer software]. Retrieved from https://cran.r-project.org. (R packages retrieved from MRAN snapshot 2020-08-24).

Rasinski, K. A., Willis, G. B., Baldwin, A. K., Yeh, W., & Lee, L. (1999). Methods of data collection, perceptions of risks and losses, and motivation to give truthful answers to sensitive survey questions. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *13*(5), 465-484.

Ravyts, S. G., Perez, E., Donovan, E. K., Soto, P., & Dzierzewski, J. M. (2021). Measurement of aggression in older adults. *Aggression and violent behavior*, *57*, 101484.

Riggs, D. S., Murphy, C. M., & O'Leary, K. D. (1989). Intentional falsification in reports of interpartner aggression. *Journal of Interpersonal Violence*, *4*(2), 220-232.

Robinson, E. A., Brower, K. J., & Gomberg, E. S. (2001). Explaining unexpected gender differences in hostility among persons seeking treatment for substance use disorders. *Journal of studies on alcohol*, *62*(5), 667-674.

Rosenbaum, J. (2010). Bayesian methods for measures of agreement. *Journal of the Royal Statistical Society Series A: A Statistics in Society*, *173*(1), 270.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225-237.

Săbăreanu, L. M., Gonța, V., & Oprea, C. E. (n.d). Factor Structure Of The Aggression Questionnaire: Study On The Romanian Delinquent Population. *European Proceedings of Educational Sciences*. https://www.europeanproceedings.com/article/10.15405/epes.23045.99

Schoenegger, P. (2023). Experimental philosophy and the incentivisation challenge: a proposed application of the Bayesian Truth Serum. *Review of Philosophy and Psychology*, *14*(1), 295-320.

Severance, L., Bui-Wrzosinska, L., Gelfand, M. J., Lyons, S., Nowak, A., Borkowski, W., ... & Yamaguchi, S. (2013). The psychological structure of aggression across cultures. *Journal of Organizational Behavior*, *34*(6), 835-865.

Sharma, M. K., & Marimuthu, P. (2014). Prevalence and psychosocial factors of aggression among youth. *Indian journal of psychological medicine*, *36*(1), 48-53.

Sherman, R. A., Rauthmann, J. F., Brown, N. A., Serfass, D. G., & Jones, A. B. (2015). The independent effects of personality and situations on real-time expressions of behavior and emotion. *Journal of personality and social psychology*, *109*(5), 872.

Shorey, R. C., Temple, J. R., Febres, J., Brasfield, H., Sherman, A. E., & Stuart, G. L. (2012). The consequences of perpetrating psychological aggression in dating relationships: A descriptive investigation. *Journal of interpersonal violence*, *27*(15), 2980-2998.

Simunović, V., & Žeželj, I. (2023). Managing Self-Presentation: How Social Cues Shape Different Forms of Socially Desirable Responding. *Studia Psychologica*, *65*(2), 103-119.

Smith, S. D., Lynch, R. J., Stephens, H. F., & Kistner, J. A. (2015). Self-perceptions and their prediction of aggression in male juvenile offenders. *Child Psychiatry & Human Development*, *46*, 609-621.

Smits, D. J., & Kuppens, P. (2005). The relations between anger, coping with anger, and aggression, and the BIS/BAS system. *Personality and Individual differences*, *39*(4), 783-793.

Stangor, C., Tarry, H., & Jhangiani, R. (2014). The biological and emotional causes of aggression. *Principles of Social Psychology-1st International Edition*.

Steakley-Freeman, D. M., Lee, R. J., McCloskey, M. S., & Coccaro, E. F. (2018). Social desirability, deceptive reporting, and awareness of problematic aggression in intermittent explosive disorder compared with non-aggressive healthy and psychiatric controls. *Psychiatry research*, *270*, 20-25.

Stewart-Williams, S. (2002). Gender, the perception of aggression, and the overestimation of gender bias. *Sex Roles*, *46*, 177-189.

Stolarski, M., Zajenkowski, M., & Zajenkowska, A. (2016). Aggressive? From time to time… uncovering the complex associations between time perspectives and aggression. *Current Psychology*, *35*, 506-515.

Surette, R. (2013). Cause or catalyst: The interaction of real world and media crime models. *American Journal of Criminal Justice*, *38*, 392-409.

Suter, J. M., Byrne, M. K., Byrne, S., Howells, K., & Day, A. (2002). Anger in prisoners: women are different from men. *Personality and Individual differences*, *32*(6), 1087-1100.

Terrell, H. K., Hill, E. D., & Nagoshi, C. T. (2008). Gender differences in aggression: The role of status and personality in competitive interactions. *Sex Roles*, *59*, 814-826.

Thomas, S. P. (2005). Women's anger, aggression, and violence. *Health care for women international*, *26*(6), 504-522.

Trautmann, S. T., & van de Kuilen, G. (2015). Belief elicitation: A horse race among truth serums. *The Economic Journal*, *125*(589), 2116-2135.

Turner, C. F., Ku, L., Rogers, S. M., Lindberg, L. D., Pleck, J. H., & Sonenstein, F. L. (1998). Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology. *Science*, *280*(5365), 867-873.

Tyrer, F., McGrother, C. W., Thorp, C. F., Donaldson, M., Bhaumik, S., Watson, J. M., & Hollin, C. (2006). Physical aggression towards others in adults with learning disabilities:

prevalence and associated factors. *Journal of intellectual disability research*, *50*(4), 295-304.

Vaillancourt, T., Brendgen, M., Boivin, M., & Tremblay, R. E. (2003). A longitudinal confirmatory factor analysis of indirect and physical aggression: Evidence of two factors over time?. *Child development*, *74*(6), 1628-1638.

van de Schoot, R., Winter, S. D., Griffioen, E., Grimmelikhuijsen, S., Arts, I., Veen, D., ... & Tummers, L. G. (2021). The use of questionable research practices to survive in academia examined with expert elicitation, prior-data conflicts, Bayes factors for replication effects, and the Bayes truth serum. *Frontiers in Psychology*, *12*, 621547.Vazire, S., & Gosling, S. D. (2004). e-Perceptions: personality impressions based on personal websites. *Journal of personality and social psychology*, *87*(1), 123.

Van Selm, M., & Jankowski, N. W. (2006). Conducting online surveys. *Quality and quantity*, *40*, 435-456.

Vigil-Colet, A., Ruiz-Pamies, M., Anguiano-Carrasco, C., & Lorenzo-Seva, U. (2012). The impact of social desirability on psychometric measures of aggression. *Psicothema*, *24*(2), 310-315.

Wang, L., Yan, J., Lin, J., & Cui, W. (2017). Let the users tell the truth: Self-disclosure intention and self-disclosure honesty in mobile social networking. *International Journal of Information Management*, *37*(1), 1428-1440.

Warburton, W. A., & Anderson, C. A. (2015). Aggression, social psychology of. *International encyclopedia of the social & behavioral sciences*, *1*, 373-380.

Watson, M., Fischer, K., Andreas, J., & Smith, K. (2004). Pathways to aggression in children and adolescents. *Harvard Educational Review*, *74*(4), 404-430.

Weaver, R., & Prelec, D. (2013). Creating truth-telling incentives with the Bayesian truth serum. *Journal of Marketing Research*, *50*(3), 289-302.

Weaver, J. R., Vandello, J. A., Bosson, J. K., & Burnaford, R. M. (2010). The proof is in the punch: Gender differences in perceptions of action and aggression as components of manhood. *Sex Roles*, *62*, 241-251.

Webster, G. D., DeWall, C. N., Pond Jr, R. S., Deckman, T., Jonason, P. K., Le, B. M., ... & Bator, R. J. (2015). The brief aggression questionnaire: Structure, validity, reliability, and generalizability. *Journal of Personality Assessment*, *97*(6), 638-649.

Weiss, B., Dodge, K. A., Bates, J. E., & Pettit, G. S. (1992). Some consequences of early harsh discipline: Child aggression and a maladaptive social information processing style. *Child development*, *63*(6), 1321-1335.

Wildeman, C. (2010). Paternal incarceration and children's physically aggressive behaviors: Evidence from the Fragile Families and Child Wellbeing Study. *Social Forces*, *89*(1), 285-309.

Wong, W. K., & Chien, W. T. (2017). Testing psychometric properties of a Chinese version of perception of aggression scale. *Asian journal of psychiatry*, *25*, 213-217.

Wyckoff, J. P. (2016). Aggression and emotion: Anger, not general negative affect, predicts desire to aggress. *Personality and Individual Differences*, *101*, 220-226.

Zimonyi, S., Kasos, K., Halmai, Z., Csirmaz, L., Stadler, H., Rózsa, S., ... & Kotyuk, E. (2021). Hungarian validation of the Buss–Perry Aggression Questionnaire—Is the short form more adequate?. *Brain and behavior*, *11*(5), e02043.

Appendix A

POAS Statements for Survey One

Subscale: Aggression as functional/comprehensible phenomenon

1. Aggression is the start of a positive relationship
2. Aggression is a healthy reaction to feelings of anger
3. Aggression is an opportunity to get a better understanding of the person's situation
4. Aggression is a form of communication and as such not destructive
5. Aggression is a way to protect yourself
6. Aggression is the protection of one's own territory

Subscale: Aggression as dysfunctional/undesirable phenomenon

7. Aggression is an unpleasant and repulsive behaviour
8. Aggression is unnecessary and unacceptable
9. Aggression is hurting others mentally and physically
10. Aggression is an actual action of physical violence of a patient against a nurse
11. Aggression is always negative and unacceptable; feelings should be expressed in another way
12. Aggression is a disturbing intrusion to dominate others

Appendix B

BPAQ-SF Statements for Survey Two

1. Given enough provocation, I may hit another person (PA)
2. I often find myself disagreeing with people (V)
3. At times I feel I have gotten a raw deal out of life (H)
4. There are people who have pushed me so far that we have come to blows (PA)
5. I can't help getting into arguments when people disagree with me (V)
6. Sometimes I fly off the handle for no good reason (A)
7. Other people always seem to get the breaks (H)
8. I have threatened people I know v
9. My friends say that I'm somewhat argumentative (V)
10. I have trouble controlling my temper (PA)
11. I wonder why sometimes I feel so bitter about things (H)
12. I sometimes feel like a powder keg ready to explode (A)
13. Once in a while I can't control the urge to strike another person (PA)
14. If somebody hits me, I hit back (PA)
15. I get into fights a little more than the average person (PA)
16. If I have to resort to violence to protect my rights, I will (PA)
17. I have become so mad I have broken things (PA)

Subscale codes:

PA – Physical Aggression

V – Verbal Aggression

A – Anger

H – Hostility