MERGING STRUCTURAL AND PROCESS-RELATED APPROACHES TO THE STUDY OF AGREEABLENESS: A PREREGISTERED REPLICATION AND EXTENSION

by

Colin E. Vize

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Psychological Sciences
West Lafayette, Indiana
August 2019

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Donald R. Lynam, Chair
Department of Psychological Sciences
Dr. Christopher I. Eckhardt
Department of Psychological Sciences
Dr. Sean P. Lane
Department of Psychological Sciences
Dr. Douglas B. Samuel
Department of Psychological Sciences

Approved by:
Dr. David Rollock
Head of the Graduate Program

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Author: Vize, Colin E. PhD
Institution: Purdue University
Degree Received: August 2019
Title: Merging Structural and Process-Related Approaches to the Study of Agreeableness: A Preregistered Replication and Extension
Committee Chair: Donald R. Lynam

Agreeableness is one of the major domains included within prominent hierarchical models of personality like the Five-factor Model (FFM). (Low) agreeableness has shown to be the strongest correlate of a variety of antisocial behaviors relative to the other FFM domains. Though there is substantial evidence that (low) agreeableness is the most important personality correlate of various antisocial behaviors, this evidence is descriptive and provides little information on the direction or processes underlying the relation. Process-related research has started to provide more insight into how agreeableness-related traits give rise to various antisocial and prosocial behaviors. The proposed study looks to first replicate previous research on some of the potential cognitive/emotional processes related to agreeableness, and then to conduct exploratory analyses to identify which, if any, of the empirically identified facets of agreeableness bear specific relations to the processes under study. Thus, the proposed project seeks to merge developments across important domains of personality research, structural research and process-based research, while also making use of open-science practices.

INTRODUCTION

Hierarchical models of personality have been essential in outlining the structural characteristics of personality traits. A substantial literature has developed around prominent hierarchical models such as the Big Five (Goldberg, 1993) and the Five-Factor Model (FFM; McCrae & John, 1992), both of which posit that personality traits are organized hierarchically. The most often studied level of the hierarchy is represented by the five broad domains of neuroticism, extraversion, openness, agreeableness, and conscientiousness. (Low) agreeableness, referred to as antagonism within clinical nosologies (American Psychiatric Association [APA], 2013), has emerged as a reliable correlate of externalizing behaviors.

Meta-analytic evidence has shown that (low) agreeableness is the most robust correlate of both antisocial and aggressive behavior relative to the other FFM domains (Jones, Miller, & Lynam, 2011; Vize, Collison, Miller, & Lynam, 2019). Within broader models of aggressive behavior, traits related to (low) agreeableness constitute important risk factors in regard to whether or not individuals may act aggressively in certain contexts (e.g., Finkel, 2007; Norlander & Eckhardt, 2005). (Low) agreeableness is also well-represented in many externalizing-related disorders including antisocial personality disorder (APD) and psychopathy (Miller et al., 2001), narcissistic personality disorder (NPD; Miller et al., 2016), and borderline personality disorder (BPD; APA, 2013). Although (low) agreeableness/antagonism consistently emerges as a correlate of antisocial behavior, this work is merely descriptive, and much of it relies on self-reported behavior. Research is needed that examines the proximal emotional and cognitive processes that can clarify *how* antagonism may increase the likelihood of aggressive behavior. Presumably, affective and cognitive processes influence these behaviors and occur further “upstream” than the behaviors themselves.

The current study will incorporate recent work that has empirically derived a hierarchical framework of agreeableness-related traits beyond the domain itself (Crowe et al., 2018) to further expand research that has focused on process accounts of agreeableness. Much of the process-based literature has focused on assessing self-reported agreeableness at the domain level (e.g., Graziano & Tobin, 2013; Wilkowski & Robinson, 2007; Wilkowski & Robinson, 2010). Specifically, we have two primary research aims: 1) a confirmatory aim in which we seek to complete a pre-registered replication of two earlier studies that have used different paradigms to investigate cognitive/emotional manifestations of agreeableness and 2) an exploratory aim to expand previous work by incorporating self-report assessment approaches that break the domain of agreeableness into its constituent parts to explore whether process-related tasks show divergent relations with empirically derived facets of agreeableness. In sum, the present study aims to incorporate calls to increase open science practices in psychological research (e.g., Nosek & Lakens, 2014) as well as recent work calling for the integration of different domains of personality research (i.e., structural and process related personality research; Baumert et al., 2017).

Theoretical Account of Agreeableness

Although agreeableness’ empirical history is shorter relative to other FFM domains like extraversion and neuroticism, agreeableness-related traits have theoretical links to behaviors that have figured prominently in human evolutionary history, such as altruism and social cooperation (Brown & Brown, 2006; Riolo, Cohen, & Axelrod, 2001;

Axelrod, 1984). Indeed, the consistent cross-cultural emergence of an agreeableness-like domain using natural language approaches (e.g., Costa & McCrae, 1992; Heaven, Connors, & Stones, 1994) suggests that agreeableness-related traits describe fundamental ways in which human beings think about and relate to one another.

Though there is not a strict, agreed-upon definition of agreeableness, most researchers broadly characterize agreeableness as a personality domain related to individual differences in motivations to maintain positive social relations with others (Graziano & Tobin, 2017; Graziano & Eisenberg, 1997). From this perspective, individuals high in agreeableness are likely to be consistently motivated to maintain harmonious relations across many interpersonal contexts, whether it be with a romantic partner or an acquaintance. At the opposite pole, individuals low in agreeableness may not place as much value on interpersonal harmony and would be more likely to sacrifice interpersonal harmony for other goals.

Consistent with this account, a substantial empirical literature has shown that high self-reported agreeableness is related to a variety of prosocial behaviors (Graziano, Habashi, Sheese, & Tobin, 2007; Habashi, Graziano, & Hoover, 2016), more effective management of interpersonal conflict, and positive functioning in interpersonal situations more generally (e.g., Jensen-Campbell et al., 2002; Jensen-Campbell & Graziano, 2001). Meanwhile, individuals who describe themselves as being low in agreeableness also report being more aggressive (e.g., Miller & Lynam, 2006; Reardon, Tackett, & Lynam, 2018) and meta-analytic evidence has indicated that relative to other FFM domains, (low) agreeableness shows the largest relations with a variety of self-reported antisocial behaviors including reactive and proactive aggression, and non-violent antisocial behavior (Vize, Miller & Lynam, 2018; Vize, Collison, Miller, & Lynam, 2019). Recent empirical work has also used multivariate techniques to show that the inclusion of antagonism-related traits within broader constructs (e.g., psychopathy, narcissism) are largely responsible for the broader constructs' relations with antisocial and aggressive behaviors (e.g., Vize et al., 2017; Vize, Miller, Collison, & Lynam, in press; Vize, Collison, & Lynam, in press).

Other personality models like the HEXACO (Lee & Ashton, 2004) and Interpersonal Circumplex (IPC; Wiggins & Broughton, 1985), have also provided insights into how personality traits related to agreeableness relate to prosocial behaviors. The HEXACO model includes both Agreeableness and Honesty/Humility (H/H) domains. The theoretical distinction between the two domains is couched in terms of interpersonal cooperation: H/H reflects a tendency to cooperate with others even when there is opportunity to exploit them while HEXACO-Agreeableness reflects a tendency to cooperate with others even when the other person has shown to be uncooperative (Ashton, Lee, & de Vries, 2014). In other words, H/H is thought to assess a tendency to actively cooperate while HEXACO-Agreeableness is thought to assess a tendency to reactively cooperate (Hilbig, Glöckner, & Zettler, 2014). Although there is little evidence that the H/H factor assesses traits that are not already adequately assessed by certain measures of the FFM (e.g., the NEO-PI-R; Costa & McCrae, 1992; Miller et al., 2011, Crowe et al., 2018), the HEXACO model has been leveraged to help explain how personality is involved in various prosocial behaviors.

Most notable is the work by Hilbig and colleagues (Heck, Thielman, Moshagen, & Hilbig, 2018; Hilbig & Zettler, 2015; Hilbig, Glöckner, & Zettler, 2014; Hilbig &

Zettler, 2009). These researchers have investigated how self-reported H/H and agreeableness (as conceptualized in the HEXACO, FFM, and BFI) relate to cheating, exploitation, and cooperative behavior using a variety of paradigms. One paradigm used in this research is the ultimatum game, in which one participant is tasked with allocating some good (e.g., money or points) between themselves and another participant, and the other participant can choose to accept or reject the offer. The consequences of the offer being rejected are manipulated by the experimenter. The consequences can be set such that the allocator does not receive any consequence if their offer is rejected (i.e., the allocator has absolute power to offer whatever amount they see fit). Alternatively, if the participant rejects an offer, both the allocator and their partner receive nothing (i.e., the allocators partner has substantial power). Across the different allocation scenarios, individuals high in self-reported H/H allocate a similar amount of points to themselves, which highlights that individuals high in self-reported H/H show cooperative, prosocial behavior even when they have the opportunity to benefit themselves without consequence (Hilbig & Zettler, 2009). In regard to cheating behaviors, a recent analysis of 16 studies utilizing basic cheating paradigms (e.g., do respondents lie about the results of a coin toss or die roll to receive a payoff) found that both self-reported FFM-Agreeableness and H/H were related to lower odds of cheating, with H/H showing a larger effect compared to FFM-Agreeableness (Heck et al., 2018).

Within the IPC, Agreeableness maps onto the Affiliation (or Love) axis, while the other primary axis of the IPC is termed Dominance which maps on to FFM extraversion. Given that Agreeableness is conceptualized as largely interpersonal in nature, there is strong correspondence between IPC-Affiliation and FFM-Agreeableness; the FFM and

IPC are thought to complement one another (McCrae & Costa, 1989; Trapnell & Wiggins, 1990). With its focus on interpersonal contexts, research utilizing interpersonal theory has provided insights into how interpersonal tendencies are important to understanding both adaptive (e.g., Hopwood et al., 2018) and maladaptive outcomes (e.g., personality pathology; Wright et al., 2012). The theoretical underpinnings of agreeableness outlined above provide important insights into the types of traits and behaviors that fall under the umbrella of agreeableness. However, much of the empirical work described above has been descriptive in nature, relying on self-report assessments of agreeableness-related traits, and has kept its focus at the level of interpersonal behaviors like altruistic helping behaviors or aggression. Greater insights may be gained from exploring the more “upstream” affective and cognitive processes of agreeableness.

Agreeableness-Related Processes

The large majority of research on agreeableness assesses the domain with self-report measures. Specifically, respondents are typically asked to report their perception of their thoughts/behaviors aggregated across extended periods of time. Process based personality assessment is notably less aggregated (i.e., confined to a specific context) and often relies on task performance. It is important to note that in the following discussion, we diverge from the typical conceptualization of agreeableness-related processes that views processes as outcomes driven by self-reported agreeableness. Research in this area tends to frame results so as to suggest trait agreeableness *influences* or *affects* cognitive/emotional processes believed to be involved in more downstream, agreeable behaviors. However, framing results in this manner leads to issues of circular reasoning—a self-reported trait influences processes that are believed to give rise to the patterns of behaviors reflected in the self-reported trait itself (Baumert et al., 2017). Instead, we emphasize that individual differences in behavior, self-reported traits, and cognitive/affective processes are all part of a broader personality domain despite differences in assessment approaches and degrees of abstraction of the broader trait.

Process based personality research has become more common as researchers look beyond descriptive personality research to better understand how various traits come to be expressed in different contexts. As Baumert et al. (2017) highlight:

> Process-oriented research in personality has primarily addressed this key task of explaining behavior. Process-oriented theories describe ideas about the particular intra-individual processes that guide behavior in transaction with situational cues and affordances. These theories propose systematic inter-individual differences in how these processes unfold. (p. 504)

Though the definition above applies equally well to interpersonal processes (e.g., the macro-level processes at play in a dyadic interaction), we are primarily focused on cognitive/emotional processes. Thus, this account of process-oriented research suggests that individual differences in cognitive and emotional processes can be thought of as part of the network that constitutes a personality domain. In turn, process oriented research takes a performance based assessment approach (e.g., reaction time during a theoretically relevant task) in order to better understand how individual differences may manifest in specific contexts.

There are a small number of researchers who have aimed to explicate which specific cognitive and affective processes may be reliable indicators of agreeableness.

Robinson and colleagues (e.g., Meier & Robinson, 2004; Meier, Robinson, & Wilkowski, 2006; Robinson, 2007; Wilkowski, Robinson, & Meier, 2006) have offered process-related understandings of multiple FFM domains, including agreeableness. These authors highlight that multiple processes intervene between the time a stimulus is first processed and subsequent emotional and behavioral outcomes. In turn, there are important stages throughout the course of these processes where individual differences relevant to agreeableness may be observed.

Robinson (2007) argues that individual differences in agreeableness are likely to manifest during two stages of processing: affect control and emotion control, where the former is related to processes that intervene to counteract priming effects related to affective thoughts and the latter is related to reducing the likelihood that the affective thoughts lead to further downstream emotional reactions (e.g., an angry facial expression). At both of these stages, individual differences in agreeableness manifest as inhibitory processes. In regard to affect control, Wilkowski and Robinson (2007) propose that individual differences in agreeableness can be demonstrated through re-appraisals of initially hostile interpretations of stimuli. In relation to emotional control, the authors argue that agreeableness-related processes are involved in actively suppressing outward manifestations of anger (e.g., aggression).

Concerning affective control, research has found that individuals high in self-reported agreeableness are able to more quickly categorize prosocial target words following a hostility-related priming word, whereas those low in self-reported agreeableness were slower to categorize the same prosocial target words following the hostility-related prime (Meier et al., 2006). In other words, the accessibility of supposed prosocial thoughts actually increases following a hostile prime for individuals high in self-reported agreeableness, which the authors suggest may be a product of an automatic “defusing” process that takes place after recognition of a hostile stimulus. In regard to emotional control, research has shown that neuroticism’s relation with trait anger and self-reported aggression is moderated by agreeableness, such that the relation is attenuated for individuals who are also high in agreeableness (Ode, Robinson, & Wilkowski, 2008). Moreover, though individuals high in self-reported agreeableness are just as quick to characterize words in terms of their blameworthiness (e.g., words like *negligence* or *malpractice*), the ability to quickly characterize such words was only related to antagonistic behavior for those low in agreeableness (Meier & Robinson, 2004). These results suggest that the *accessibility* of hostile thoughts may not differ between individuals low and high in agreeableness, but agreeable individuals may inhibit the initial hostile thoughts from leading to further emotional reactions, and in turn, aggressive behaviors.

Additional research, focusing on Gross’ (1998) model of emotion regulation stages, found that individual differences in self-reported agreeableness were related to Gross’ first stage of emotion regulation, situation selection. Using a picture viewing task where participants could view positively and negatively valenced pictures for as long as they wanted, Bresin & Robinson (2015) found that individuals low in agreeableness spent a longer amount of time viewing negative pictures compared to positive pictures, while individuals high in agreeableness showed no such preference (i.e., highly agreeable individuals were quicker to move on to the next picture compared to disagreeable individuals). In a separate study (Study 3; Bresin & Robinson, 2015), the authors found that individuals high in agreeableness showed a significantly stronger preference for positively valenced activities (e.g., having to choose between watching a movie described as a “gut-busting comedy” versus a “fast-paced, violent thriller”). These results were taken to suggest that agreeableness manifests in the earliest stage of emotion regulation: the types of stimuli one is likely to be exposed to in the first place.

Other process related accounts of agreeableness are largely complementary to the views of Robinson and colleagues. Graziano and colleagues (2010; 2013) have emphasized the role of agreeableness in dual-process models of behavior which have been proposed to help explain behaviors with close ties to agreeableness (e.g., helping behaviors and prejudice). The dual-process model emphasizes that two competing cognitive and emotional processes are activated by an external stimulus. In the case of helping, the stimulus may be someone in distress that requires assistance. The two processes activated are that of personal distress and empathic concern, with the former being more strongly related to neuroticism and the latter related to agreeableness. Graziano and Habashi (2010) argue that personal distress occurs first, followed by empathic concern and these two states are opponent processes. In other words, they argue that empathic concern works to inhibit personal distress, which otherwise may result in escape or avoidance from the situation if left unchecked, assuming escape/avoidance behaviors are available.

In sum, process models of agreeableness suggest that inhibitory processes may help explain *how* people come to act in agreeable ways. Specifically, these processes have been invoked to explain how prosocial behaviors may come about (e.g., helping others in distress), as well as the avoidance of antisocial behaviors (e.g., aggression).

These models are also consistent with developmental research that has shown traits related to self-regulation and inhibitory control are the developmental precursors to traits related to agreeableness and conscientiousness (e.g., Jensen-Campbell et al., 2002; De Pauw & Mervielde, 2010). However, all of the empirical support for these models considered thus far has used assessment approaches that only consider agreeableness at the domain level. Recent work within personality assessment has provided a compelling rationale for looking beyond the domain level.

Studying Personality at Different Levels of Specificity

In recent decades, researchers have examined self-reported personality at varying degrees of specificity within the broader FFM hierarchy. For example, at the most fine-grained level of analysis one can examine individual items within self-report personality measures (termed “nuances”; Mõttus et al., 2017). At a less specific level, particular measures of the FFM like the NEO-PI-R (Costa & McCrae, 1992) allow one to make use of facet scales that assess related but conceptually distinct traits within a domain (e.g., the Trust and Modesty facet scales from the agreeableness domain). Other work has identified ten personality “aspects” that lie between the facets and domains (DeYoung, Quilty, & Peterson, 2007). These self-report assessment approaches have demonstrated that moving beyond domain level analyses can provide useful information about more specific, homogenous features of personality that exist within the broader domains (Smith, McCarthy, & Zapolski, 2009). Of most relevance to the present paper is work that has examined the predictive utility of the FFM facet scales.

In relation to the facets of agreeableness, meta-analytic evidence has shown that although all agreeableness facets are negatively related to antisocial behavior and aggression, some facet scales show larger negative relations than others (e.g., the Compliance and Straightforwardness facets; Vize, Miller, & Lynam, 2018). Other work has shown that the facet scales of agreeableness account for significant variance in externalizing behavior outcomes, beyond that accounted for by the domain alone (e.g., Klimstra, Luyckx, Hale, & Goossens, 2014). Using dominance analysis, Vize and colleagues (2017) showed that the antagonism facet scales of an FFM-based measure of narcissism showed differential importance in their relations to various self-reported externalizing outcomes.

Importantly, facet scales within broader domains may also demonstrate divergent or even opposing relations. Thus, domain-focused assessment approaches may lead to inappropriate estimates of the broad domains' empirical relations. The clearest example of such findings is found for extraversion. Within extraversion, facet scales related to positive emotionality (e.g., the Warmth and Positive Emotions scales from the NEO-PI-R) tend to show negative relations with externalizing outcomes, while facet scales related to excitement/thrill seeking are positively related to the same externalizing outcomes with the end result being null relations between the domain and these outcomes (Watson, Stasik, Ellickson-Larew, & Stanton, 2015; Jones, Miller, & Lynam, 2011).

However, current FFM instruments include facet scales that are conceptually-based rather than empirically derived (Costa, McCrae, & Dye, 1991). Work has now been conducted to empirically explicate the hierarchical structure of traits within the broader domains of extraversion (Watson, Stasik, Ellickson-Larew, & Stanton, 2015), conscientiousness (Roberts, Chernyshenko, Stark, & Goldberg, 2005), and most recently, agreeableness (Crowe et al., 2018). Crowe and colleagues (2018) used the “bass-ackwards” approach (Goldberg, 2006) to empirically identify the underlying structure of the agreeableness domain based on 22 self-report agreeableness scales (initial item pool of *N*=131). In doing so, Crowe and colleagues were able to identify, in a step-wise fashion, the structure of agreeableness that emerged starting at the broadest level (i.e., the agreeableness domain) to more specific but still conceptually coherent factors while also estimating the interrelations of the extracted factors at each step of the analysis.

The results for agreeableness showed that at the final level of the hierarchy at which the factors remained interpretable and contained relatively homogenous content, 5 factors were shown to underlie the broader domain. These factors were labeled compassion (vs. callousness), morality (vs. immorality), modesty (vs. arrogance), affability (vs. combativeness), and trust (vs. distrust). The five factors demonstrated divergent validity with relevant outcomes like proactive and reactive aggression (*r* range = -.13 to -.50; the affability factor showed the strongest negative relations to both outcomes), and drug use and criminal behavior (*r* range among the factors = .07 to -.34; the morality factor showed the strongest negative relations to both outcomes).

The Current Study

Although the merits of studying self-reported agreeableness at more fine-grained levels of analysis have been supported in numerous studies, work on agreeableness-related processes has yet to investigate how these finer-grained assessments may inform process-based research. Thus, the aim of the present study is to first replicate results from process-focused work on agreeableness using the same experimental designs and paradigms in previously published research. Next, we will examine whether these paradigms relate to specific facets of agreeableness.

In regard to the first aim, calls for more replication efforts and increased methodological rigor have become more prominent in light of the fact that many psychological findings appear to be less robust than initially thought (Open Science Collaboration, 2015). A variety of reasons have been proposed to explain why many findings may not replicate (e.g., Hoenig & Heisey, 2001; Nosek, Spies, & Motyl, 2017; Simonsohn, Nelson, & Simmons, 2014). In response, many researchers have argued for incentivizing replication studies and the use of open science practices which emphasize transparency throughout the research and publication process (e.g., Nosek & Lakens, 2014; Simonsohn, 2015). Partly in response to these arguments, the present project focuses on replicating previous research findings and also on incorporating the best available open science practices.

In regard to assessment of agreeableness processes, there are two primary reasons for incorporating recent advancements in self-reported agreeableness assessment. First, it remains an open question whether task performance assessments of agreeableness will be related to distinct, self-reported facets of agreeableness. Much of the work that has demonstrated the discriminant utility of facet scales has tended to use self-reported behavior outcomes, and to a lesser extent, actual behavioral outcomes (Paunonen & Ashton, 2001).

Second, incorporating the use of facets in studying agreeableness-related processes has the potential to inform and expand upon theoretical accounts of agreeableness. If agreeableness facets do show positive relations with process tasks (e.g., the Affability facet is the primary driving force of the relation the broader agreeableness domain has shown with specific cognitive tasks), it suggests that there are characteristics unique to the facet that may be differentially important to understanding the processes that lead to antisocial behavior. However, results showing a positive relationship with the domain but no divergence across facets suggest that the task-based paradigms are likely assessing agreeableness content that is not unique to a particular self-reported facet of agreeableness. In order for such conclusions to be drawn, however, it requires the use of different paradigms that allow for potential differences in relations to emerge in the first place. The present study will make use of 2 separate paradigms examining agreeableness-related processes. The paradigms are taken from previously published research. The utility in first replicating previous effects found for the process-based paradigms is that it allows greater confidence that the paradigms are validly assessing processes that fall within the network of agreeableness. In other words, as research on the FFM moves beyond descriptive work, it is essential that paradigms used to assess purported processes of specific personality domains are valid.

Though there are a variety of process-based tasks that could be chosen, we chose the paradigms for the proposed study based on their purported location within process-based models of behavior and emotion. Specifically, the two paradigms fall at different stages of stimuli processing (Crick & Dodge, 1996; Gross, 1998) ranging from the earliest stages (i.e., choosing what situations one exposes oneself to) to later stages involving attentional control. Thus, it is important to note that the cognitive-emotional process paradigms chosen for the current study assess very specific processes underlying differences in agreeableness; numerous other processes are also relevant to this area of research.

METHOD

Participants

Participants were recruited from Amazon’s Mechanical Turk, an online crowd-sourcing platform (MTurk; Paolacci & Chandler, 2014), and were paid a total of $2.50 for their participation in the study. The planned sample size was approximately 500 based on power requirements (see Power Analyses section below). We planned to recruit 600 MTurk workers, and 599 workers complete the protocol. After removing participant data based on exclusion criteria (described in the results section), the final sample size was $N = 517$.

In the studies we sought to replicate, undergraduate participants were used, as opposed to MTurk workers. However, based on the theoretical underpinnings of the original studies, there is no reason to expect that personality processes will function differently across undergrad versus MTurk samples. Furthermore, the recently published Many Labs 2 project found that very little variability in effect sizes could be attributable to “hidden moderators” (e.g., lab-based study vs. online, WEIRD vs. less WEIRD samples, etc.; Klein et al., 2015).

Self-Report Measures

Demographics Questionnaire

The demographics questionnaire asked basic questions regarding gender, age, ethnicity, education level, and employment status.

Goldberg’s Big Five Markers

Goldberg’s Big Five Markers (BFM; Goldberg, 1992) is a 50-item self-report inventory that is widely used to assess the five domains of the Big Five personality model. The 10-item Agreeableness scale from the BFM was used to assess Agreeableness in the studies the current project seeks to replicate. Thus, only the items from the BFM that assess Agreeableness were administered to participants. The internal consistency (α) of BFM-Agreeableness was .90.

International Personality Item Pool-NEO (IPIP-NEO)

The IPIP-NEO-120 (Maples et al., 2014) is a freely available self-report measure of personality designed to assess the FFM. Like the NEO-PI-R, the IPIP-NEO includes facet scales for each FFM domain. Previous work has shown that the IPIP-NEO reliably and validly assesses the FFM domains and facets (Maples, et al., 2014). All domains of the IPIP-NEO were administered.

As previously mentioned, Crowe et al. (2018) identified 5 subfactors underlying the domain of agreeableness: compassion, morality, modesty, affability, and trust. The results from Crowe et al. (2018) also showed that the agreeableness facet scales from the IPIP-NEO strongly relate to each of the empirically derived agreeableness factors (*r* range = .82 to .89) and can thus serve as proxies of the factors identified in their article. The IPIP-NEO facet scales include altruism, sympathy, morality, trust, cooperativeness, and modesty. The one exception was the modesty facet scale from the IPIP-NEO which showed a less than ideal relation with the modesty (vs. arrogance) factor (*r* = .55). Thus, we substituted the Modesty subscale (10 items) of the Facet Inventory of the Five Factor Model (FI-FFM; Simms, 2009), which was strongly correlated with the modesty factor (*r* = .89) identified in the Crowe et al. (2018) study. Internal consistencies of the facet scales were .72 (Morality), .75 (Altruism), .76 (Cooperativeness), .87 (Trust), and .89 (Modesty) while the internal consistency of the IPIP-Agreeableness domain scale was .87.

Conditional Reasoning Test for Aggression (CRT-A)

The CRT-A (James et al. 2005) is a 22-item conditional reasoning questionnaire that is designed to assess participants' cognitive precursors to aggression. Participants' are presented with 22 reasoning problems and must select one of four options. Of the four options, there is one "aggressive" response option, a non-aggressive response option, and two illogical response options. Scores on the CRT-A are calculated by giving participants a 1 on every item where they choose the aggressive option, a 0 for choosing an illogical response option, and a -1 for the non-aggressive option. Two example items from the CRT-A are displayed in Appendix C. The CRT-A was chosen for inclusion due to its purpose of indirectly assessing (it is described to participants as a task of logical reasoning) the implicit biases that make certain individuals more likely to engage in aggressive behavior. The CRT-A has been validated across a range of samples; its relation with behavioral criteria across these samples (mean correlation =.44) has exceeded those typically seen for tests of aptitude and self-reported personality. The internal consistency of the CRT-A total score was low (α = .33; but see Lebreton, Grimaldi, & Schoen, 2018 for a discussion of internal consistency as it applies to conditional reasoning tasks).

Elemental Psychopathy Assessment-Short Form (EPA-SF)

The EPA-SF (Lynam et al., 2013) is a 72-item short form of the full-length EPA (Lynam et al., 2011). It is a self-report measure of psychopathy that provides scores on 18 psychopathic traits as well as a global psychopathy score. A global score can be calculated by taking the mean of the scores on all 72 items, yielding an aggregate score ranging from one to five. Participants rate the items on a 5-point Likert scale ranging from disagree strongly to agree strongly. The EPA-SF is underlain by four factors: Antagonism (consisting of Callousness, Coldness, Distrust, Manipulation, and Self-centeredness), Disinhibition (Opposition, Rashness, Thrill-Seeking, Urgency, Disobliged, Impersistence), Emotional Stability (Unconcern, Self-contented, Invulnerable), and Narcissism (Anger, Dominance, Self-assured, Arrogance). Although the EPA-SF was included in our battery of self-report assessments, analyses focused on the EPA-SF were not the focus of the pre-registered report and thus were not included in our planned analyses nor in any results in the registered report.

However, in order to assess for careless or invalid patterns of responding, two validity scales (Too Good to be True and Infrequency; 8 items each) from the original EPA (Lynam et al., 2011) were administered to participants. Example items include "I have never in my life been angry at another person" and "I try to eat something almost every day."

Reactive-Proactive Aggression Questionnaire (RPQ)

The RPQ (Raine et al., 2006) is a 23-item self-report measure which measures both reactive and proactive aggression. Participants endorse items by choosing either 0 (never), 1 (sometimes), or 2 (often). Examples include "Had fights with others to show who was on top" and "Reacted angrily when provoked by others." Internal consistencies for the subscales were .82 (Reactive Aggression) and .84 (Proactive Aggression).

Crime and Analogous Behavior Scale (CAB)

The CAB (Miller & Lynam, 2003) is a self-report measure that asks respondents whether or not they have engaged in a range of activities over the past year and the frequency of such activities. Items assess a variety of externalizing behaviors including substance use, physical aggression, stealing, and gambling. In line with previous research, four subscales were computed: aggressive/violent behavior (α = .43), non-violent antisocial behavior (α = .62), substance use (α = .67), and gambling (α = .76).

Paradigms

Situation Selection Task

The situation selection paradigm used by Bresin and Robinson (2015) involves examining the amount of time spent looking at negatively and positively valenced pictures from the International Affective Picture System (IAPS; Lang, Bradley, & Cuthbert, 1997). Specifically, 50 pictures of each valence (i.e., positive and negative; matched on arousal) were presented to participants in a random sequence. The same 100 IAPS pictures used in the original study were used in the current study. Participants simply pressed the spacebar when they were done viewing the picture. Responses faster than 300 ms resulted in a text display encouraging participants to take time to view each picture. Viewing time of the pictures served as the dependent variable. In their original study, Bresin and Robinson (2015) found that there was a main effect of picture valence such that on average, participants viewed negatively valenced pictures longer than positively valenced pictures. However, this effect was moderated by agreeableness: individuals low in agreeableness spent longer amounts of time viewing negative images compared to positive images while individuals high in agreeableness displayed no preference for negative pictures compared to positive pictures.

The reliabilities of reaction times were computed separately for positive and negative pictures using split half reliabilities, split by even and odd trial numbers. The reliability of reaction time differences between positive and negative pictures (i.e., RT for positive pictures – RT for negative pictures) was also calculated using the same split-half procedure. All split-half reliabilities were then corrected using the Spearman-Brown formula. The corrected split-half reliabilities of negative, positive, and the difference between negative-positive trials were .63, .64, and .36, respectively.[1]

Spatial Attention Paradigm

Based on an extension of well-established spatial attention paradigms (Posner et al., 1980), we made use of the task implemented in Wilkowski et al. (2006) to examine participants' ability to disengage attention from prosocial and antisocial cues in order to respond to a target stimulus. For each trial of this task, a white fixation cross was displayed at the center of the screen, with two rectangular boxes, depicted as transparent with yellow borders, displayed to the left and right of the fixation cross. The boxes were displayed 250 pixels to the left and right of the fixation cross, allowing for the targets to be displayed at least one inch from the center across a range of different monitor sizes. Trials were presented on a full screen (1000 x 800 pixel resolution) with a black background. Participants were asked to categorize the prosocial or antisocial cue word that is presented within either the left or right box by pressing the "q" or "p" key. A total of 10 different cue words were used. The five antisocial cue words were shoot, hit, stab, kill, kick and the five prosocial cue words were hug, smile, praise, help, give. Participants were asked to categorize these words as either "helpful" or "hurtful"[2]. The association

[1] Though the split-half reliabilities of these variables were not reported in the original publication, they were computed using the data made available by the original study authors. The corrected split-half reliabilities of negative, positive, and the difference between negative-positive trials were .74, .80, and .36, respectively. Thus, the split-half reliabilities were larger for positive and negative trials in the original study, and the reliability of the difference between negative and positive trials was equivalent in the original and current study.

[2] These words are chosen over prosocial and antisocial because they are more likely to be familiar to participants.

between key (q or p) and cue type (helpful/hurtful) was counter-balanced across participants. In addition, a banner was displayed at the top of the screen indicating “q = helpful & p = hurtful” or “q = hurtful & p = helpful.” The cue word remained on the screen until the computer registered a response. Once a response was registered, a 50 ms blank delay occurred before the spatial target was presented. The spatial target appeared randomly at either the same or opposite location of the cue word, and the target was either the letter “q” or “p”. Participants were asked to press the corresponding key as quickly as possible while maintaining accuracy. Following a response, a 200 ms delay occurred before the start of the next trial. In cases where participants pressed the wrong key when identifying the cue or target, a 4000 ms error message was displayed encouraging participants to respond as accurately as possible. The sequence of cues (antisocial or prosocial) and target locations (same or different) and target identities (q or p) were determined randomly for each participant. Participants first completed 20 practice trials before completing 160 trials (approximately 40 trials per cell), the latter of which were used for analyses.

Importantly, the spatial attention task described above differed slightly from the one used in the Wilkowski et al. (2006) publication. In the original study, participants wore a headset with a microphone and audibly stated whether the cue was a helpful or hurtful word. After a vocal response had been registered, the target (either a q or p) was then presented and participants had to press the corresponding key as quickly as possible. Because the study was completed online, vocal responses were not considered. As mentioned, participants instead categorized the cue as helpful/hurtful by pressing either the “q” or “p” keys. In order to ensure that this deviation did not substantively alter the fidelity of the replication effort, the corresponding author was contacted by email and the potential modification was explained while planning the study. The corresponding author offered that the banner indicating what letter corresponded to what type of cue (i.e., the banner that reads “q = helpful & p = hurtful”) can be displayed to minimize potential difficulties of the task but otherwise stated that the modification should not impact the validity of the paradigm (B. Wilkowski, personal communication, March 23, 2018).

An additional concern is whether reaction time data collected over online platforms allowed for the necessary precision required for most reaction time-based tasks. Recent work has shown that online platforms can indeed be relied on to produce reliable results for cognitive and perceptual experiments (e.g., Germine et al., 2012). Relatedly, Crump, McDonnell, & Gureckis (2013) made use of MTurk samples to examine the validity of results for a variety of reaction time tasks that require millisecond control, including the Posner cuing task. The Posner cueing task was presented to workers as an HTML web page with the task flow controlled by JavaScript code which was run locally in each MTurker’s web browser. The results for the Posner task showed that the typically observed cuing effects were replicated, and even small reaction time effects (~20 ms) could be reliably measured despite the high likelihood that there was variability in web browsers and computer hardware used by the participants. The software used for the present project also made use of JavaScript code to run the experimental tasks, and was run locally on participants’ web browsers.

The reaction time outcome examined was the 3-way interaction effect of agreeableness, cue type (prosocial vs. antisocial), and location (target and cue appear at the same location vs. target and cue appear at opposite locations) on target reaction time.

Wilkowski et al. (2006) found that the nature of the three-way interaction was such that while individuals low in agreeableness showed a higher disengagement cost (i.e., the reaction time difference between same vs. different target/cue location trials) after categorizing antisocial cue words, individuals high in agreeableness show a higher disengagement cost after categorizing prosocial cue words. The reliability for four components of the spatial attention were computed: reaction time on valid trials, reaction time on invalid trials, reaction time on antisocial cue/invalid trials, and reaction time on prosocial cue/invalid trials using corrected split-half reliabilities. The split-half reliabilities for valid trials was .52 and .59 for invalid trials. For antisocial cue/invalid trials and prosocial cue/invalid trials, the respective reliabilities were both .56.[3]

Planned Analyses and Hypotheses

Situation Selection Task

Due to the repeated nature of the situation selection task, a multilevel model (MLM) was used to test hypotheses regarding the situation selection task. In contrast to the original publication, we included the maximum number of random effects parameters. There are various theoretical and statistical reasons to include the maximum number of random effects parameters (e.g., Judd, Westfall, & Kenny, 2012; Westfall, Kenny, & Judd, 2014), and the inclusion of the maximum number of random effects is generally recommended in confirmatory testing contexts (Barr et al., 2013). Thus we aimed to replicate the original findings under optimal testing conditions.

[3] The split-half reliabilities of these components were not reported in the original publication. Using the data made available by the corresponding author, the split-half reliabilities of valid and invalid trials were .44 and .46 while the reliabilities of the antisocial cue/invalid trials and prosocial cue/invalid trials were .46 and .47, respectively. Thus, the split-half reliabilities in the present study were slightly larger than in the original publication.

For the picture viewing task, the random effects included random slope parameters for the within-subject effect of picture and picture valence. Random intercepts for subject, picture, and picture valence were also included. Picture valence was entered as a level 1 predictor, with grand-mean centered agreeableness entered as a level 2 predictor. A two-way, cross-level interaction term (agreeableness x valence) was entered to examine the effect of agreeableness and valence (coded 0/1 for positive/negative) on picture viewing times.

Hypothesis 1. We aimed to replicate the finding that agreeableness will interact with picture valence to predict viewing times of the IAPS pictures (original interaction effect: $F(1, 7191) = 4.52$, $p = .03$) with individuals low in agreeableness viewing negative pictures longer than positive pictures while individuals at high levels of agreeableness will not show a preference for viewing negative pictures relative to viewing positive pictures (Hypothesis 1a). Figure 1 displays the millisecond differences observed in the original study (Bresin & Robinson, 2015) for picture viewing times as a function of Agreeableness and picture valence. Similar to the original study, we tested for a main effect of picture valence, predicting that participants, in general, will view negative pictures longer than positive pictures (Hypothesis 1b).

Spatial Attention

Similar to the situation selection tasks, a MLM was used to analyze the results of the spatial attention task, and the maximum number of random effects parameters were included in the model. Specifically, we included random slopes for the within-person effect of cue words (hit, kick, hug, etc), cue type (prosocial vs. antisocial), cue-target location (same vs. different), and the interaction of cue type and cue-target location. We included random intercepts for participants, cue words, cue type, and cue-target location. Cue type and cue/target location were entered as level 1 predictors, with grand-mean centered agreeableness entered as a level 2 predictor. A three-way, cross-level interaction term (agreeableness x cue x location) was entered in order to examine the effects of cue and agreeableness on attention disengagement costs (the reaction time difference between same vs. different cue/target location trials).

Hypothesis 2. We aimed to replicate the initial finding that agreeableness will be significantly related to disengagement costs for individuals both high and low in Agreeableness (original interaction effect: $F(1, 64) = 6.52$, $p = .01$), such that individuals low in agreeableness will show disengagement costs for antisocial cues while individuals high in agreeableness will show disengagement costs for prosocial cues (Hypothesis 2a; see Figure 1 for disengagement costs as a function of Agreeableness and cue type observed in the original study).

Exploratory Analyses

All primary analyses outlined above were confirmatory in nature and dealt with replicating previous research findings. However, we also conducted similar analyses for the facet scales of agreeableness in order to examine their potential unique predictive abilities. Unfortunately, previous work did not lead to straightforward hypotheses for facet level analyses, and thus, the facet analyses were exploratory. However, we offered directional hypotheses for our exploratory analyses and tentative, facet-specific hypotheses in some cases. All exploratory facet analyses were conducted regardless of whether the primary analyses that focus on the broader domain showed significant effects.

Situation selection task. For the picture viewing task, the five agreeableness facet scales were substituted for the broader domain and their respective interactions with picture valence were examined. We did not have any *a priori* hypotheses for which facets of Agreeableness would be differentially related to viewing positive versus negatively valenced pictures. However, we expected that the nature of the relationship, if present, will be the same as the relation predicted for the broader domain: those who are low on the Agreeableness facet(s) will show a preference for negatively valenced pictures (Hypothesis 3).

Spatial attention. Following the replication analysis focusing solely on the agreeableness domain, the additional analyses that were conducted involved substituting the five facet scales of agreeableness for the single agreeableness domain as level two predictors. Then, the three-way, cross-level interactions between each of the facets and cue x location was examined. This allowed for the exploration of whether any one of the agreeableness facets shows unique effects above and beyond the other facets of agreeableness. If the attentional processes involved in the spatial attention paradigm outlined above are thought to index a necessary precursor to aggressive behavior, then the facet most likely to predict disengagement costs would be (low) Cooperativeness (Hypothesis 4) given previous meta-analytic evidence that has shown this facet to be most strongly related to such behavior (e.g., Vize, Miller, Lynam, 2018). The nature of the relation would be similar to the broader domain: individuals low in Cooperativeness would show a larger disengagement cost on antisocial cue trials (Hypothesis 4a). We had no *a priori* hypotheses regarding which agreeableness facets would be related to disengagement costs on prosocial trials.

To further explore the validity of the spatial attention paradigm as an indicator of attentional biases that may increase aggression, we examined the relation between disengagement costs from the spatial attention task and our measures of antisocial and externalizing behavior (the CRT-A, RPQ, and CAB). We expected the disengagement cost for trials with antisocial cue words (i.e., the difference in reaction times on trials with same vs. different cue-target location when antisocial word cues are displayed) to be significantly positively related to scores on the CRT-A (Hypothesis 5), and that the relation between CRT-A and disengagement costs for antisocial cues would be accounted for by their shared relation with agreeableness (Hypothesis 5a).

Two separate analyses were conducted. In the first, we examined the zero-order correlation between disengagement cost for antisocial cue trials and the total score of the CRT-A and compared it with the semipartial correlation between disengagement cost and the CRT-A, accounting for agreeableness’ overlap with the CRT-A.

Next, we computed zero-order correlations between disengagement costs (for both antisocial and prosocial cue word trials), the total score on the CRT-A, reactive and proactive aggression subscales from the RPQ, and aggression, non-violent antisocial behavior, substance use, and gambling subscales from the CAB. We expected to find a positive manifold among our measures of antisocial behavior (the CRT-A, RPQ, and CAB) and the disengagement cost on antisocial cue trials (Hypothesis 6). We hypothesized that the correlations should be larger for subscales assessing aggression compared to scales measuring self-reported gambling and substance use. We planned to test the differences among these correlations using Steiger’s test of dependent correlations (Steiger, 1980). We also hypothesized Disengagement costs on *prosocial* cue trials would be negatively related to measures of aggression.

Supplementary Analyses

In order to ensure that the primary results (i.e., replication-focused results) are not due to the use of MLMs that include the maximum number of random effects parameters, we conducted additional analyses with MLMs that only contain a single random effect (random intercept for participants) for both the picture viewing and spatial attention tasks. These supplementary models were equivalent to the models used in the original studies. This allowed to examine whether the replicability of effects was reliant on modeling decisions.

Based on the planned analyses, there was a total of 37 tests planned for the exploratory hypotheses. To limit the proportion of Type I errors, we made use of the false-discovery rate (FDR) correction (Benjamini et al., 2006), which limits the proportion of false positives among significant findings, compared to other corrections (e.g., Bonferroni) which guards against making *any* Type I error. The former correction allows for greater statistical power while still taking into account a large number of tests. The FDR was set to 5%, such that if all tests are statistically significant, 1 result is likely to be a false positive.

Power Analyses

Because the present project is partially concerned with replicating previous research on agreeableness-related processes, power analyses were conducted for each of the primary planned analyses to ensure adequate power in detecting previously published effects. Power analyses were conducted for both the spatial attention and picture viewing tasks. Power analyses were conducted using the ‘simr’ package (Green & Macleod, 2016) in R (R Core Team, 2013). The ‘simr’ package allows for Monte Carlo simulation-based power analyses to be conducted for linear mixed models and generalized linear mixed models. Power analyses are conducted by first fitting a multilevel model to either pilot or simulated data using the ‘lme4’ package (Bates, Machler, Bolker, & Walker, 2015). Next, three steps are repeated over a pre-specified number of iterations: 1) a simulated value for the response variable is generated based on the initial model fitting; 2) the model is refit; 3) a significance test is conducted for the effect of interest. Power is computed based on the number of successes (i.e., significant results) and failures (i.e., non-significant results) derived from the simulations.

In order to derive estimates for the mixed model parameters, the researchers that served as corresponding authors for the Wilkowski et al. (2006) and Bresin & Robinson (2015) publications were contacted by email. The rationale for the present study was explained, and the original data used for the publication was requested for the purposes of power analyses.[4] All authors provided the requested data. Although these publications included multiple independent studies, only the data from Study 2 from Wilkowski et al. (2006) and Study 2 from Bresin and Robinson (2015) were used for power analyses as these samples used the tasks that will be used in the present study.

Multilevel models were then estimated for each of the tasks (spatial attention and picture viewing) using the same model specification outlined in the planned analyses section above (i.e., models with the maximum number of random effects). Results for the

[4] All pilot data and syntax necessary to reproduce the power analyses are available at https://osf.io/acqxu/. The corresponding authors were contacted by e-mail and the authors provided consent to having the pilot data posted on the OSF website.

multilevel model power analyses can be seen in Table 1. Initial power analyses were conducted using the raw data in order to estimate power for the previously published results. However, assuming that published effect size estimates are reflective of the true size of the effect can lead to biases in power estimation (Hoenig & Heisey, 2001). Thus, once power was estimated for the original data, smaller effect sizes were specified for the effect of interest (i.e., the interaction terms in the models) to examine the published studies' power to detect an effect if the 'true' effect was in fact smaller than the one observed in the sample.

The specified decrease in effect size magnitude was based on substantive considerations of what constituted a meaningful effect size. In other words, these considerations were made in line with recommendations that power analyses ought to consider the question, "What effect would one care about detecting?" (Morey & Lakens, 2016). To this end, we used the "detectability" heuristic (Simonsohn, 2015) to determine the smallest effect size of interest. The "detectability" heuristic outlined by Simonsohn (2015) focuses on statistical power and the sample size of the original study when determining effects one would care about detecting. Specifically, the effects we cared about detecting were those that the original studies would have had at least 33% power to detect. If we were to have observed an effect size in our data that the original study would have had a 20% chance of detecting (i.e., an effect that falls below the 33% criteria), it suggests that the original study would not have been able to meaningfully distinguish the effect from zero. This criterion remains informative even if we observe a small effect that is nonetheless statistically significant; such a result would suggest that even though the sign of the effect was replicated, the original study did not provide a meaningful understanding of the phenomenon. Thus, the effects we care about detecting are those that the original studies would have had at least 33% power to detect. However, we note that no single “replication statistic” exists—careful consideration of results is always necessary in order to determine whether or not the original results are consistent or inconsistent with the replication results.

Picture Viewing

The maximum number of random effect parameters were included in the model for the spatial attention data with the exception of a random slope parameter for the within-person effects of picture, as there were not enough observations in the data to estimate the parameter. Assuming an effect size equivalent to the observed effect size in the sample (which was marginally significant, compared to the significant fixed effect observed when only a random intercept for participant was included in the model), power was estimated at .48. We then substituted a smaller effect size for our power analyses and conducted sensitivity analyses using the ‘simr’ package to determine what effect sizes the original studies would have had a 33% chance of detecting, and used that effect size to determine the necessary sample size to achieve 90% power. The results showed that with a valence x agreeableness fixed effect of -.037, a sample size of approximately 500 participants was needed to achieve 90% power.

Spatial Attention

The same procedure described above was used to estimate power for the spatial attention task. However, when including random effects parameters for the spatial attention model, the inclusion of the parameters were a) estimated to be zero or close to zero, b) lead to model convergence issues, or c) had no effect on the cue type x location x agreeableness fixed effect. Thus, power analyses were based on a MLM with only a random intercept for participant included. Power was estimated at .60 for the original study, assuming the true effect size was equal to the effect size observed in the sample. After substituting a smaller effect size for the cue type x location x agreeableness effect (that which the original study would have had a 33% chance of detecting), the results showed that a sample of 500 would result in approximately 99% power to detect such an effect. Because a sample of 500 would ensure adequate power for both the picture viewing and spatial attention tasks, we planned to have a final sample of at least 500 participants.

Preregistration of Model Reduction Steps

As previously noted, including the maximal number of random effects parameters lead to convergence issues for the spatial attention task. It is unclear whether similar convergence issues would arise with a larger sample size. However, fitting maximal models can lead to convergence issues as a consequence of using a model that is too complex for the data (Bates, Kliegl, Vasishth, & Baayen, 2015). As a result, we applied an iterative process of reducing model complexity when convergence issues arose for either the spatial attention or picture viewing multilevel models. First, we constrained the correlation among random effects parameters to zero, and tested whether this constraint significantly decreases goodness of fit for the model. Should the models still not converge after imposing the zero-correlation constraint, we examined whether the elimination of random effects parameters lead to significant decreases in goodness of fit. We first tested higher order random effects before testing the lower order random effects nested within them. For example, the random slope parameter for the Cue Type x

Location interaction was tested before either the random slope parameters of Cue Type or Location in the spatial attention model. We tested for convergence each time a random effect parameter is eliminated, and repeated this process until convergence occurs.

Procedure

The situation selection and spatial attention paradigms were programmed using PsyToolkit (Stoet, 2010; Stoet, 2017). PsyToolkit is a free to use online platform that allows for researchers to design and implement a variety of cognitive experimental paradigms online which than can be integrated with crowdsourcing platforms such as MTurk. Participants were recruited through MTurk, and first completed the picture viewing task and spatial attention task through the PsyToolkit website.[5] The order of completion between the two tasks was decided randomly for each participant. Following the completion of those tasks, participants completed the remainder of the protocol using PsyToolkit's survey platform. The remainder of the protocol included the self-report measures and the CRT-A. The presentation of these measures was counter-balanced across participants.

Participant Inclusion Criteria

To participate in the study, we required that MTurk workers be at least 18 yrs. old, be English speakers, and have at least a 95% MTurk approval rating. In addition, PsyToolkit allows for researchers to specify that a real keyboard and/or specific internet browsers are required to complete the protocol. For the current study, participants were required to have a real keyboard and to use either Google Chrome or Mozilla FireFox

[5] This process was completed by having MTurk participants following a link to the PsyToolkit website, which then assigns a unique ID to each participant which can be used to link information between the sites.

(required for full screen display of the tasks) as their internet browser in order to participate.

In order for participant data to be included in analyses, participants must have completed the entirety of the protocol. Completion of the protocol was determined by whether or not the participants have successfully entered a randomly generated code, made available at the end of the protocol, into MTurk. Participant data from participants who complete the protocol in an unrealistic amount of time was excluded. We expected the protocol to take approximately 60-90 minutes to complete. Thus, participant data from participants who completed the protocol in under 20 minutes was rejected.

A total of five attention check items were administered over the course of the protocol (example item: “Please click the option that says “Very accurate””) in addition to five “captcha” questions (i.e., open ended questions that have verifiable responses: “What is 2+2?”). A failure of more than one of these ten items resulted in the participant’s data being rejected. As previously mentioned, two validity scales (8 items each; 1-5 Likert scale) from the EPA were also be administered. Similar to Lynam et al. (2011), participants were considered invalid responders and their data rejected if their responses for either scale were 2.75 standard deviations above the sample average. Last, participant data could not be filtered through a recently developed tool to flag suspicious GPS coordinates that are present in the data (Prims & Motyl, 2018) as we mistakenly assumed MTurk provided such information. Thus, geolocation checks for the data were not conducted.

Inclusion Criteria for Trial Data

To ensure that data from “button mashing” during the reaction time tasks was excluded, participants with response times faster than 300 ms for more than 10% of the total trials for either task were rejected. For the spatial attention paradigm, participants who made errors on more than 20% of the total trials had their data excluded from analyses.

In regard to the exclusion of outliers, we made use of the same outlier rules used in the previous publications. Specifically, in Wilkowski et al. (2006), reaction time values that were 2.5 SDs above or below the sample mean were windsorized before analyses were conducted. In Bresin & Robinson (2015), any trials with reaction times faster than 300 ms were excluded from analyses and the same exclusion rule was used in the proposed study. No data modification steps were described for longer picture viewing times, and the range of reaction times in the data was reasonable (i.e., the longest viewing time was 3,703 ms). Thus, we had no *a priori* rules for excluding longer viewing times.

Data Preparation

As is common for reaction time outcomes, the reaction times showed positive skew. Assumptions of linear models include normality of the outcome variable, and therefore the reaction time outcomes for the spatial attention task and picture viewing task were log-transformed to correct for the skewness of the data (Zandt, 2008). In addition, log-transformations of reaction time data were used in both the Wilkowski et al. (2006) and Bresin & Robinson (2015) publications.

The preregistered protocol described above received in principle acceptance before data collection began. The study registration can be found at the following link: https://osf.io/gxnys.

RESULTS

Data Exclusion

Data collection began on May 18th, 2019 with the goal of recruiting 600 MTurk participants. Data collection was completed Friday, June 14th, 2019 and a total of 599 participants completed the protocol. Based on the preregistered exclusion criteria regarding attention checks, time-to-completion, and re-captchas, 27 participants' data were excluded. The EPA Infrequency and Validity scales were then examined for the 572 remaining participants. Both scales showed relatively low levels of item endorsement (Infrequency *M* = .75, *SD* = 1.28, Range = 0 to 6; Validity *M* = .98, *SD* = 1.41, Range = 0 to 7). Based on the preregistered exclusion criteria, 15 participants' data were excluded based on Infrequency scale scores, and 18 participants' data excluded based on Virtue scale scores. Next, reaction time data was examined in order to exclude any additional participant data that showed an excessive number of mistakes (i.e., mistakes on 20% or more of total trials on the Posner cueing task). A total of 22 participants' data were excluded based on this criterion leaving a final sample of *N* = 517, slightly above the planned sample size of 500 participants. Due to a coding error, demographic information was not available for 53 participants. In regard to the remaining 464 participants, the average age of the participants was 37.52 years old (*SD* = 11.55) and participants were predominantly Euro American (78%). The sample was 57% female, and 38% of the participants reported having a bachelor's degree.

Data Preparation

In accordance with the study preregistration, all reaction time data were log-transformed for primary analyses. The spatial attention task reaction time data were also windsorized, in line with the original study (Wilkowski et al., 2006) and the preregistered analytical plan. In regard to reaction time outliers for the picture viewing task, a total of 333 trials (less than 0.1% of the total trials) were excluded from analyses due to being faster than 300 ms. Additional picture viewing trial data that was excluded from analyses were trials with reaction times longer than 20,000 milliseconds. The excessively long reaction times suggested that the participant was not paying attention during the particular trial. A total of 155 trials were removed for being longer than 20,000 milliseconds. In regard to the Posner cueing task, trials where participants made a mistake categorizing the cue were also excluded (3,651 trials; less than 5% of total trials), given that a 4,000 ms error message was displayed which would interfere with any cueing effect. Because the latter two trial exclusion criteria were not preregistered, all reaction time based results were conducted with and without the excluded trials. Any differences or lack thereof are noted in the reported results. All agreeableness scales (IPIP-NEO-Agreeableness, Big Five Markers-Agreeableness, all agreeableness facet scales) were grand-mean centered for analyses.

Primary Confirmatory Analyses

A summary of all hypotheses, both confirmatory and exploratory, can be found in Table 2. In addition, Table 2 contains summaries of both the statistical tests used to test each hypothesis and the results for all focal hypotheses. Descriptive statistics for the agreeableness scales are presented in Table 3.

Hypothesis 1: Agreeableness and Picture Viewing

To test hypothesis 1, we conducted a multilevel regression with the maximum number of random effects. The random effects included random intercepts for subject, IAPS picture, and valence along with random slopes for the within-subject effect of valence.[6] This model failed to converge, so in line with the preregistered steps to address model convergence issues, the correlation between the random slopes for picture valence was constrained to zero. The updated model failed to converge, so the random slopes parameter for picture valence was removed. Model convergence was still not achieved so a model was estimated only with two random intercepts: one for participant and one for IAPS picture. This model converged and was used to test Hypothesis 2. The results showed that there was no significant interaction between picture valence (0 = positive; 1 = negative) and IPIP-Agreeableness (B = -.001; 95% CI = -.004 to .001; *t* = -.42; *p* = .68), contrary to Hypothesis 1. There was a main effect of IPIP-Agreeableness (B = .02; 95% CI = .01 to .03; *t* = 2.84; *p* = .005) such that as individuals reported higher levels of agreeableness, they viewed positively valenced pictures for longer periods of time. The main effect of agreeableness was not consistent with Hypothesis 1a, which predicted that individuals high in agreeableness would not show a preference for negatively valenced pictures. Additionally, Hypothesis 1a was predicated on the interaction between agreeableness and picture valence. There was also a main effect of picture valence (B = .06; 95% CI = .01 to .12; *t* = 2.40; *p* = .02), and was consistent with Hypothesis 1b such that individuals viewed negatively valenced pictures longer than positively valenced pictures. The results were nearly identical when including trials where reaction times exceeded 20,000 milliseconds.

[6] A random slope for the within-person effect of picture could not be included because once trials were excluded based on outlier exclusion rules, there were not enough observations in the data to estimate the parameter.

Hypothesis 2: Agreeableness and Spatial Attention

Similar to the picture viewing task, the maximum number of random effects were initially specified for the spatial attention task. This included random intercepts for subject, word (i.e., hit, kick, hug, etc.), cue type (prosocial vs. antisocial), and cue/target location (same vs. different). In addition, random slopes were included for the within-subject effects of word, cue type, and location. The maximally specified model failed to converge, so the correlations among the random slopes were first constrained to zero. The constraint allowed for model convergence, but also resulted in a singular fit. Inspection of results showed that nearly all random effects were estimated to be zero or approximately zero. In line with preregistered steps to deal with model convergence issues, the random effects parameters were dropped with the exception of the random slope for participant. The updated model showed no convergence issues and was used for testing the spatial attention hypotheses.

The results showed no support for Hypothesis 2, as the three-way interaction between agreeableness, cue type (0 = prosocial; 1 = antisocial), and cue/target location (0 = same; 1 = different) was approximately zero (B = .000; 95% CI = -.001 to .002; *t* = .26, *p* = .79). There were also no significant two-way interactions. There were two main effects that were significant. The main effect of cue/target location (B = .07; 95% CI = .06 to .08; *t* = 26.52; *p* < .001) showed that participants had slower reaction times when the cue and target were presented in different locations compared to when they were presented in the same location (i.e., the Posner cueing effect). The other main effect was observed for agreeableness (B = -.007; 95% CI = -.012 to -.001; *t* = -2.21; *p* = .03) such that as individuals reported higher levels of agreeableness, their reaction times were faster on trials where a prosocial cue was shown at the same location as the target.

When the analyses were rerun while not excluding trials where participants incorrectly identified the cue (3,651 trials), the results were essentially identical with only the two main effects of agreeableness and cue/target location being significant. Both effects were of similar magnitudes compared to when the trials were excluded.

Exploratory Analyses

Hypothesis 3: Agreeableness Facets and Picture Viewing

Although the confirmatory results showed no significant interaction between agreeableness and picture valence, the preregistered analytical plan was to expand the analyses to include the empirically identified facets of agreeableness (A1: Trust; A2: Morality; A3: Altruism; A4: Cooperativeness; A5: Modesty) regardless of findings at the confirmatory stage. Thus, using the same random effects specified in the model used to test Hypothesis 1 (i.e., random intercepts for subject and picture), Hypothesis 3 was tested. The results show no significant interactions between picture valence and agreeableness facets (*t* range = -1.91 to 1.91; *p* range = .06 to .75). Two main effects were observed. One main effect was observed for the Morality facet (B = .17; 95% CI = .09 to .24; *t* = 4.29; *p* < .001) such that individuals who reported higher levels of Morality spent a longer amount of time viewing positively valenced pictures compared to negatively valenced pictures. The other main effect was observed for picture valence, such that individuals viewed negative pictures longer than positive pictures (B = .06; 95% CI = .01 to .12; *t* = 2.40; *p* = .018). Similar to Hypothesis 1, the data did not support the preregistered Hypothesis 3 predictions in regard to the interaction between picture valence and agreeableness. When the analyses were rerun with reaction times greater than 20,000 milliseconds included (155 trials), the results did not change.

Hypothesis 4: Agreeableness Facets and Spatial Attention

Facet-level analyses were also conducted for the spatial attention task. The same model used for the domain-level analysis was used for the facet-level analysis (only 1 random slope included for Subject). No three-way interaction between the agreeableness facets, cue type, and location were observed (*t* range = -1.23 to 1.36; *p* range = .17 to .88). One significant two-way interaction was observed for the Cue x Modesty interaction (B = -.008; 95% CI = -.016 to -.000; *t* = -2.05; *p* = .04) such that individuals who reported higher levels of Modesty were quicker to respond to targets after an antisocial cue was presented at the same location as the target. There were two significant main effects. The first was a main effect for Compliance (B = -.04; 95% CI = -.07 to -.01; *t* = -2.83; *p* = .005), with individuals who reported higher levels of compliance being faster to respond to prosocial cues presented at the same location as the target. The second main effect was observed for location (B = .07; *t* = 29.76; 95% CI = .06 to .07; *p* < .001) and like the main effect of location observed in the domain-level analysis, the main effect showed that participants were slower to respond to the target when the cue and target were presented at different locations. When analyses were rerun while including trials where mistakes were made identifying the cue (3,651 trials), the results were largely the same with the exception that there was also a significant main effect of Modesty (B = -.034; 95% CI = -.065 to -.003; *t* = -2.16; *p* = .03) but this main effect was qualified by the two-way interaction observed for Modesty and cue type which was slightly larger when the additional trials were included (B = -.009; 95% CI = -.017 to -.001; *t* = -2.29; *p* = .02). Thus, the data did not provide support for Hypothesis 4 nor Hypothesis 4a.

Hypothesis 5: Spatial Attention and the CRT-A

Despite the lack of replicating the hypothesized three-way interaction in the spatial attention task, the zero-order correlation between the Conditional Reasoning Test of Aggression (CRT-A) and a disengagement cost score was still calculated. The disengagement cost score was computed for each participant based on their average reaction time on trials where an antisocial cue was presented at the opposite location of the target minus the average reaction time on trials where an antisocial cue was presented at the same location as the target. The zero-order correlation between disengagement cost and the CRT-A total score was small, negative, and non-significant (*r* = -.04; 95% CI = -.12 to .05). Because there was no correlation between disengagement cost and the CRT-A, the test for Hypothesis 5a was not conducted. However, there was a significant relationship between IPIP-Agreeableness and the CRT-A (*r* = -.18; 95% CI = -.09 to -.26).

Hypothesis 6: Spatial Attention and Antisocial Behavior

The correlations among the disengagement cost (both antisocial and prosocial disengagement cost, with the prosocial disengagement variable being computed in the same manner as the antisocial disengagement variable) variables and the measures of self-reported antisocial behavior (the Reactive-Proactive Aggression Questionnaire, the CRT-A, and the Crime and Analogous Behaviors Scale) were also tested. The results are presented in Table 4, and show that there was a positive manifold among the self-reported antisocial behavior scales (*r* range = .10 to .83) but the CRT-A was unrelated to the self-report measures of antisocial behavior. Additionally, none of the self-report measures nor the CRT-A showed significant relations with the disengagement cost variables with the exception of a positive correlation between disengagement costs on prosocial trials and CAB-Gambling (*r* = .10; 95% CI =.02 to .19) . However, the disengagement cost variables were significantly correlated with one another (*r* = .43; 95% CI = .36 to .50). Due to the lack of relations between the disengagement cost variables and the self-reported antisocial variables, the preregistered Steiger tests of dependent correlations were not conducted. In sum, the correlational results did not provide support for Hypothesis 6.

Adjusting for Type I Error

In the preregistered analytical plan, the adjustment for the large number of significance tests among the exploratory analyses was based on the false discovery rate (FDR; Benjamini et al., 2006) adjustment. Furthermore, the adjustment was planned for a total of 37 tests (10 interaction terms in the multilevel models; 27 correlation tests). However, because main effects were interpreted in all the multilevel models, while the tests of dependent correlations were not conducted, the total number of tests conducted was 50 (all main effects and interactions in multilevel models = 35 tests; all correlations = 15 tests). Based on the FDR adjustment, with a false discovery rate of 5%, three effects did not meet the modified *p* value cutoff to be considered statistically significant: the correlation between prosocial disengagement cost and CAB-Gambling, the two-way interaction between Cue type and Modesty that was observed in the picture viewing multilevel model, and the main effect of picture valence observed in the picture viewing facet model. Notably, the FDR is a less conservative correction than the Bonferroni correction, and with 50 tests and a FDR set at 5%, approximately 2 tests will be false positives. Given that there were only 5 tests that remained significant, a more conservative FDR rate (e.g., 1%) may be justified. When the FDR is set at 1% (i.e., out of 50 tests, approximately one false positive may arise), the main effect of Compliance in the spatial attention task no longer was considered statistically significant. After applying the FDR adjustment, four out of the 50 effects from the exploratory analyses remained significant: the main effect of location in the spatial attention task (i.e., the Posner cueing effect), the main effect of Morality in the picture viewing task, the correlation between prosocial and antisocial disengagement costs, and the correlation between IPIP-Agreeableness and the total score of the CRT-A.

Supplementary Analyses

Although all confirmatory analyses aimed to replicate previously published results using best practices in regard to modeling random effects in multilevel models, we conducted supplementary analyses in order to exactly replicate the picture viewing and spatial attention tasks. These analyses were conducted using the exact same model specification as the original studies (only including a random intercept for Subject) and also made use of the same agreeableness scale: the 10-item Agreeableness scale from Goldberg’s Big Five Markers.

Picture Viewing

The results of the direct replication of the picture viewing task showed that the interaction between picture valence (positive = 0; negative = 1) and BFM-Agreeableness to be approximately zero (B = -.005; 95% CI = -.017 to .007; *t* = -.76; *p* = .45). There was no main effect for BFM-Agreeableness, but there was a main effect of picture valence, such that at average levels of BFM-Agreeableness, participants tended to view negatively valenced pictures longer (B = .06; 95% CI = .05 to .07; *t* = 13.73; *p* < .001). Thus, the main effect of picture valence found in the original study was replicated, but not the focal hypothesis regarding the interaction between picture valence and agreeableness. These results were nearly identical when trial data with reaction times longer than 20,000 ms were included in the analyses.

Spatial Attention

The results of the spatial attention task direct replication showed no significant two-way interactions, nor the hypothesized three-way interaction between BFM-Agreeableness, Cue/Target Location, and Cue Type (B = .002; *t* = .54; 95% CI = -.006 to .010; *p* = .59). There was a significant main effect of location (i.e., the Posner cueing effect), such that participants were slower to identify the target when the cue appeared at the opposite location of the target (B = .07; 95% CI = .06 to .07; *t* = 29.77; *p* < .001). There was a small but significant two-way interaction between cue type (0 = Prosocial; 1 = Antisocial) and agreeableness (B = -.006; 95% CI = -.012 to -.000; *t* = -2.00; *p* = .045), such that as participants' agreeableness increased, the reaction time to the target after an antisocial cue was presented decreased. The nature of the two-way interaction was not consistent with the effect of agreeableness observed in the original study.

When including trials where participants incorrectly identified the cue (3,651 trials), the main effect of location was slightly smaller (B = .06; 95% CI = .06 to .07; *t* = 27.32; *p* < .001) and the two-way interaction between cue type and agreeableness was not significant. All other results remained essentially unchanged.

DISCUSSION

The goal of the present study was to merge typically disconnected domains of personality research in order to better understand agreeableness-related processes. Focusing on two previously published studies that examined two purported processes of agreeableness (situation selection and spatial attention), we aimed to replicate and extend previous findings. As Table 2 highlights, with the exception of the main effect of picture valence, no evidential support was found for the focal hypotheses investigated in the current study, despite having notably high power (>94%) to detect effects that the original studies would have had a 33% chance of detecting. Based on post-hoc sensitivity analyses where the interaction effects observed for the tests of Hypotheses 1 and 2 in the current study were substituted for the interaction effects observed in the original studies, Bresin and Robinson (2015) would have had a 5% chance of detecting the two-way interaction between agreeableness and picture valence. Wilkowski et al. (2006) would have had a 4.6% chance of detecting the three-way interaction between cue type, location, and agreeableness.

One could also substitute the lower bound of the confidence intervals on the interaction effects, given that the original effects both had a negative sign. When substituting the lower bounds of the confidence intervals (-.006 for the picture viewing interaction; -.002 for the spatial attention interaction), the original studies would have had a 7.8% (picture viewing) and 7.5% chance (spatial attention) of detecting interactions of such magnitudes. These results highlight that the interaction effects examined in the current study are close enough to zero that the original studies were unable to meaningfully study them. In turn, the original evidence of these interactions does not appear to be informative.

It is also worth noting that both paradigms used in the current study had been supported by conceptual replications in the original publications. Our results are in line with other work that has found that conceptual replications do little to increase the chances of independent, direct replication effects that are consistent with the original findings (Kunert, 2016). In sum, the results suggest that under rigorous testing scenarios, the indicators utilized in previous studies on agreeableness-related processes are not robust measures of such processes.

Agreeableness and Situation Selection

We were unable to replicate the effect of primary interest in the picture viewing task (i.e., the interaction between agreeableness and picture valence) but main effects of agreeableness and picture valence were observed. In regard to the main effect of agreeableness, it was in the expected direction given the previous results as the main effect showed that as participants reported higher levels of agreeableness, they viewed all pictures (positive and negative) for longer periods of time. The main effect of picture valence, however, was consistent with the original effect such that participants viewed negatively valenced pictures longer than positively valenced pictures. The main effect of picture valence was observed in analyses where the maximum number of random effects were included (i.e., random slopes for participant and picture) and IPIP-Agreeableness was used to assess agreeableness, as well as in our supplemental analysis using the same multilevel model and measure of agreeableness as the original publication. However, it is worth noting that there is greater variability surrounding the point estimate in the multilevel model including a random intercept for picture, compared to the multilevel model where this random effect is not included.

As previously noted, there are various reasons to include the maximum number of random effects parameters (e.g., Judd, Westfall, & Kenny, 2012; Westfall, Kenny, & Judd, 2014), and the inclusion of the maximum number of random effects is recommended for confirmatory testing (Barr et al., 2013). Specific to the stimuli used in the present (i.e., the 100 pictures from the IAPS), not including random effects for IAPS pictures ignores two important features of the IAPS stimuli. First, it ignores that the 100 pictures are drawn from a larger population of pictures. Second, it ignores that different pictures may have different effects on viewing times. The former feature of the IAPS pictures is clearly true (1,196 pictures make up the collection of IAPS pictures; Lang et al. 2008), and it seems plausible that different pictures will have different effects on viewing times (e.g., there are differences in arousal ratings across the pictures). Without accounting for these sources of variance in the data leads to Type I error rates well above the standard threshold of .05 (Judd et al., 2012). Relatedly, if one is interested solely in the fixed effects of a multilevel model, it is important to establish that the fixed effects are robust to other systematic sources of variance in the data. Ultimately, the more conservative estimate of the main effect of picture valence is likely to be more realistic, compared to the main effect observed in the exact replication result where only a random intercept for participant was included.

Agreeableness and Spatial Attention

The spatial attention task results showed that while the current study replicated the Posner cueing effect found in the original study (i.e., participants took longer to react to the target on trials when the cue had been presented at an opposite location than the target), no other effects from the original study replicated. There was a main effect of agreeableness, such that individuals who reported higher levels of agreeableness had faster reaction times on trials where a prosocial cue was shown at the same location as the target. The nature of the effect was not in line with any hypotheses nor the results from the original study and thus it is questionable if it is worth significant attention.

However, two considerations are worth mentioning concerning the inability to replicate the initial findings of Wilkowski et al. (2006). First, a straightforward explanation is that the original research was conducted before the field had begun to more seriously address concerns over statistical power and the various issues regarding small samples which can lead to overestimates of effects and/or false positives (Nelson, Simmons, & Simonsohn, 2018). Alternatively, recent work has highlighted that leveraging robust, cognitive paradigms to function as measures of reliable individual difference processes may be misguided. Hedge, Powell, and Sumner (2018) highlight that historically, two distinct approaches to psychological research have been implemented—experimental and correlational approaches. Experimental research seeks to decrease heterogeneity in the between-participant effect of interest. In other words, the Posner cueing paradigm, like other paradigms from the experimental tradition, is designed to minimize individual differences in order to produce robust, experimental effects that consistently replicate across studies. This can be contrasted with the individual differences approach to research (i.e., the correlational tradition), where there must be reliable between-participant variability on some trait/construct so that measures of individual differences can reliably rank-order individuals relative to one another. Thus, given the goals of many tasks derived from experimental research (minimize between-subject variance), they are likely suboptimal tools for discovering robust individual differences related to personality traits. Ultimately, what likely makes the Posner effect robust across studies is its tendency not to be influenced by participant characteristics like personality traits.

Despite the lack of replication of previous results regarding processes of agreeableness, the goal of process-based approaches to studying agreeableness remains important. Specifically, (low) agreeableness has shown to be the strongest correlate of various antisocial behaviors (Vize et al., 2018) but much of this research, predominantly composed of data from self-reported traits and outcomes, offers little in way of explaining *how* agreeableness comes to be such a robust correlate of antisocial behavior. Recent calls to integrate research on personality structure and personality processes (e.g., Baumert et al., 2018) provide compelling reasons to move towards integration of these research areas in order to explain behavior and advance personality science. A variety of interesting questions have yet to be explored regarding processes of agreeableness such as whether processes organize themselves in the same way the traits are organized at the population level (e.g., are there specific processes that underlie the Morality factor compared to the Humility factor?), and how such processes may lead to the emergence of agreeableness-related traits when studied over the course of development. Thus, an important first step is to ensure that there are robust measures that can identify specific processes that correspond to self-reported agreeableness.

Future Directions and Recommendations

Though the current study was unable to replicate the previous effects found in Bresin and Robinson (2015) and Wilkowski et al. (2006), the study focused on two paradigms within a research area where a variety of paradigms have been employed. Furthermore, the paradigms employed in the current study take a particular approach to operationalizing personality "processes." The paradigms used in the current study focus on cognitive-emotional processes assessed by reaction times on behavioral tasks, which were drawn from a larger research area that is concerned with developing robust cognitive-processing models of personality (e.g., Wilkowski & Robinson, 2010; Robinson, 2007; 2010).

In regard to cognitive-emotional processes and the paradigms used to operationalize them, the current results suggest that further validation steps are likely necessary to ensure that personality processes can be reliably assessed. In the current context, the use of the term reliable encompasses both how the term is typically used in experimental research (i.e., to describe an effect that consistently replicates across studies with relatively consistent effect sizes) and how the term is used in individual differences research (i.e., the measure consistently rank-orders individuals). This measurement-based issue is essential in order to integrate structure and process-based research on personality (Baumert et al., 2018).

Additionally, improved theoretical development of cognitive-emotional process models will allow for clearer benchmarks in regard to the effects that process-based measures ought to be able to detect in order to be considered reliable in the sense defined above. Motivated by the lack of replicability of psychological findings, researchers have called for greater consideration of what effect sizes one would care about detecting (or the smallest effect size of interest), and to power one's study accordingly (Morey & Lakens, 2016; Lakens, Scheel, & Isager, 2018). For example, in attempting to identify the smallest effect size that was clinically relevant for patients undergoing treatment for major depressive disorder symptoms, Cuijpers et al. (2014) identified an effect size of d = .24 as corresponding to the smallest effect size that was still deemed clinically relevant. It is not clear if the smallest effect of interest is clearly identified by existing theories on cognitive-emotional processes of personality. For example, what degree of millisecond difference is large enough to trigger more downstream cognitive and emotional processes that in turn increase the propensity to act aggressively in a given scenario? An answer to this question requires substantial elucidation of how statistical observations (e.g., an effect size or statistical significance) map onto meaningful external criteria. It may be the case that more exploratory work be done in this area. We argue that incorporating the rapidly evolving methods of open science can help strengthen how robust such exploratory work is and also clearly delineate between exploratory and confirmatory work (Frankenhuis & Nettle, 2018).

Last, when considering agreeableness-related processes more specifically, it is likely that if processes that are uniquely tied to downstream, agreeable behaviors can be reliably assessed, they will be present in interpersonal situations. As noted in the introduction, of the five domains of the Five-factor Model, agreeableness has the most robust empirical and theoretical relations to behaviors, thoughts, and feelings rooted in interpersonal contexts (Graziano & Tobin, 2017). In turn, another direction future research can take is to study agreeableness-related processes at a more macro level (i.e., in interpersonal contexts using momentary sampling techniques). Such research may in turn inform which measures of cognitive-emotional are at play in interpersonal situations.

In a recent example of such work, Sun and Vazire (2019) had undergraduate participants wear electronically activated recording devices that captured short audio recordings of participants' behaviors throughout the day over the course of 15 days. Their results showed that while there was agreement between participants and coders in how extraverted and conscientious participants acted, there was little agreement observed for agreeableness suggesting that participants had relatively poor insight into how agreeable or disagreeable they were acting in the moment. Other examples can be drawn from research on interpersonal theory, where affiliation can be understood to be a marker of agreeableness-related traits (McCrae & Costa, 1998). Using a sample of married couples, Demody et al. (2017) investigated patterns of complementarity, which describes the process by which as one individual acts more affiliatively, the other does as well and as one individual acts more dominantly, the other acts more submissively. Using time-varying models, Demody and colleauges found that patterns of complementarity changed over time within the dyads as they discussed their favorite aspects of their relationship. The results showed that there were no changes in average affiliation for wives and husbands over time, but on average, the complementarity of affiliative behaviors increased over the course of their observed interactions.

Both studies provide examples of how agreeableness-related processes may be studied in such a way to help inform other areas of process-based research. Importantly, these more macro-focused processes are firmly grounded in interpersonal contexts, which is likely a key feature of agreeableness-related processes. Ultimately, process-based accounts of agreeableness will be essential in order to explicate *how* agreeableness comes to be related to a wide range of interpersonal outcomes (e.g., antisocial behaviors). Research within psychological science continues to work towards heeding the growing call to build a more robust research literature (Nosek et al., 2018). Process-based personality research, like other research fields in psychology, can leverage such improvements, such as preregistration, in order to effectively advance the empirical and theoretical integration of process and structure-based personality research.

LIST OF REFERENCES

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: American Psychiatric Association.

Axelrod, R. (1984). *The evolution of cooperation*. New York, NY: Basic Books, Inc.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255-278.

Bates, D., Machler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1-48.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *ArXiv150604967 Stat*. Available online at: http://arxiv.org/abs/1506.04967

Baumert, A., Schmitt, M., Perugini, M., Johnson, W., Blum, G., Borkenau, P., . . . Wrzus, C. (2017). Integrating personality structure, personality process, and personality development. *European Journal of Personality*, *31*, 503-528.

Benjamini, Y., Krieger, A. M., Yekutieli, D., & Krieger, A. M. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*, 491-507.

Bresin, K., & Robinson, M. D. (2015). You are what you see and choose: Agreeableness and situation selection. *Journal of Personality*, *83*, 452-463.

Brown, S. L., & Brown, R. M. (2006). Selective Investment Theory: Recasting the functional significance of close relationships. *Psychological Inquiry*, *17*, 1-29.

Costa, P. T., & McCrae, R. R. (1992a). Four ways five factors are basic. *Personality and Individual Differences, 13*, 653-656.

Costa, P. T., & McCrae, R. R. (1992b). *Professional manual: Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI)*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., McCrae, R., & Dye, D. A. (1991). Domains and facets scales for agreeableness and conscientiousness: A revision of the NEO personality inventory. *Journal of Personality Assessment*, *12*, 887-898.

Crick, N. R., & Dodge, K. A. (1996). Social information-processing mechanisms in reactive and proactive aggression. *Child Development, 67*, 993-1002.

Crowe, M. L., Lynam, D. R., & Miller, J. D. (2018). Uncovering the structure of agreeableness from self-report measures. *Journal of Personality, 86,* 771-787.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon’s Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, *8*.

Cuijpers, P., Turner, E. H., Koole, S. L., van Dijke, A., & Smit, F. (2014). What is the threshold for a clinically relevant effect? The case of major depressive disorders. *Depression and Anxiety*, *378*, 374-378. https://doi.org/10.1002/da.22249

De Pauw, S. S. W., & Mervielde, I. (2010). Temperament, personality and developmental psychopathology: A review based on the conceptual dimensions underlying childhood traits. *Child Psychiatry and Human Development*, *41,* 313-329.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology, 93,* 880-896.

Eckhardt, C. I., & Cohen, D. J. (1997). Attention to anger-relevant and irrevlevant stimuli following naturalistic insult. *Personality and Individual Differences*, *23,* 619-629.

Evans, T. D., Cullen, F. T., Dunaway, R. G., & Benson, M. L. (1997). The social consequences of self-control: Testing the general theory of crime. *Criminology*, *35*, 475-504.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160.

Finkel, E. J. (2007). Impelling and inhibiting forces in the perpetration of intimate partner violence. *Review of General Psychology, 11*, 193-207.

Frankenhuis, W. E., & Nettle, D. (2018). Open science is liberating and can foster creativity. *Perspectives on Psychological Science*, *13*, 439-447.

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19,* 847-857.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*, 26-42.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *The American Psychologist*, *48*, 26-34.

Goldberg, L. R. (2006). Doing it all bass-ackwards: The development of hierarchical factor structures from the top down. *Journal of Research in Personality*, *40,* 347-358.

Graziano, W. G., & Eisenberg, N. (1997). Agreeableness: A dimension of personality. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 795-824). San Deigo, CA: Academic Press.

Graziano, W. G., & Habashi, M. M. (2010). Motivational processes underlying both prejudice and helping. *Personality and Social Psychology Review, 3*, 313-331.

Graziano, W. G., Habashi, M. M., Sheese, B. E., & Tobin, R. M. (2007). Agreeableness, empathy, and helping: A person × situation perspective. *Journal of Personality and Social Psychology*, *93*, 583-599.

Graziano, W. G., & Tobin, R. M. (2013). The cognitive and motivational foundations underlying agreeableness. In M. D. Robinson (Ed.), *Handbook of cognition and emotion* (pp. 347-364). New York, NY: Guilford Press.

Graziano, W. G., & Tobin, M. (2017). Agreeableness and the Five-Factor Model. In T. A. Widiger (Ed.), *The Oxford handbook of the Five Factor Model* (pp. 105-132). Oxford, England: Oxford University Press.

Green, P., & Macleod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*, 493-498.

Gross, J. J. (1998). The emerging field of emotion regulation: an integrative review. *Review of General Psychology*, *2*, 271-299.

Habashi, M. M., Graziano, W. G., & Hoover, A. E. (2016). Searching for the prosocial personality: A big five approach to linking personality and prosocial behavior. *Personality and Social Psychology Bulletin*, *42*, 1177-1192.

Heaven, P. C. L., Connors, J., & Stones, C. R. (1994). Three or five personality dimensions? An analysis of natural language terms in two cultures. *Personality and Individual Differences*, *17*, 181-189.

Heck, D. W., Thielmann, I., Moshagen, M., & Hilbig, B. E. (2018). Who lies ? A large-scale reanalysis linking basic personality traits to unethical decision making. *Judgment and Decision Making*, *13*, 356-371.

Hilbig, B. E., Glöckner, A., & Zettler, I. (2014). Personality and prosocial behavior: Linking basic traits and social value orientations. *Journal of Personality and Social Psychology*, *107*, 529-539.

Hilbig, B. E., & Zettler, I. (2009). Pillars of cooperation: Honesty-Humility, social value orientations, and economic behavior. *Journal of Research in Personality*, *43*, 516-519.

Hilbig, B. E., & Zettler, I. (2015). When the cat's away, some mice will play: A basic trait account of dishonest behavior. *Journal of Research in Personality*, *57*, 72-88.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, *55*, 19-24.

Hopwood, C. J., Harrison, A. L., Amole, M., Girard, J. M., Wright, A. G., Thomas, K. M., . . . & Crowley, M. J. (2018). Properties of the continuous assessment of interpersonal dynamics across sex, level of familiarity, and interpersonal conflict. *Assessment*, Advance online publication.

James, L. R., McIntyre, M. D., Glisson, C. A., Green, P. D., Patton, T. W., LeBreton, J. M., . . . Williams, L. J. (2005). A conditional reasoning measure for aggression. *Organizational Research Methods*, *8*, 69-99.

Jensen-Campbell, L. A., Adams, R., Perry, D. G., Workman, K. A., Furdella, J. Q., & Egan, S. K. (2002). Agreeableness, extraversion, and peer relations in early adolescence: Winning friends and deflecting aggression. *Journal of Research in Personality*, *36*, 224-251.

Jensen-Campbell, L. A., & Graziano, W. G. (2001). Agreeableness as a moderator of interpersonal conflict. *Journal of Personality, 69*, 323-362.

Jensen-Campbell, L. A., Rosselli, M., Workman, K. A., Santisi, M., Rios, J. D., & Bojan, D. (2002). Agreeableness, conscientiousness, and effortful control processes. *Journal of Research in Personality, 36*, 476-489.

Jones, S. E., Miller, J. D., & Lynam, D. R. (2011). Personality, antisocial behavior, and aggression: A meta-analytic review. *Journal of Criminal Justice*, *39*, 329-337.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*, 54-69.

Klein, R. A., Vianello, M., Hasselman, F., Alper, S., Aveyland, M., Axt, J. R., . . . Nosek, B. A. (2015). *Many Labs 2: Investigating variation in replicability across sample and setting*. Manuscript in preparation.

Klimstra, T. A., Luyckx, K., Hale, W. W., & Goossens, L. (2014). Personality and externalizing behavior in the transition to young adulthood: The additive value of personality facets. *Social Psychiatry and Psychiatric Epidemiology*, *49*, 1319-1333.

Kunert, R. (2016). Internal conceptual replications do not increase independent replication success. *Psychonomic Bulletin & Review*, *23*, 1631-1638.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1997). International Affective Picture System (IAPS): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, 39-58.

Lebreton, J. M., Grimaldi, E. M., & Schoen, J. L. (2018). Conditional reasoning : A review and suggestions for future test development and validation. *Organizational Research Methods*, (Advance online publication), 1-31.

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, *39*, 329-358.

Lynam, D. R., Gaughan, E. T., Miller, J. D., Miller, D. J., Mullins-Sweatt, S., & Widiger, T. A. (2011). Assessing the basic traits associated with psychopathy: Development and validation of the Elemental Psychopathy Assessment. *Psychological Assessment*, *23*, 108-124.

Lynam, D. R., Sherman, E. D., Samuel, D., Miller, J. D., Few, L. R., & Widiger, T. A. (2013). Development of a short form of the elemental psychopathy assessment. *Assessment*, *20*, 659-669.

Maples, J. L., Guan, L., Carter, N. T., & Miller, J. D. (2014). A test of the international personality item pool representation of the revised NEO personality inventory and development of a 120-item IPIP-based measure of the five-factor model. *Psychological Assessment*, *26*, 1070-1084.

McCrae, R. R., & Costa, P. T. (1989). The structure of interpersonal traits: Wiggin’s Circumplex and the Five-Factor Model. *Journal of Personality and Social Psychology*, *56*, 586-595.

McCrae, R. R., & John, O. P. (1992). An introduction to the Five-Factor model and its applications. *Journal of Personality*, *60*, 175-215.

Meier, B. P., & Robinson, M. D. (2004). Does quick to blame mean quick to anger? The role of agreeableness in dissociating blame and anger. *Personality and Social Psychology Bulletin*, *30*, 856-867.

Meier, B. P., Robinson, M. D., & Wilkowski, B. M. (2006). Turning the other cheek: Agreeableness and the regulation of aggression-related primes. *Psychological Science*, *17*, 136-143.

Miller, J. D., Gaughan, E. T., Maples, J., & Price, J. (2011). A comparison of agreeableness scores from the Big Five Inventory and the NEO PI-R: Consequences for the study of narcissism and psychopathy. *Assessment*, *18,* 335-339.

Miller, J. D., & Lynam, D. R. (2006). Reactive and proactive aggression: Similarities and differences. *Personality and Individual Differences*, *41*, 1469-1480.

Miller, J. D., Lynam, D. R., McCain, J. L., Few, L. R., Crego, C., Widiger, T. A., & Campbell, W. K. (2016). Thinking structurally about narcissism: An examination of the Five-Factor Narcissism Inventory and its components. *Journal of Personality Disorders*, *29*, 1-18.

Miller, J. D., Lynam, D. R., Widiger, T. A., & Leukefeld, C. (2001). Personality disorders as extreme variants of common personality dimensions: can the Five-Factor Model adequately represent psychopathy? *Journal of Personality, 69,* 253-276.

Morey, R. D., & Lakens, D. (2016). *Why most of psychology is statistically unfalsifiable*. Manuscript submitted for publication.

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology, 112,* 474-490.

Norlander, B., & Eckhardt, C. (2005). Anger, hostility, and male perpetrators of intimate partner violence: A meta-analytic review. *Clinical Psychology Review*, *25*, 119-152.

Nosek, B. A., Ebersole, C. R., Dehaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115,* 2600-2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*, 137-141.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615-631.

Ode, S., Robinson, M. D., & Wilkowski, B. M. (2008). Can one’s temper be cooled ? A role for agreeableness in moderating Neuroticism ’s influence on anger and aggression. *Journal of Research in Personality*, *42*, 295-311.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 1-8.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*, 184-188.

Paunonen, S. V., & Ashton, M. C. (2001). Big Five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology*, *81*, 524-539.

Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, *109*, 160-174.

Prims, J., & Motyl, M. (2018). A tool for detecting low quality data in internet research. GitHub: https://github.com/SICLab/detecting-bots

Raine, A., Dodge, K., Loeber, R., Gatzke-Kopp, L., Lynam, D., Reynolds, C., . . . Liu, J. (2006). The reactive-proactive aggression questionnaire: Differential correlates of reactive and proactive aggression in adolescent boys. *Aggressive Behavior*, *32*, 159-171.

Reardon, K. W., Tackett, J. L., & Lynam, D. R. (2018). The personality context of relational aggression: A Five-Factor Model profile analysis. *Personality Disorders : Theory , Research , and Treatment*, *9*(3), 228-238.

Riolo, R. L., Cohen, M. D., & Axelrod, R. (2001). Evolution of cooperation without reciprocity. *Nature*, *414*, 441-443.

Roberts, B. W., Chernyshenko, O. S., Stark, S., & Goldberg, L. R. (2005). The structure of conscientiousness: An empirical investigation based on seven major personality questionnaires. *Personnel Psychology*, *58*, 103-139.

Robinson, M. D. (2007). Personality, affective processing, and self-regulation: Toward process-based views of extraversion, neuroticism, and agreeableness. *Social and Personality Psychology Compass*, *1*, 223-235.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359-1366.

Simms, E. E. (2009). *Assessment of the facets of the five factor model: Further development and validation of a new personality measure*. Unpublished dissertation.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*, 559-569.

Smith, G. T., McCarthy, D. M., & Zapolski, T. C. B. (2009). On the value of homogeneous constructs for construct validation, theory testing, and the description of psychopathology. *Psychological Assessment*, *21*, 272-284.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245-251.

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods*, *42*, 1096-1104.

Stoet, G. (2017). PsyToolkit: A vovel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, *44*, 24-31.

Sun, J., & Vazire, S. (2019). Do people know what they're like in the moment ? *Psychological Science, 30*, 405-414. https://doi.org/10.1177/0956797618818476

Trapnell, P. D., & Wiggins, J. S. (1990). Extension of the Interpersonal Adjective Scales to include the Big Five dimensions of personality. *Journal of Personality and Social Psychology*, *59*, 781-790.

Vize, C. E., Collison, K. L., Crowe, M. L., Campbell, W. K., Miller, J. D., & Lynam, D. R. (2017). Using dominance analysis to decompose narcissism and its relation to aggression and externalizing outcomes. *Assessment, 26,* 260-270.

Vize, C. E., Collison, K. L., & Lynam, D. R. (in press). The importance of antagonism: Explaining similarities and differences in psychopathy and narcissism's relations with aggression and externalizing outcomes. *Journal of Personality Disorders.*

Vize, C. E., Collison, K. L., Miller, J. D., & Lynam, D. R. (2019). Using Bayesian methods to update and expand the meta-analytic evidence of the Five-factor Model's relation to antisocial behavior. *Clinical Psychology Review, 67,* 61-77.

Vize, C. E., Miller, J. D., Collison, K. L., & Lynam, D. R. (in press). Untangling the relation between narcissistic traits and behavioral aggression following provocation using an FFM framework. *Journal of Personality Disorders.*

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

Watson, D., Stasik, S. M., Ellickson-Larew, S., & Stanton, K. (2015). Extraversion and psychopathology: A facet-level analysis. *Journal of Abnormal Psychology*, *124,* 432-446.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143,* 2020-2045.

Wilkowski, B. M., & Robinson, M. D. (2007). The cognitive basis of trait anger and reactive Aggression: An integrative analysis. *Personality and Social Psychology Review*, *12*, 3-21.

Wilkowski, B. M., & Robinson, M. D. (2010). The anatomy of anger: An integrative cognitive model of trait anger and reactive aggression. *Journal of Personality*, *78,* 9-38.

Wilkowski, B. M., Robinson, M. D., & Meier, B. P. (2006). Agreeableness and the prolonged spatial processing of antisocial and prosocial information. *Journal of Research in Personality*, *40*, 1152-1168.

Wright, A. G. C., Pincus, A. L., Hopwood, C. J., Thomas, K. M., Markon, K. E., & Krueger, R. F. (2012). An interpersonal analysis of pathological personality traits in DSM-5. *Assessment*, *19*, 263-275.

Zandt, V. (2008). Effective analysis of reaction time data. *The Psychological Record, 58,* 475-482.

APPENDIX A

Table 1

Complete Power Results for Multilevel Models

Paradigm	Sample Size	Coefficient	*SE*	Estimated Power (95 % CI)
Spatial Attention				
Published Data	66	-.01	.005	.60 (.57 to .63)
Decreased ES	66	-.007	—	.33 (.30 to .36)
Decreased ES/Increased *N*	500	-.007	—	.99 (.98 to .99)
Picture Viewing				
Published Data	122	-.048	.02	.48 (.45 to .52)
Decreased ES	122	-.037	—	.33 (.30 to .36)
Decreased ES/Increased *N*	500	-.037	—	.94 (.92 to .96)

Note: The dependent variable for the spatial attention and picture viewing tasks is log-transformed reaction time; Decreased ES = effect size that the original studies would have had 33% power to detect; power estimates are based on 1000 simulations; all estimates are the effect of most interest within each multi-level model. For the spatial attention task, the estimate refers to the agreeableness x cue type x location effect. For the picture viewing task, the estimate refers to the agreeableness x valence effect.

Table 2

Hypotheses, Planned Analyses, and Results for Primary and Exploratory Research Questions

Primary Analyses		
	How Hypothesis Will be Tested	Result
Hypothesis 1: Agreeableness and Picture Viewing There will be a significant, two-way interaction between Agreeableness and IAPS picture valence. Based on our coding of picture valence, we expect the sign of the interaction to be negative.	MLM with maximum number of random effects included in the model. Picture valence will be entered as a level 1 predictor, with grand-mean centered agreeableness entered as a level 2 predictor. A two-way, cross-level interaction term (agreeableness x valence) will be entered to test Hypothesis 1.	Contrary to Hypothesis 1, the results showed that there was no significant interaction between picture valence (0 = positive; 1 = negative) and IPIP-Agreeableness (B = -.001; 95% CI = .006 to .00; *t* = -.42; *p* = .68).
Hypothesis 1a: Individuals low in Agreeableness will show a preference for negatively valenced pictures, viewing them for a significantly longer time. However, the effect of valence will decrease at higher levels of agreeableness.	Simple slope analysis examining the effect of picture valence on viewing time for individuals low (-1 *SD*) and high (+1 *SD*) in agreeableness.	Simple slope analysis not conducted because there was no significant interaction between agreeableness and picture valence. Despite the lack of significant interaction, there was a main effect of IPIP-Agreeableness (B = .02; 95% CI = .01 to .03; *t* = 2.84; *p* = .005) such that as individuals reported higher levels of agreeableness, they viewed positively valenced pictures for longer periods of time.

(*table continues*)

Primary Analyses		
Hypothesis 1b: There will be a main effect of picture valence, such that individuals will view negatively valenced pictured longer than positively valenced pictures.	The main effect of picture valence will be tested using the same MLM specified for Hypothesis 1.	There was a main effect of picture valence (B = .06; 95% CI = .01 to .12; *t* = 2.40; *p* = .02), contrary to Hypothesis 1b.
<u>Hypothesis 2: Agreeableness and Spatial Attention</u> There will be a significant three-way interaction Between cue type, cue/target location, and Agreeableness	MLM with maximum number of random effects included in the model. Cue type (prosocial/antisocial) and cue/target location (same/different) will be entered as level 1 predictors; grand-mean centered agreeableness entered as a level 2 predictor. Three-way, cross-level interaction term (agreeableness x cue x location) included to test Hypothesis 2.	The results showed no support for Hypothesis 2, as the three-way interaction between agreeableness, cue type (0 = prosocial; 1 = antisocial), and cue/target location (0 = same; 1 = different) was approximately zero (B = .00; 9% CI = -.001 to .002; *t* = .26; *p* = .79).
Hypothesis 2a: Individuals low in Agreeableness will show higher disengagement costs when presented with antisocial cue words compared to when they are presented with prosocial cue words.	Simple slope analyses specifically examining the effect of cue word when there is a mismatch between cue/target location for individuals high (+1 *SD*) and low (-1 *SD*) in agreeableness.	Simple slope analyses are not conducted due to no significant interaction.

(*table continues*)

Exploratory Analyses		
Hypothesis 3: Agreeableness Facets and Picture Viewing We do not have an *a priori* hypothesis for which facets of Agreeableness will be differentially related to viewing positive versus negatively valenced pictures. However, should an effect be observed, we expect that the nature of the relationship will be the same as that predicted for the broader domain: those who score low on the Agreeableness facet(s) will show a preference for negatively valenced pictures.	MLM with maximum number of random effects included. Picture valence will be entered as a level 1 predictor, with grand-mean centered Agreeableness facets (5 total; Altruism, Sympathy, Morality, Trust, Cooperativeness, and Modesty) entered as level 2 predictors. Two-way, cross-level interaction terms (Agreeableness facet x valence) will then be entered to test Hypothesis 3.	The results showed no significant interactions beween picture valence and any of agreeableness facets (*t* range = -1.91 to 1.91; *p* range = .06 to .75).
Hypothesis 4: Agreeableness Facets and Spatial Attention Because the spatial attention paradigm has been argued to assess biases in processing that are related to forms of aggression, we expect that the three-way interaction between the Cooperativeness x cue type x location to be significant, given the Cooperativeness facet's	MLM with maximum number of random effects included. Cue type (prosocial/antisocial) and cue/target location (same/different) will be entered as level 1predictors; grand-mean centered facets of Agreeableness (5 total; Altruism, Sympathy, Morality, Trust, Cooperativeness, and Modesty) entered as level 2 pre0--dictors. Three-way, cross-level interaction	No three-way interactions between the agreeableness facets, cue type, and location were observed (*t* range = -1.23 to 1.36; *p* range = .17 to .88).

(*table continues*)

Exploratory Analyses		
strong relation to aggression.	Terms (Agreeableness facet x cue x location) included to test Hypothesis 3.	
Hypothesis 4a: Individuals low in Cooperativeness will show higher disengagement costs when presented with an antisocial cue word compared to when they are presented with a prosocial cue word.	Simple slope analyses specifically examining the effect of cue word when there is a mismatch between cue/target location for individuals high (+1 *SD*) and low (-1 *SD*) in Cooperativeness.	Simple slope analyses not conducted due to no significant interaction effects.
Hypothesis 5: Spatial Attention and the CRT-A Given the spatial attention task’s purported relation with aggression processes, we expect the disengagement cos for trials with antisocial cue words (i.e., the difference in reaction times on trials with same vs. different cue-target location when antisocial words cues are displayed) will be significantly positively related to scores on the CRT-A.	Zero-order correlation between disengagement cost for antisocial cue trials and the total score of the CRT-A.	The zero-order correlation between disengagement cost and the CRT-A total score was small, negative, and non-significant (*r* = -.04; 95% CI = -.12 to .05).

(*table continues*)

Exploratory Analyses		
Hypothesis 5a: Should there be a positive correlation between the CRT-A, Agreeableness, and disengagement cost, the relation between CRT-A and disengagement costs for antisocial cues can be accounted for by their shared relation with Agreeableness.	Comparison of the zero-order correlation between disengagement cost and the CRT-A with the semipartial correlation between disengagement cost and the CRT-A, accounting for agreeableness’ overlap with the CRT-A.	Because there was no correlation between disengagement cost and the CRT-A, the test for hypothesis 5a was not conducted. However, there was a significant relationship between IPIP-Agreeableness and the CRT-A (*r* = -.18; 95% CI = -.09 to -.26).
Hypothesis 6: Spatial Attention and Antisocial Behavior We expect to find a positive manifold among our measures of antisocial behavior (the CRT-A, RPQ, and CAB) and the disengagement cost on antisocial cue trials. The correlations should be larger for subscales assessing aggression compared to scales measuring self-reported gambling and substance use. Disengagement costs on *prosocial* cue trials should be negatively related to measures of aggression.	Zero-order correlations between disengagement costs, the total score on the CRT-A, reactive and proactive aggression subscales from the RPQ, abd aggression, non-violent antisocial behavior, substance use, and gambling subscales from the CAB. Differences among externalizing correlations (aggression/antisocial correlations) vs. CAB substance use/gambling correlations) will be tested using Steiger’s (1980) method for dependent correlations.	There was a positive manifold among the self-reported antisocial behavior scales (*r* range = .10 to .83) but the CRT-A was unrelated to the self-report measures. None of the self-report measures nor the CRT-A showed significant relations with the disengagement cost variables. However, the disengagement cost variables were significantly correlated with one another (*r* = .43; 95% CI = .36 to .50). Tests of dependent correlations were not conducted due to lack of correlation between disengagement cost variables and externalizing behavior measures.

(*table continues*)

Note. MLM = Multilevel model; CRT-A = Conditional Reasoning Test of Aggression; RPQ = Reactive Proactive Aggression Questionnaire; CAB = Crime and Analogous Behavior Scale. Agreeableness domain scores will be assessed using Goldberg’s 10-item scale of Agreeableness which was used in the original studies to assess Agreeableness.

Table 3

Descriptive Statistics of Agreeableness Scales/Facets

	M (*SD*)	BFM-A	IPIP-A	Trust	Morality	Altruism	Cooperativeness	Modesty
BFM-A	3.93 (.75)	(.90)						
IPIP-A	3.68 (.57)	.70**	(.87)					
Trust	3.18 (.96)	.33**	.42**	(.87)				
Morality	3.84 (.83)	.41**	.75**	.05	(.72)			
Altruism	3.95 (.76)	.75**	.75**	.34**	.41**	(.75)		
Cooperativeness	4.01 (.87)	.50**	.81**	.19**	.68**	.51**	(.76)	
Modesty	4.05 (.71)	.46**	.61**	.04	.53**	.45**	.55**	(.89)

Note. Values in parentheses represent the internal consistency (α) of the self-report agreeableness scales/facets.

** = p < .01; the Modesty facet is scored from the Modesty subscale of the Faceted Inventory of the Five Factor Model (FI-FFM).

Table 4

Descriptive Statistics and Correlations of Antisocial Behavior Measures

	M (SD)	BFM-A	IPIP-A	RPQ-React	RPQ-Proact	CAB-Violent	CAB-Nonviolent	CAB-Substance	CAB-Gambling	CRT-A	Disengage-Anti	Disengage-Pro
BFM-A	3.93 (.75)	(.90)										
IPIP-A	3.68 (.57)	.70**	(.87)									
RPQ-React	1.48 (.30)	-.26**	-.43**	(.82)								
RPQ-Proact	1.53 (.30)	-.28**	-.49**	.83**	(.84)							
CAB-Violent	.57 (.73)	-.07	-.18**	.36**	.40**	(.43)						
CAB-Nonviolent	.66 (.94)	-.08	-.18**	.41**	.38**	.47**	(.62)					
CAB-Substance	1.57 (1.37)	.00	-.05	.22**	.17**	.38**	.45**	(.67)				
CAB-Gambling	2.42 (1.78)	.01	-.03	.16**	.10*	.33**	.31**	.36**	(.76)			
CRT-A	5.37 (2.26)	-.14**	-.18**	.02	.05	-.04	-.08	-.11*	-.06	(.33)		
Disengage-Anti	47.56 (72.93)	.03	.02	-.04	-.03	.03	.07	.01	.03	-.04	—	

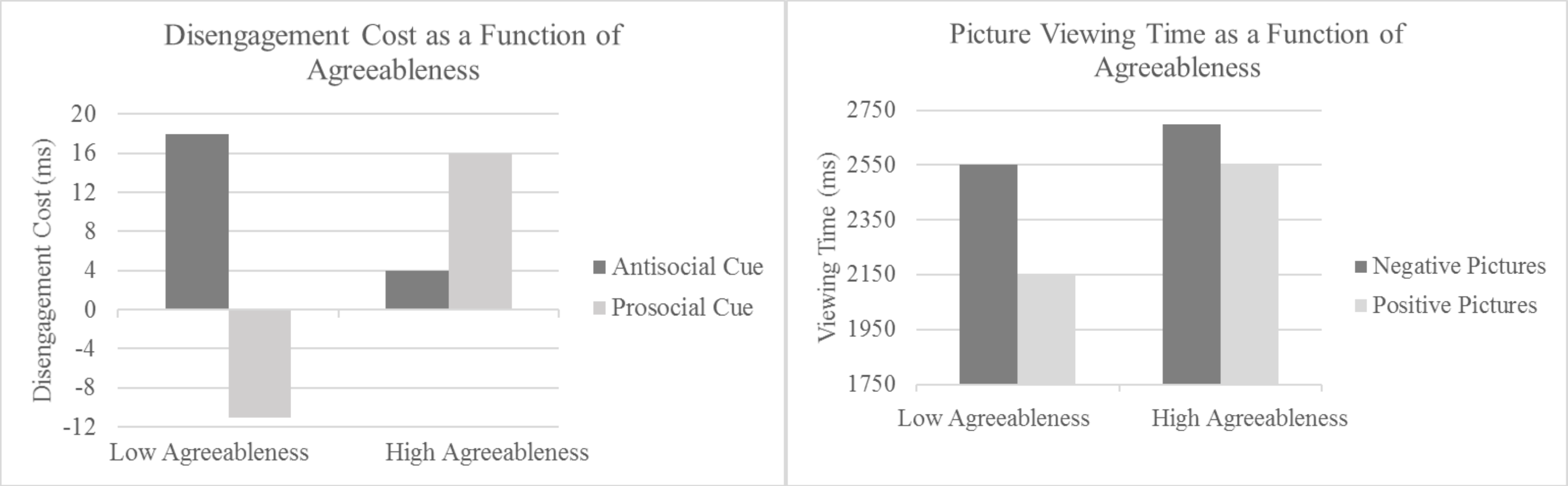
(*table continues*)

	M (SD)	BFM-A	IPIP-A	RPQ-React	RPQ-Proact	CAB-Violent	CAB-Nonviolent	CAB-Substance	CAB-Gambling	CRT-A	Disengage-Anti	Disengage-Pro
Disengage-Pro	48.33 (72.52)	.01	.00	.01	-.01	.02	.02	.00	.10*	-.06	.43**	—

Note. Values in parentheses represent the internal consistency (α) of the self-report agreeableness scales/facets. BFM-A = Big Five Markers Agreeableness Scale; IPIP-A = IPIP-NEO-120 Agreeableness Scale; RPQ-React = Reactive Proactive Aggression Questionnaire-Reactive Aggression Scale; RPQ-Proact = Reactive Proactive Aggression Questionnaire-Proactive Aggression Scale; CAB = Crime and Analogous Behaviors Scale; CRT-A = Conditional Reasoning Test of Aggression; Disengage-Anti = Reaction time difference between Posner cueing trials where an antisocial cue is show on the opposite versus same location as target; Disengage-Pro = Reaction time difference between Posner cueing trials where prosocial cue is shown on the oppose versus same location as target; the Modesty facet is scored from the Modesty subscale of the Faceted Inventory of the Five Factor Model (FI-FFM); Means and SDs of the Disengagement Cost variables are displayed in milliseconds.

* = p < .05. ** = p < .01.

APPENDIX B



Note. ms=millisecond; disengagement cost is the difference in response times on trials when the cue and target are presented on the same side (left/left) versus when the cue and target are presented on different sides (right/left)

Figure 1. Graphs depicting expected interactions for spatial attention and picture viewing tasks.

APPENDIX C

Two Sample Items from the CRT-A, as presented in James et al. (2005)

1. **American cars have gotten better in the past 15 years. American carmakers started to build better cars when they began to lose business to the Japanese. Many American buyers thought that foreign cars were better made. Which of the following is the most logical conclusion based on the above?**

   a. America was the world's largest producer of airplanes 15 years ago.
   b. Swedish carmakers lost business in America 15 years ago.
   c. The Japanese knew more than Americans about building good cars 15 years ago.
   d. American carmakers built cars to wear out 15 years ago so they could make a lot of money selling parts.

2. **The old saying, "an eye for an eye," means that if someone hurts you, then you should hurt that person back. If you are hit, then you should hit back. If someone burns your house, then you should burn that person's house. Which of the following is the biggest problem with the "eye for an eye" plan?**

   a. It tells people to "turn the other cheek."
   b. It offers no way to settle a conflict in a friendly manner.
   c. It can be used only at certain times of the year.
   d. People have to wait until they are attacked before they can strike.

Note: For both items, response "d" is indicative of an aggressive response; for item 1, response "c" is the non-aggressive response while for item 2, response "b" is the non-aggressive response; all other response options represent non-logical responses.