Journal of Research in Personality 87 (2020) 103986

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Agreeableness and the common core of dark traits are functionally different constructs

Morten Moshagen [a,*], Ingo Zettler [b], Luisa K. Horsten [c], Benjamin E. Hilbig [c]

[a] Ulm University, Germany
[b] University of Copenhagen, Denmark
[c] University of Koblenz-Landau, Germany

ARTICLE INFO

Article history:
Received 29 December 2019
Revised 15 June 2020
Accepted 16 June 2020
Available online 24 June 2020

Keywords:
Agreeableness
D Factor
Dark factor of personality
Dark traits

ABSTRACT

The Dark Factor of Personality (D) has been suggested as the basic disposition underlying dark traits, thereby representing their common core. However, it has also been argued that such commonalities reflect the low pole of Agreeableness. The present study (N = 729) employed five established inventories to model the Agreeableness construct and considered seven theoretically derived criterion variables, including one behavioral outcome. Results indicate that Agreeableness and D exhibit a substantial, but far from perfect, association of r = −.64. Further, D incrementally improved the prediction of all but one criterion measure. These results speak against the notion that the commonalities of dark traits can be reduced to low Agreeableness and rather support the contention to consider Agreeableness and D as functionally distinct constructs.

1. Introduction

The past decades witnessed an upsurge of interest in personality traits related to malevolent behaviors, as—most prominently—represented in the components of the "Dark Triad" (Paulhus & Williams, 2002) as well as many other such "dark" traits such as Greed (Seuntjens, Zeelenberg, van de Ven, & Breugelmans, 2015), Sadism (O'Meara, Davies, & Hammond, 2011), or Spitefulness (Marcus, Zeigler-Hill, Mercer, & Norris, 2014), to name just a few examples. In light of their importance for a variety of outcomes (O'Boyle, Forsyth, Banks, & McDaniel, 2012; Vize, Lynam, Collison, & Miller, 2018) and given that dark traits exhibit a substantial theoretical and empirical overlap (Muris, Merckelbach, Otgaar, & Meijer, 2017; O'Boyle, Forsyth, Banks, Story, & White, 2015), various attempts have been made to describe their commonalities (e.g., Diebels, Leary, & Chon, 2018; Jonason, Li, Webster, & Schmitt, 2009; Jones & Figueredo, 2013). Recently, Moshagen, Hilbig, and Zettler (2018) provided an integrative and extended account of the common core of dark personality by defining the basic disposition that gives rise to *all* dark traits (and thus, the underlying disposition responsible for the observed commonalities across dark traits). Their conceptualization of the Dark Factor of Personality (D) as the "general tendency to maximize one's individual utility—disregarding, accepting, or malevolently provoking disutility for others—, accompanied by beliefs that serve as justifications" (p. 657) was empirically supported by studies showing that (a) the majority of common variance pertaining to the indicators of 9–12 different dark traits was subsumed by D; (b) dark traits rarely predicted relevant outcomes beyond D; and (c) item loadings on D as well as (d) the relations of D to relevant outcome measures were in agreement with the very theoretical definition of D (Moshagen et al., 2018; Moshagen, Zettler, & Hilbig, 2020).

Despite accumulating evidence in support of the notion that much of the behaviorally relevant variance (in terms of outcomes related to malevolent behavior) of dark traits can be succinctly described through their commonalities as represented in D, it is important to note that the conceptualization of D is only one out of several suggestions concerning the theoretical definition and properties of the common core of dark traits. In particular, regarding the Dark Triad components, it has been repeatedly argued that the positive manifold of Machiavellianism, Narcissism, and Psychopathy can be understood through basic models of personality such as the Five-Factor Model (FFM; McCrae & Costa, 2008). Specifically, reiterating previous notions that "the dark dimension of personality can be described in terms of low Agreeableness" (Jakobwitz & Egan, 2006, p. 331; Paulhus & Williams, 2002; Stead & Fekken, 2014), Vize, Lynam, Collison, and Miller (in press) recently concluded that the "most parsimonious account of the core of the [Dark Triad] is that it is Antagonism

* Corresponding author at: Research Methods, Institute of Psychology and Education, Ulm University, Albert-Einstein-Allee 47, 89081 Ulm, Germany.
E-mail address: morten.moshagen@uni-ulm.de (M. Moshagen).

https://doi.org/10.1016/j.jrp.2020.103986
0092-6566/

(vs. Agreeableness) from the Five-Factor Model" (p. 22; see also Vize, Miller, & Lynam, 2019). Thus, according to this view, the commonalities of the Dark Triad components can essentially be seen as a reflection of the opposite pole of the Agreeableness dimension.

Importantly, both theorizing and empirical evidence concerning the proposition that the commonalities of dark traits essentially reflect low Agreeableness (as defined in the FFM) almost exclusively refer to the former in terms of the Dark Triad components. However, the Dark Triad components represent just a subset of all dark traits, so a natural extension is to broaden the notion previously confined to the Dark Triad and presume that low Agreeableness represents the commonalities of not only Machiavellianism, Narcissism, and Psychopathy, but actually all dark traits. If the positive manifold of all dark traits, and thus D, could indeed adequately be described as the opposite pole of an established (and fundamental) personality dimension like Agreeableness, theoretical parsimony obviously dictates to adopt this view and thus to discard other notions concerning the common core of all dark traits.

In the only study so far jointly considering Agreeableness and D, Moshagen et al. (2018; Study 3; N = 1,261) reported a latent correlation of $r = -0.69$ between Agreeableness (as assessed through the respective subscale of the NEO-FFI; Borkenau & Ostendorf, 1994; Costa & McCrae, 1992) and D (as assessed as the general factor arising in a bifactor model comprising 9 dark traits). Thus, at the very least, the overlap between Agreeableness and D is substantial.

However, the theoretical origins of Agreeableness and D fundamentally differ with Agreeableness being part of models of basic personality structure as derived from lexical studies with the aim to describe all major sets of individual differences by as few independent dimensions as possible (e.g., Saucier, 2002). In contrast, D is defined to represent the commonalities of all dark traits, so that it is possible and indeed plausible that D represents a blend of several characteristics across basic dimensions of personality. Correspondingly, empirical evidence concerning both, the Dark Triad components (Furnham, Richards, Rangel, & Jones, 2014; Muris et al., 2017; Vize et al., 2018) and indeed the common core of all dark traits (Moshagen et al., 2018) indicates substantial associations not only with Agreeableness, but other FFM dimensions as well. This is in line with the fact that the theoretical conceptualization of D also refers to features that are typically thought to reflect other FFM dimensions (in terms of the NEO-PI-R facets, for instance, warmth is part of Extraversion and hostility is part of Neuroticism; Costa, McCrae, & Dye, 1991), in turn suggesting that "dismiss[ing] the Dark Triad as simply low Agreeableness is not warranted" (Furnham et al., 2014, p. 116).

Further supporting this notion, Moshagen et al. (2018) also showed that D incrementally predicted 7 out of 11 external criteria over all five FFM dimensions (including assessments of dishonest behavior and various relevant outcomes in the domain of socially aversive patterns of behavior, i.e., aggression, dominance, impulsivity, insensitivity, self-centeredness, and power). Thus, despite Agreeableness and D sharing approximately 50% of variance, these results indicate that D comprises behaviorally relevant meaning contained in neither Agreeableness nor indeed the entire space spanned by the FFM (and, vice versa, the FFM clearly comprises variance not contained in D), which also maps on behavioral genetic evidence suggesting that pro- and antisocial behavior are independent tendencies with distinct etiologies (Krueger, Hicks, & McGue, 2001). As such, these results rather suggest interpreting Agreeableness and D as related, but functionally different constructs (in the sense that these comprise different behaviorally relevant variance components).

This view is also corroborated by several conceptual differences between Agreeableness and D. As noted by Graziano and Tobin (2017), theoretical definitions of the construct of Agreeableness are rather sparse, somewhat incoherent, and rarely go beyond defining a list of trait or facet word descriptors. A more elaborate account, which is largely compatible with and largely subsumes other prominent definitions (e.g., Buss, 1991; Hogan, 1996; John, Naumann, & Soto, 2008), has been provided by Graziano and Tobin (2009, 2013). They broadly (albeit somewhat vaguely) define Agreeableness as the "motivation to maintain positive relations with others" (Graziano & Tobin, 2009, p. 46), tying it to individual differences in social accommodation in terms of an opponent process model comprising elements of approach and avoidance. This conceptualization differs from that of D in at least three respects. First, as a consequence of defining Agreeableness through predominantly motivational terms, there is hardly a reference to individual differences in social cognition.[1] By contrast, the definition of D directly highlights the importance of attitudes and beliefs that are used to justify malevolent behavior (and empirically, D indeed strongly relates to such beliefs; Moshagen et al., 2020). Second, whereas Graziano and Tobin's (2009, 2013) account can immediately be used to explain certain classes of relevant behaviors (such as helping others), it is rather difficult to reconcile with behaviors that impose disutility on others in absence of an explicit receiver/other (such as tax fraud or conservation behavior). Also, it seems less suited to account for sadistic or spiteful behaviors, i.e., behaviors directed at deriving utility from the very act of inflicting disutility on others – as is part of the conceptualization of D. Third, individuals with high levels in D will often be poorly described by resorting to mere avoidance. On the contrary, the core defining feature of D – seeking to maximize individual utility – is very clearly approach behavior, especially in social settings (e.g., seeking recognition, reputation, or status), as is perhaps most evident in specific dark traits such as Narcissism. Finally, it should also be noted that Graziano and Tobin's (2009, 2013) conceptualization of Agreeableness allows for a rather substantial overlap with the theoretical content of other FFM dimensions. Most obviously, individuals with high levels in Extraversion can be expected to show a pronounced motivation to maintain positive relations with others (and to exhibit strong approach tendencies, e.g., Wilt & Revelle, 2009). On a theoretical level, such a conflation is unsatisfactory given the presumed independence of the FFM dimensions.

Overall, based on the theoretical considerations sketched above, there are various reasons to motivate the assumption that Agreeableness and D show meaningful differences, which is also corroborated by initial empirical evidence provided in Moshagen et al. (2018). However, the study by Moshagen et al. (2018) was not primarily designed to dissociate D from Agreeableness (but rather to locate D in the personality spectrum overall), so that further investigation on the similarity and differences between Agreeableness and D is warranted. In particular, the criteria considered therein were not selected on theoretical grounds with the purpose to distinguish Agreeableness and D (but to distinguish D from specific dark traits), so that it might be argued that some of the criteria lack theoretical relevance. For example, lack of impulse control is a rather tangential theoretical feature of Agreeableness and D alike. Whereas the finding that both relate differently to impulsivity indicates that certain variance components differ across these constructs, this result is hardly illuminating on a theoretical level. A superior approach thus seeks criteria that allow for a theoretically grounded dissociation between Agreeableness and D.

[1] Unlike as the definition by Graziano and Tobin (2009, 2013), Agreeableness in the NEO-framework (McCrae & Costa, 2003) contains references to specific cognitions, as "Agreeableness is seen in selfless concern for others and in trusting and generous sentiments" (p. 46). This nevertheless strongly differs from the conceptualization of D which involves a much broader range of beliefs (any belief that individuals may use to justify malevolent behavior), rather than being limited to one particular belief such as distrust.

Moreover, Moshagen et al. (2018) considered only one particular operationalization of Agreeableness (via the NEO-FFI). However, there are various established operationalizations of Agreeableness beyond the one provided by the NEO-FFI, in particular the respective subscales of the Big Five Aspects Scales (BFAS; DeYoung, Quilty, & Peterson, 2007), the Big Five Inventory (BFI; Soto & John, 2017), and the International Personality Item Pool (IPIP) Big Five scales (Goldberg, 1992). These operationalizations share many key aspects inherent in the theoretical conceptualization of Agreeableness, but also display some differences regarding content and emphasis of certain features (such as BFAS-Agreeableness placing a strong weight on compassion and NEO-Agreeableness emphasizing straightforwardness; e.g., Crowe, Lynam, & Miller, 2018). Correspondingly, whereas these scales show adequate convergent validities and thus can be reasonably employed to measure an overarching Agreeableness dimension, differences in content and focus yield slightly varying measurements thereof, which, in turn, can lead to different predictive abilities for certain outcome criteria (see, e.g., the meta-analyses by Decuyper, De Pauw, De Fruyt, De Bolle, & De Clercq, 2009; Sibley & Duckitt, 2008; Thielmann, Spadaro, & Balliet, 2020). Correspondingly, in order to investigate Agreeableness vis-à-vis D on the construct level (rather than relative to any one particular instance), it is imperative to consider multiple established operationalizations of the former to capture the commonalities across different operationalizations and therefore the theoretical gist of the Agreeableness construct.

1.1. The present study

The purpose of the present study was to test whether Agreeableness and D can be considered to reflect different poles of an essentially identical single dimension or whether they should rather be considered as functionally distinct constructs. The latter position prescribes that (a) Agreeableness and D must exhibit a correlation that is substantially smaller than 1 and—if this holds—that (b) D captures behaviorally relevant variance beyond Agreeableness implying that D must incrementally predict theoretically meaningful and consequential outcome criteria over and above Agreeableness. Note that the comparison of zero-order correlations to outcome criteria is only partly informative to investigate the distinctiveness of constructs. Two dimensions may exhibit the very same zero-order correlation to an outcome, but still represent entirely different, non-overlapping variance components (unless their zero-order correlation is −1 or 1). For instance, a recent meta-analysis showed that Agreeableness and Conscientiousness show highly similar zero-order correlations to workplace deviance (r = −0.30; Pletzer, Bentvelzen, Oostrom, & de Vries, 2019), yet contribute independently to the prediction thereof, in turn illustrating that Agreeableness and Conscientiousness are functionally different. Rather than merely considering zero-order correlations, a more appropriate test thus seeks to demonstrate that one construct incrementally improves the prediction of a criterion to a substantial extent.

To obtain a comprehensive coverage of the Agreeableness construct, we assessed Agreeableness via five different established inventories. Although the present study primarily focuses on FFM-Agreeableness, we also included a measure of Agreeableness as per the HEXACO Model of Personality (Ashton & Lee, 2007) to assess Agreeableness in full breadth.[2] To test the hypothesis that D and Agreeableness are functionally different constructs, we further assessed seven criterion variables (including one behavioural outcome) in a separate session to avoid biases due to consistent reporting. The outcome criteria were selected to represent theoretically implied differences between Agreeableness and D. Specifically, to the extent that D differs from Agreeableness, D must improve the prediction of criteria that immediately reflect one (or more) of its theoretical core characteristics. Correspondingly, we considered behavioral dishonesty (maximizing own utility *disregarding* disutility of others), stereotypical sexualized behaviors (maximizing own utility *accepting* disutility of others), internet trolling (deriving own utility from malevolently *provoking* disutility on others), (lack of) guilt proneness as a consequence of the availability of justifying beliefs, and competitive and dangerous worldviews as prominent instances of such beliefs. In addition, we investigated (lack of) empathy as a psychological characteristic that is often considered to be closely linked to dark personality and thus D (Jones & Figueredo, 2013; Paulhus, 2014).

2. Methods

The data and analyses scripts are available at the open science framework at https://osf.io/xkgfp/. The study has not been preregistered.

2.1. Participants and procedure

Participants were recruited using a professionally managed online panel (prolific.ac) realizing a convenience sampling scheme. Members were eligible to participate when their approval rate exceeded 0.95 and they were born in either Ireland, the UK, or the US. We implemented two measurement occasions, each starting with participants providing informed consent and ending with demographic information and debriefing. Participants received a flat fee for every measurement occasion completed and an additional bonus of 3 GBP depending on their behavior in the mind-game (see below).

At the first measurement occasion, 729 participants (65% female; mean age = 37.06, SD = 12.96 years) completed the items measuring Agreeableness and D, respectively. Participants were native (95%) or fluent (5%) in English and showed diverse educational backgrounds with 37% holding a certificate of secondary education, 44% a college bachelor, and 13% a university degree (6% other). Of the participants, 69% were currently employed in part- or full-time. Approximately seven days after the first measurement (mean lag 6.98 days, SD = 0.17), participants were reinvited to complete the second part of the study, which yielded N = 598 valid responses (response rate 82%). Data were matched using anonymous random codes (which was additionally verified using demographic data). There was no indication of selective drop-out concerning Agreeableness or D; however, responders tended to be older than non-responders (d = 0.41, p < .05).

2.2. Measures

At the first measurement occasion, five different measures of Agreeableness (presented in random order) and a measure of D were administered. The order of the Agreeableness block and the measure of D was random. At the second measurement occasion, the self-report criterion measures were presented in random order, followed by the behavioral measure of dishonesty (the mind-game, see below) at the end. The order of the items within each scale was random. To maintain consistency, a five-point Likert response scale ranging from 1 = *strongly disagree* to 5 = *strongly agree* was used for all questionnaires (the anchors for the guilt proneness scale ranged from 1 = *extremely unlikely* to 5 = *extremely likely*).

[2] HEXACO-Agreeableness can be considered as a rotated variant of FFM-Agreeableness (Ashton et al., 2014). Although some content is shared across these conceptualizations, the former lacks the sentimentality aspect of FFM-Agreeableness (which is assigned to HEXACO-Emotionality), but additionally covers even-temper (which is assigned to FFM-Neuroticism).

Agreeableness was assessed via the corresponding scales of the BFAS (20 items; e.g., "I avoid imposing my will on others."; DeYoung et al., 2007), the BFI-2 (12 items; e.g., "I have a forgiving nature"; Soto & John, 2017), the HEXACO-100 (12 items; e.g., "I tend to be lenient in judging other people."; Lee & Ashton, 2018), the IPIP (20 items; e.g., "I think of others first"; Goldberg, 1992), and the NEO-FFI (12 items; e.g., "I generally try to be thoughtful and considerate."; Costa & McCrae, 1992; McCrae & Costa, 2004), leading to a total of 76 items representing Agreeableness.

D was assessed via a set of 70 items (D70; e.g. "My own pleasure is all that matters.") as identified in Moshagen et al. (2020) by applying rational item selection techniques on a pool of over 180 items from established scales designed to assess 12 different dark traits. The measure has been shown to possess favorable psychometric properties and exhibited substantial associations to various criterion measures, including actual behavior.

Behavioral Dishonesty was assessed via a variant of the mind-game (Jiang, 2013; Schild, Heck, Ścigała, & Zettler, 2019) which is structurally equivalent to paradigms widely used in behavioral ethics research (e.g., Gerlach, Teodorescu, & Hertwig, 2019; Heck, Thielmann, Moshagen, & Hilbig, 2018). Participants were informed that a number between 1 and 8 was going to be drawn at random (with equal probabilities) and that predicting this target number correctly would incur an additional payoff of 3 GBP. Participants were asked to choose and memorize one of these numbers. On the next screen, the randomly drawn target number was displayed. Participants were asked to indicate whether the displayed number matched their chosen number (in which case they received the additional payoff) or not (in which case they did not receive any bonus payment). Given the known baseline probability of choosing the same number as subsequently displayed (1/8), basic probability calculations allow for determining the proportion of dishonest responders (see Moshagen & Hilbig, 2017, for details). Nonetheless, responses are completely non-incriminating as any single affirmative response may always stem from actual luck (i.e., having predicted the target number correctly).

Competitive and Dangerous Worldviews are beliefs characterizing the world as a "ruthless, amoral struggle for resources and power" and "dangerous and threatening place", respectively (Duckitt, Wagner, du Plessis, & Birum, 2002, p. 78). We assessed competitive (e.g., "It's a dog-eat-dog world where you have to be ruthless at times.") and dangerous (e.g., "There are many dangerous people in our society who will attack someone out of pure meanness, for no reason at all.") worldviews by 6 items each (Duckitt et al., 2002; Sibley & Duckitt, 2009).

Empathic Concern was assessed via the respective 7-item scale by Davis (1983). A sample item is "Other people's misfortunes do not usually disturb me a great deal" (reversed).

Guilt Proneness was assessed via the five-item guilt proneness scale (GP-5; Cohen, Panter, Turan, Morse, & Kim, 2014; Cohen, Wolf, Panter, & Insko, 2011). A sample item is "You lie to people but they never find out about it. What is the likelihood that you would feel terrible about the lies you told?".

Internet Trolling was assessed via the Global Assessment of Internet Trolling (GAIT; Buckels, Trapnell, & Paulhus, 2014). The scale consists of 4 items (e.g., "I like to troll people in forums or the comments section of websites.").

Stereotypical Sexualized Behaviors were assessed via the respective 8-item scale by Jewell and Brown (2013; see also Jewell, Spears Brown, & Perry, 2015). A sample item is "During the last year I brushed up against someone in a sexual way on purpose".

2.3. Statistical analyses

The hypotheses were investigated using structural equation modeling. We estimated both Agreeableness and D using bifactor modeling (e.g., Reise, 2012). Bifactor models posit that each observed indicator of a certain construct (such as Agreeableness) loads both on a general factor representing said construct and on a specific factor representing the remaining covariances between the items of a particular measure that are not attributable to the general factor. Concerning Agreeableness, the specific factors were defined by the respective Agreeableness scale (e.g., the Agreeableness items of the BFI loaded both on the general Agreeableness factor and on a specific factor representing the specifics of the BFI; see Fig. 1). We modeled D as a general factor along with specific factors representing five dark themes (Bader et al., 2019). Thus, the complete model included two general factors (Agreeableness and D), five specific factors representing a certain measure of Agreeableness, and another five specific factors representing a certain theme of D. The specific factors representing a certain Agreeableness measure (and D theme, respectively) were mutually orthogonal and also independent from the general Agreeableness (and D, respectively) factor. Note that the specific factors represent common variance residualized for the general factor. For example, the specific factor for the BFI items reflects the remaining covariances among the BFI items that is not explained by the general Agreeableness factor. Generally, the specific factors are therefore difficult to interpret, in particular in the presence of a strong general factor (Sellbom & Tellegen, 2019), so we do not consider the specific factors in further detail. To evaluate how much of the common variance is explained by the general (vs the specific) factors, we considered the explained common variance (*ECV*; Ten Berge & Sočan, 2004), which gives the proportion of common variance explained in the items of a particular scale by the general factor relative to the specific factor. An *ECV* of 1 thus indicates that the entire shared variance of the items of a particular measure can be explained by the general factor. The relations to the criteria were investigated by adding a single latent factor for each criterion measure.

In all models, the general factors were assigned a scale by fixing their variance to 1 and the specific factors were identified by setting one unstandardized loading to 1. To address the fact that the BFAS and the IPIP contain 6 identical items, we allowed the respective residuals to correlate. Further, given that modification indices suggested localized areas of model misfit associated with pairs of items of the same Agreeableness scale, we added 8 residual correlations between items that reflected the same Agreeableness aspect in that particular measure.[3]

All models were estimated based on the raw scores using Mplus (version 7.11; Muthén & Muthén, 2015). Full information maximum likelihood estimation was employed to address incomplete data at the second measurement occasion. Model fit was evaluated through the log-likelihood ratio test statistic, while correcting for non-normality using Huber-White sandwich estimated standard errors and corresponding test-statistics (Yuan & Bentler, 2000). The models involving categorical outcomes were estimated using diagonally (robust) weighted-least squares estimation (implementing a probit link function; Muthén, duToit, & Spisic, 1997). Nested models were compared based on the scaled chi-square difference (Asparouhov & Muthén, 2006; Satorra & Bentler, 2010). In the present study, the power of the log-likelihood ratio test to detect global misspecifications of the estimated models corresponding to *RMSEA* = 0.01 on α = 0.05 was very high, $1 - \beta > 99\%$ (MacCallum, Browne, & Sugawara, 1996; Moshagen

[3] For example, "I seek conflict" and "I love a good fight" represent the politeness aspect of BFAS-Agreeableness. These items exhibited substantial correlations beyond both the commonalities of all Agreeableness items (as reflected in the Agreeableness factor) and the commonalities of the BFAS-Agreeableness items (as reflected in the specific factor for BFAS-Agreeableness), so that their residuals showed excess covariance, which was captured by allowing the residuals of these items to covary.

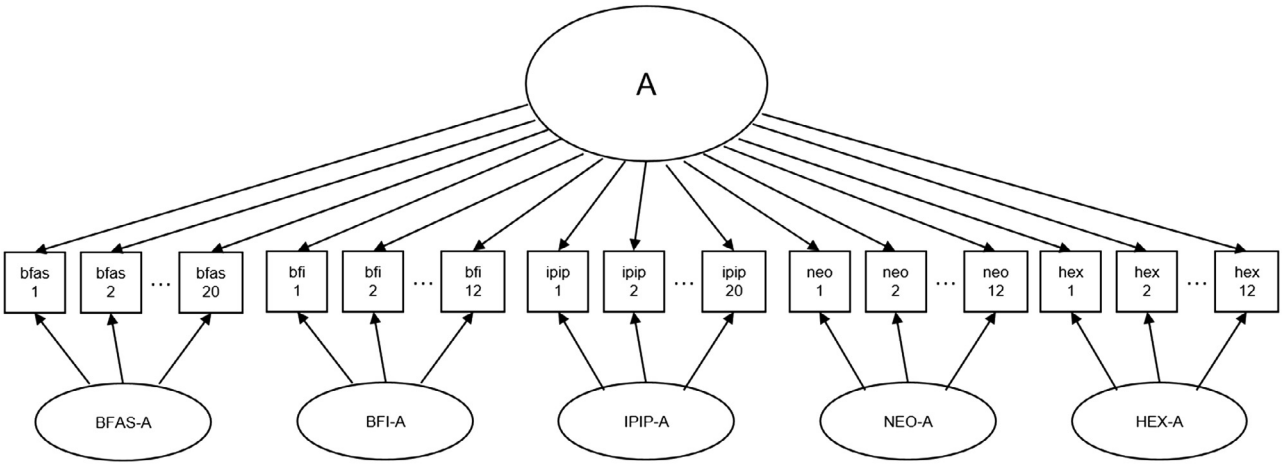

Fig. 1. Bifactor structure modeling Agreeableness (A) as a general factor affecting all indicators of all of the measures of Agreeableness. The specific factors (such as BFAS-A) represent the remaining covariances between the items of a particular scale (such as the BFAS) that are not attributable to the general factor. Note that all specific factors are mutually independent and also independent of the general factor to render the model identifiable.

Table 1
Latent correlations and internal consistencies.

Variable	α	1	2	3	4	5	6	7
1 Agreeableness	0.96							
2 D	0.95	−0.64						
3 BFI-A	0.84	–	−0.64					
4 BFAS-A	0.88	–	−0.68	0.87				
5 IPIP-A	0.92	–	−0.56	0.89	0.96			
6 NEO-A	0.81	–	−0.74	0.93	0.77	0.72		
7 HEXACO-A	0.82	–	−0.39	0.71	0.44	0.45	0.60	
Behavioral Dishonesty[1]	–	−0.17	0.26	−0.12	−0.18	−0.14	−0.16	−0.09
Competitive Jungle Beliefs	0.75	−0.62	0.84	−0.60	−0.65	−0.57	−0.62	−0.39
Dangerous World Beliefs	0.82	−0.09	0.30	−0.19	-0.08	−0.09	−0.21	−0.16
(Lack of) Empathic Concern	0.85	−0.82	0.55	−0.75	−0.81	−0.81	−0.61	−0.38
(Lack of) Guilt Proneness	0.78	−0.50	0.62	−0.47	−0.54	−0.47	−0.48	−0.32
Internet Trolling	0.70	−0.44	0.54	−0.45	−0.46	−0.41	−0.47	−0.23
Stereotypical Sexualized Behaviors	0.85	−0.02	0.18	−0.11	−0.06	0.03	−0.19	−0.02

Note. α = Cronbach's alpha estimate of internal consistency. BFI-A, BFAS-A, IPIP-A, NEO-A, and HEXACO-A give the correlations of the respective primary factors (rather than specific factors residualized for the general Agreeableness factors). All $|r| \geq 0.09$ (and $|r| \geq 0.12$ concerning the polychoric estimates, respectively) significantly differ from zero at $p < .05$.

[1] Polychoric correlations estimates.

& Erdfelder, 2016). We therefore also considered the *RMSEA* and the *SRMR* as descriptive indicators of model fit and normalized evidence ratios (ER; Wagenmakers & Farrell, 2004) to aid model comparisons. The *ER* is computed from *BIC* model weights (e.g., Bollen, Harden, Ray, & Zavisca, 2014) and expresses the likelihood that a less restricted model is superior to a more restricted comparison model. For example, an *ER* of 0.80 means that the less restricted model is 0.80/(1–0.80) = 4 times more likely than the comparison model, given the data and in terms of the degree of belief that it reflects the true model.

3. Results

The model specifying bifactor structures for both Agreeableness and D yielded a satisfactory fit to the data, $\chi^2(10{,}244) = 23{,}127$, $p < .01$; *SRMR* = 0.067; *RMSEA* = 0.042 (90%-CI: 0.041 - 0.042). All Agreeableness items loaded significantly on the general Agreeableness factor (range: 0.09–0.78; mean 0.46) and all items measuring D showed adequate loadings on the D factor (range: 0.24–0.66; mean 0.45; detailed loading estimates are provided in the osf repository). Both the Agreeableness and the D factor proved to be highly reliable (ω_H = 0.91 and ω_H = 0.90; Rodriguez, Reise, & Haviland, 2016). The *ECV* indicated that the Agreeableness factor accounted for 72% of the common variance among all Agreeableness items. Similarly, the D factor accounted for 70% of the common variance among the D items. Thus, the factors indicating Agreeableness and D, respectively, exhibited highly similar psychometric properties. The Agreeableness factor was most strongly reflected in the items of the IPIP (*ECV* = 0.89), the BFAS (*ECV* = 0.80), the BFI (*ECV* = 0.75), and—somewhat less—of the NEO-FFI (*ECV* = 0.55), but showed comparatively weaker relations to the items of the HEXACO (*ECV* = 0.25), thereby mirroring differences in the theoretical conceptualizations of FFM- versus HEXACO-Agreeableness (Ashton, Lee, & De Vries, 2014). However, the results generally support the idea that the most prominent measures of Agreeableness converge with respect to a single construct.

The bivariate latent correlations between Agreeableness (modeled as the general factor in the bifactor specification), D, all specific Agreeableness operationalizations (modeled as primary factors), and the criterion measures are shown in Table 1. Most importantly, the correlation between the Agreeableness factor and the D factor was estimated at $r = -.64$ ($r^2 = 41\%$) and was thus similar in magnitude to the one reported in Moshagen et al. (2018). Further, the correlations between the FFM-Agreeableness measures were generally higher than the one between Agreeableness and D with a median latent correlation between the FFM

Table 2
Latent regression results.

Outcome variable	β_A	β_D	$R2_{(A, D)}$	$\Delta R2$	$\Delta\chi2$	*ER*	*q*
Behavioral Dishonesty[1]	−0.08	0.23*	0.07	0.04	6.0*	0.927	0.06*
Competitive Jungle Beliefs	−0.16*	0.74*	0.71	0.29	207.6*	>0.999	0.49*
Dangerous World Beliefs	0.10	0.34*	0.08	0.07	31.2*	>0.999	0.21*
(Lack of) Empathic Concern	−0.79*	0.06	0.69	<0.01	0.6	0.112	0.54*
(Lack of) Guilt Proneness	−0.16*	0.55*	0.43	0.17	49.1*	>0.999	0.17*
Internet Trolling	−0.14*	0.47*	0.32	0.12	58.0*	>0.999	0.13*
Stereotypical Sexualized Behaviors	0.19*	0.33*	0.06	0.06	32.8*	>0.999	0.16*

Note. A = Agreeableness. *q* gives the difference in the (absolute) zero-order correlations to an outcome between A and D as measured by Cohen's *q* with associated (one-sided and Holm-Bonferroni corrected) *p*-value (Williams, 1959). $\Delta\chi^2$ is the (scaled) log-likelihood ratio test, *ER* the evidence ratio, and ΔR^2 refers to the increase in the variance explained; all comparing a model predicting the criterion by Agreeableness and D versus a model predicting the criterion by Agreeableness only.
[1] Probit regression coefficients; ΔR^2 gives the increase in the variance explained in the latent response variable of behavioral dishonesty.
* $p < .05$.

Agreeableness scales of 0.88, which was thus about 38% stronger than the correlation between Agreeableness and D.

As a formal test of whether Agreeableness and D can be considered to reflect merely opposite poles of the same dimension, we estimated a model restricting their correlation to (negative) unity (which is equivalent to assuming a single factor comprising both Agreeableness and D). This led to a significant decrease in model fit, $\Delta\chi^2(1) = 23.7$, $p < .01$, and was associated with an evidence ratio of $ER > 0.999$ showing that the unrestricted model is over 1,000 times as likely as the model assuming a perfect correlation between Agreeableness and D, thereby disconfirming that (low) Agreeableness and D represent a unitary dimension. In light of the differences between FFM- and HEXACO-Agreeableness, we also estimated a model specifying a bifactor structure for Agreeableness excluding the respective HEXACO items. However, this led to virtually the same correlation to D ($r = -.65$). In sum, these results suggest that Agreeableness and D share a substantial proportion—though less than half—of variance, yet are separate constructs that cannot be considered opposite poles of a single dimension.

To scrutinize the conclusion that Agreeableness and D are functionally distinct constructs, we further considered how they relate to the seven criterion measures. If Agreeableness and D can essentially be considered to reflect opposite poles of the same dimension, no systematic differences between these two would be expected to occur concerning their relation to other theoretically relevant psychological attributes, that is, they would have to exhibit a high degree of nomological consistency (Hilbig, Moshagen, & Zettler, 2016; Thielmann & Hilbig, 2019) and would have to show extrinsic convergent validity (Gonzalez, MacKinnon, & Muniz, 2020).

We first consider the behavioral measure of dishonesty. Of the participants, 37% indicated to have correctly predicted the displayed target number in the mind-game and thus received the additional payoff. Based on the baseline probability of 1/8, 24.5% are thus estimated to have cheated (Moshagen & Hilbig, 2017). Whereas the observed responses were related to both Agreeableness ($r = -0.17$) and D ($r = 0.26$)[4], only D significantly predicted responses in a latent probit regression using both Agreeableness and D as predictors. Likewise, a log-likelihood ratio test versus a model omitting D indicated the inferiority of the model only including Agreeableness, $\Delta\chi^2(df = 1) = 6.03$, $p = .01$, as did the evidence ratio in favor of the model including D, $ER = 0.927$.

Concerning the self-report criteria, Agreeableness exhibited a significantly stronger correlation to empathic concern ($r = 0.82$ vs. $r = -0.55$), but significantly weaker correlations than D to all remaining criteria (Table 1). Correspondingly, latent regressions (Table 2) revealed that D incrementally predicted most criteria to a substantial extent ($0.06 \leq \Delta R^2 \leq 0.29$), again with the exception of empathic concern ($\Delta R^2 < 0.01$), and the evidence ratios indicated to prefer the model including D as predictor (except for empathic concern, $ER = 0.112$). Taken together, the relations to the criteria rather suggest nomological inconsistency between Agreeableness and D, thereby suggesting that they are functionally distinct constructs.

To further gauge the similarities of Agreeableness and D with respect to their pattern of correlations to the criteria, we evaluated the extrinsic convergent validity hypothesis using Cohen's *q* as effect size measure (and associated Holm-Bonferroni corrected *p*-values according to Williams, 1959) and further considered the double-entry *ICC* along with measures of shape, scatter, and elevation similarity as recommended by Furr (2010), as well as the root-mean-square error (*RMSE*)[5] as measures of profile similarity. Over the seven criteria, profile similarity was estimated at $ICC = 0.747$, shape similarity was $r = 0.828$, scatter similarity was 0.033, and elevation similarity was 0.090.[6] By comparison, the measures of FFM-Agreeableness were associated with a median $ICC = 0.957$, a median shape $r = 0.979$, a median scatter 0.025, and a median elevation 0.023. The average deviation between Agreeableness and D in the correlational patterns to the criteria was $RMSE = 0.179$. Finally, all correlations to the outcomes of Agreeableness versus D significantly differed, with the magnitude of difference corresponding to a medium effect on average ($q = 0.25$), thereby uniformly indicating to reject the extrinsic convergent validity hypothesis.

As a robustness check and to address predictor-criterion contamination, we repeated the regression analyses omitting items from the measurement model for Agreeableness and D, respectively, which yielded a substantial content overlap to the items of a particular criterion measure. For instance, the IPIP-item "I have a soft heart" is highly similar to the item "I would describe myself

[4] In interpreting these results, it is important to note that a certain fraction of the participants actually did predict the correct target number and were thus honest in obtaining the additional payoff. This leads to attenuated correlation estimates and regression results that require correction using modified analytic procedures (Moshagen & Hilbig, 2017). However, these approaches are not yet available in the context of latent variable modeling, so that the obtained correlations and regression results (based on the observed responses in the mind-game) underestimate the true relationships. Indeed, applying the correction factor (Moshagen & Hilbig, 2017, Eqn. 6) to the attenuated correlation estimates reported above (of $r = -0.17$ and $r = 0.26$, respectively) yields corrected estimates of $r = -0.20$ and $r = 0.32$, respectively, and thus Cohen's $q = 0.13$ ($p < .01$).

[5] The RMSE is the root of the mean squared difference, $RMSE = \sqrt{\frac{1}{k}\sum(r_{A,i} - r_{D,i})^2}$, where *k* denotes the number of criteria.

[6] Agreeableness and D were coded to point in the same direction before computing all measures of profile similarity. Leaving the direction of Agreeableness and all outcomes as implied by their label (e.g., so that low Agreeableness corresponds to high D), the resulting measures of similarity were notably different, namely $ICC = -0.970$, shape similarity $r = -0.971$, scatter similarity 0.035, elevation similarity 0.139, and $RMSE = 0.971$. Concerning the measures of FFM-Agreeableness, the same approach yielded a median $ICC = 0.988$, a median shape $r = 0.994$, a median scatter 0.029, a median elevation 0.047, and a median $RMSE = 0.069$.

as a pretty soft-hearted person" from the measure of empathic concern, as is the D70-item "I'm not very sympathetic to other people or their problems" to "Sometimes I don't feel very sorry for other people when they are having problems" (also from empathic concern). Correspondingly, we excluded such items to investigate whether content overlap drives the correlations to the criteria, which, in turn, might bias the comparison between Agreeableness and D. Unsurprisingly, the correlations to the criteria proved to be slightly weaker when overlapping items were omitted. However, the regressions yielded equivalent results throughout, i.e., adding D to the model improved the prediction of all criteria, except for empathic concern, to approximately the same extent as in the results without any item omissions, $0.05 \leq \Delta R^2 \leq 0.27$ (see osf repository for details).

Finally, we investigated whether the superior predictive performance of D versus Agreeableness can be traced back to different abilities of the underlying item-sets to differentiate at particular positions on the latent trait spectra. To this end, we estimated graded item response models to obtain test-information functions for the Agreeableness items and the D items (which were recoded to point in the same direction as the Agreeableness items). Results revealed that both item-sets were associated with highly similar test information functions showing a peak at rather low latent trait levels (detailed results are provided in the osf repository), thereby indicating that both item-sets yield the highest information at approximately the same latent trait levels.

4. Discussion

The Dark Factor of Personality (D) has been suggested as the basic disposition responsible for the emergence of dark traits, thereby representing their commonalities. However, considering the Dark Triad components in particular, it has also been argued that their commonalities represent the low pole of Agreeableness (e.g., Jakobwitz & Egan, 2006; Paulhus & Williams, 2002; Stead & Fekken, 2014; Vize et al., 2019, in press) as included in models of basic personality structure, especially the FFM. In the present study, we investigated whether this logic extends to the common core of all dark traits and thus whether Agreeableness and D can be considered as merely opposite poles of an essentially identical dimension or whether they can rather be assumed to represent functionally different constructs in terms of comprising different behaviorally relevant variance components.

Relying on a broad measurement of Agreeableness using the respective scales of five established inventories, results suggest that Agreeableness and D are best understood as related, but functionally different constructs. In support of the position that Agreeableness and D are related, their shared variance was estimated at approximately 41%, thereby indicating substantial similarities and shared content in some respects. However, results further illustrated that the proportions of variance unique to either Agreeableness or D also carry psychologically relevant meaning, as evident in the fact that both relate differently to a host of relevant criterion measures.

In particular, D was shown to exhibit stronger correlations to and to improve the prediction of criterion variables that immediately relate to the definitional core aspects of D. According to the theoretical definition of D, individuals with high levels are thought to maximize their individual utility "disregarding, accepting, or malevolently provoking disutility for others" (Moshagen et al., 2018, p. 657). These aspects were corroborated by the findings that D related stronger than (and beyond) Agreeableness to behavioral dishonesty (disregarding others' disutility), stereotypical sexualized behaviors (accepting others' disutility), and internet trolling (deriving utility from malevolently provoking disutility). In addition, the final aspect inherent in the definition of D that individuals will hold "beliefs that serve as justifications" (p. 657) was supported by stronger relations to guilt proneness as well as to competitive and dangerous worldviews, thereby highlighting the importance of attitudes and beliefs that can be used to justify malevolent behaviors. As such, the results are aligned with the theoretical definition of D and rather speak against regarding (low) Agreeableness as a substitute of D.

Beyond the criteria selected to reflect a definitional core aspect of D (as reviewed above), empathy was also considered as a psychological characteristic that has often been suggested to relate strongly to the core of dark traits (e.g., Jones & Figueredo, 2013; Paulhus, 2014). However, D exhibited lower (though still substantial) correlations to empathic concern than Agreeableness and did not incrementally predict empathic concern over Agreeableness. In hindsight, it is actually plausible that a certain degree of cognitive empathy is required to display malevolent behaviors that aim at deriving utility from the disutility inflicted on others, as hinted by findings that the Dark Triad/Tetrad components (and Sadism in particular) show stronger (negative) relations to affective as compared to cognitive aspects of empathy (Kajonius & Björkman, 2019; Pajevic, Vukosavljevic-Gvozden, Stevanovic, & Neumann, 2018). Nevertheless, although unexpected, the very fact that Agreeableness displayed substantially stronger correlations to the measure of empathy employed herein is another indication of functionally different variance components inherent in Agreeableness and D, with the former apparently capturing individual differences in empathy in a more general way.

Considering the overall pattern of how Agreeableness versus D were associated with the criteria, the conclusion that these constructs comprise different behaviorally relevant variance components was further supported by consistent evidence against the extrinsic convergent validity hypothesis (Gonzalez et al., 2020) with an average Cohen's *q* of 0.25 and an *RMSE* of 0.18, thus indicating rather substantial differences. The assessment of profile similarity exhibited conflicting results, however. Depending on which measures of similarity (and direction of scales) are considered, the correlational profiles could be interpreted as being more or less in line with the view to consider Agreeableness and D equivalent. Nonetheless, it should be kept in mind that the criteria were not selected with the aim to yield different profiles of Agreeableness versus D (but rather to show stronger correlations of D, which does not necessarily translate to profile dissimilarity). Clearly, it would be useful to extend the present study by relating Agreeableness and D to a wider array of relevant behaviors as criterion measures beyond the one considered herein to shed further light on this issue.

It should be noted that the observed pattern of results cannot be explained by arguing that stronger relationships are to be expected when predictor and criterion occupy the same pole (i.e., Agreeableness better predicts positively connoted outcomes whereas D better predicts negatively connoted outcomes) or by arguing that the measure of D comprises more extremely worded items. Indeed, guilt proneness is a positively connoted attribute, yet its prediction was vastly improved when adding D (ΔR^2 = 0.17), and all of the considered criterion measures comprise rather moderately worded items, in turn being more aligned with the item wording realized in the measures of Agreeableness. Likewise, the latent variables for Agreeableness and D exhibited highly similar psychometric properties, were measured with a comparable number of items (if anything, factor saturation and reliability of Agreeableness was higher than that of D), and both the items indicating Agreeableness and D were almost perfectly balanced with respect to the keyings (thus making an effect of polarity implausible), so the results can neither be explained by resorting to any of such arguments. However, it should be noted that both Agreeableness

and D were assessed online using self-reports, so it might be worthwhile to consider peer-reports as a complementary data source.

It should also be kept in mind that the results can only be considered valid to the extent that the chosen operationalizations can be seen as comprehensive indicators of the constructs they intend to represent. This might be particularly the case concerning the assessment of Agreeableness, given that multiple operationalizations exist that place a different emphasis on certain features and thus may represent different aspects of the broader construct of Agreeableness. For example, most common measures of Agreeableness contain little content related to humility or straightforwardness, both of which are arguably particularly relevant concerning D, so that measuring Agreeableness by other instruments might yield different conclusions. Although we attempted to realize comprehensive construct coverage of both Agreeableness by resorting to five commonly used and well established measures and D, it is still possible that some features were not well represented in the chosen measures with the consequence that the correlation obtained herein might have under- or overestimated (depending on which features were underrepresented) the true relation between Agreeableness and D on construct level.

An alternative interpretation of our results would be to argue that the employed measure of D merely offers a broader (and perhaps superior) representation of the construct of (FFM-) Agreeableness. That is, the imperfect association between Agreeableness and D would not be interpreted to imply that these represent distinct dimensions. Rather, referring to the likewise imperfect association between various operationalizations of Agreeableness, it might be argued that the items used to measure D simply represent another operationalization of an overarching Agreeableness dimension. However, we argue that this interpretation falls short for three reasons. First, the results indicate that the considered operationalizations of FFM-Agreeableness show stronger correlations and are more similar to each other (with a median of r = 0.88) than to the measure of D (with a median of r = 0.66). Thus, despite the differences in content and focus inherent in common FFM-Agreeableness measures, it seems fair to conclude that all largely converge on a single construct, whereas the measure of D appears somewhat off. Second, the items contained in the measure of D generally correspond to aspects inherent in the theoretical conceptualization of D, which, however, differs in several respects of the account of Agreeableness as provided by Graziano and Tobin (2009, 2013). Viewing the measure of D as an instance of the Agreeableness construct in the sense of Graziano and Tobin (2009, 2013) would require an explication of their conceptualization to cover cognitions related to justifying beliefs in a more comprehensive way, utility maximization in absence of an explicit other, and behaviors related to sadism and spite, as well as an explanation of strategic social accommodation for purely egoistic motives. Finally, as a consequence of its theoretical origin as part of an integrated five-factor system (assuming approximately independent basic dimensions), Agreeableness cannot be seen in isolation, but must be viewed in the context of the remaining dimensions of the FFM. However, the measure of D comprises various features that bear resemblance to facets commonly assigned to other FFM dimensions, such as warmth (Extraversion), self-discipline (Conscientiousness), or hostility (Neuroticism) in the NEO-PI-R. Regarding the measure of D as a mere expression of (low) Agreeableness would thus require to rotate (and thus change the content of) the remaining dimensions in order to maintain approximate independence, which is generally desired to meet the purpose of basic models of personality structure to provide comprehensive description of individual differences by as few and non-redundant dimensions as possible (e.g., Goldberg, 1992; McCrae & Costa, 2003; Saucier, 2002). Relatedly, a recent investigation (Vize, Miller, & Lynam, 2020) based on 104 Agreeableness-related items (also including content from constructs other than Agreeableness, such as the Altruism and Honesty-Humility scales from the HEXACO-PI-R) indicated that a thereby obtained factor relates more strongly to D than Agreeableness in the present study. Crucially, and in line with the present arguments, the associations between this factor and all remaining FFM-dimensions were substantial (correlations of -0.35 to Neuroticism, 0.28 to Extraversion, 0.42 to Openness, and 0.57 to Conscientiousness)[7] and thus notably stronger than the typical associations between FFM-Agreeableness and the remaining FFM-dimensions (see Park et al., 2020, for a recent second order meta-analysis). The factor obtained in Vize et al. (2020) thus does not seem to be a representation of Agreeableness as defined within the FFM, but a notably broader construct that is not approximately independent of the remaining FFM dimensions in the same range as FFM-research typically suggests. Given that D theoretically and indeed empirically overlaps substantially with some of these remaining FFM dimensions (especially Conscientiousness; Moshagen et al., 2018), it is unsurprising that the factor Vize et al. termed Agreeableness more closely corresponds to D. Thus, whereas it thus might well be possible to construct a factor based on Agreeableness-related items that closely mimics D, such a resulting factor also carries substantial content of other FFM dimension and thus cannot be readily interpreted as one of few basic and largely orthogonal dimensions of personality as conceptualized in the FFM. Correspondingly, rather than trying to broaden Agreeableness so that it covers D in its entirety, the purpose of the FFM is done more justice when describing D as a blend of several fundamental personality dimensions, in line with the notion that "D is not well suited for inclusion in a more general model of personality dimensions" (Moshagen et al., 2018, p. 682).

Overall, the results are rather difficult to be reconciled with the proposition that Agreeableness (as an approximately orthogonal dimension in a model of basic personality structure) and D essentially reflect opposite poles from the same dimension and are more aligned with interpreting Agreeableness and D as functionally distinct constructs that comprise different behaviorally relevant variance components. As indicated by their substantial association, Agreeableness may serve as a reasonable proxy for D within the FFM. However, given that D also comprises features typically assigned to other FFM dimensions and given that Agreeableness and D share less than half of the variance, it would be inappropriate to treat them as interchangeable constructs. For example, despite strong correlations between body height and body weight (about r = 0.80; Heinz, Peterson, Johnson, & Kerk, 2003) one would hardly argue that both represent the same entity. Thus, rather than considering a substantial (latent) correlation as sufficient evidence for collapsing different constructs, differences in their respective nomological net needs to be thoroughly evaluated (Borsboom, Mellenbergh, & van Heerden, 2004; Gonzalez et al., 2020; Hilbig et al., 2016; Thielmann & Hilbig, 2019), as we have done herein.

In conclusion, using a broad measurement approach and considering various relevant criteria theoretically derived to reflect core characteristics of D, the results of the present study are rather difficult to reconcile with the assumption that the commonalities of dark traits can be seen as mere reflection of low Agreeableness. Rather, the results are better aligned with the contention to consider Agreeableness and D as functionally distinct constructs.

[7] The preprint only reports the correlations to the facet scores. The script used to compute the correlation to the FFM domains is available at https://osf.io/xkgfp/.

Author Note

All authors contributed equally to this work. Preparation of this manuscript was supported by a grant from the Carlsberg Foundation (CF16-0444) to Ingo Zettler, as well as Grants HI 1600/1-2 and HI 1600/6-1 funded by the German Research Foundation (DFG) to Benjamin E. Hilbig. Luisa K. Horsten was supported by the research-training group *Statistical Modeling in Psychology* (GRK 2277), funded by the German Research Foundation (DFG).

References

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 11*, 150–166.

Ashton, M. C., Lee, K., & De Vries, R. E. (2014). The HEXACO honesty-humility, agreeableness, and emotionality factors: A review of research and theory. *Personality and Social Psychology Review, 18*, 139–152.

Asparouhov, T., Muthén, B. (2006). Robust chi square difference testing with mean and variance adjusted test statistic. Retrieved from http://statmodel2.com/download/webnotes/webnote10.pdf.

Bader, M., Hartung, J., Hilbig, B. E., Zettler, I., Moshagen, M., Wilhelm, O. (2019). Themes of the Dark Core of Personality. Submitted for Publication. Retrieved from https://osf.io/aq4j3.

Bollen, K. A., Harden, J. J., Ray, S., & Zavisca, J. (2014). BIC and Alternative Bayesian Information Criteria in the Selection of Structural Equation Models. *Structural Equation Modeling, 21*, 1–19.

Borkenau, P., & Ostendorf, F. (1994). *NEO-Fünf-Faktoren Inventar (NEO-FFI): Handanweisung*. Göttingen: Hogrefe.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061.

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences, 67*, 97–102. https://doi.org/10.1016/j.paid.2014.01.016.

Buss, D. M. (1991). Evolutionary personality psychology. *Annual Review of Psychology, 42*, 459–491. https://doi.org/10.1146/annurev.ps.42.020191.002331.

Cohen, T. R., Panter, A. T., Turan, N., Morse, L., & Kim, Y. (2014). Moral character in the workplace. *Journal of Personality and Social Psychology, 107*, 943–963. https://doi.org/10.1037/a0037245.

Cohen, T. R., Wolf, S. T., Panter, A. T., & Insko, C. A. (2011). Introducing the GASP scale: A new measure of guilt and shame proneness. *Journal of Personality and Social Psychology, 100*(5), 947–966. https://doi.org/10.1037/a0022641.

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

Costa, P. T., McCrae, R. R., & Dye, D. A. (1991). Facet scales for agreeableness and conscientiousness: A revision of the NEO personality inventory. *Personality and Individual Differences, 12*, 887–898. https://doi.org/10.1016/0191-8869(91)90177-D.

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. https://doi.org/10.1111/jopy.12358.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*, 113–126.

Decuyper, M., De Pauw, S., De Fruyt, F., De Bolle, M., & De Clercq, B. J. (2009). A meta-analysis of psychopathy-, antisocial PD- and FFM associations. *European Journal of Personality, 23*(7), 531–565. https://doi.org/10.1002/per.729.

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.

Diebels, K. J., Leary, M. R., & Chon, D. (2018). Individual differences in selfishness as a major dimension of personality: A reinterpretation of the sixth personality factor. *Review of General Psychology, 22*, 367–376.

Duckitt, J., Wagner, C., du Plessis, I., & Birum, I. (2002). The psychological bases of ideology and prejudice: Testing a dual process model. *Journal of Personality and Social Psychology, 83*, 75–93.

Furnham, A., Richards, S., Rangel, L., & Jones, D. N. (2014). Measuring malevolence: Quantitative issues surrounding the Dark Triad of personality. *Personality and Individual Differences, 67*, 114–121. https://doi.org/10.1016/j.paid.2014.02.001.

Furr, R. M. (2010). The double-entry intraclass correlation as an index of profile similarity: Meaning, limitations, and alternatives. *Journal of Personality Assessment, 92*(1), 1–15. https://doi.org/10.1080/00223890903379134.

Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin, 145*, 1–44.

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

Gonzalez, O., MacKinnon, D. P., & Muniz, F. B. (2020). Extrinsic convergent validity evidence to prevent jingle and jangle fallacies. *Multivariate Behavioral Research*. https://doi.org/10.1080/00273171.2019.1707061.

Graziano, W. G., & Tobin, R. M. (2009). Agreeableness. In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior* (pp. 46–61). New York: Guilford.

Graziano, W. G., & Tobin, R. M. (2013). The cognitive and motivational foundations underlying agreeableness. In M. D. Robinson, E. R. Watkins, & E. Harmon-Jones (Eds.), *Handbook of cognition and emotion* (pp. 347–364). New York: Guilford.

Graziano, W. G., & Tobin, R. 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: Oxford University Press.

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.

Heinz, G., Peterson, L. J., Johnson, R. W., & Kerk, C. J. (2003). Exploring relationships in body dimensions. *Journal of Statistics Education, 11*(2).

Hilbig, B. E., Moshagen, M., & Zettler, I. (2016). Prediction consistency: A test of the equivalence assumption across different indicators of the same construct.: Prediction consistency. *European Journal of Personality, 30*, 637–647. https://doi.org/10.1002/per.2085.

Hogan, R. (1996). A socioanalytic perspective on the five-factor model. In J. S. Wiggins (Ed.), *The five factor model of personality: Theoretical perspectives* (pp. 163–179). New York: Guilford.

Jakobwitz, S., & Egan, V. (2006). The dark triad and normal personality traits. *Personality and Individual Differences, 40*, 331–339.

Jewell, J. A., & Brown, C. S. (2013). Sexting, catcalls, and butt slaps: How gender stereotypes and perceived group norms predict sexualized behavior. *Sex Roles, 69*(11–12), 594–604. https://doi.org/10.1007/s11199-013-0320-1.

Jewell, J., Spears Brown, C., & Perry, B. (2015). All my friends are doing it: Potentially offensive sexual behavior perpetration within adolescent social networks. *Journal of Research on Adolescence, 25*, 592–604. https://doi.org/10.1111/jora.12150.

Jiang, T. (2013). Cheating in mind games: The subtlety of rules matters. *Journal of Economic Behavior & Organization, 93*, 328–336.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm Shift to the Integrative Big-Five Trait Taxonomy: History, Measurement, and Conceptual Issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality, third edition: Theory and research* (pp. 114–158). New York: Guilford Press.

Jonason, P. K., Li, N. P., Webster, G. D., & Schmitt, D. P. (2009). The dark triad: Facilitating a short-term mating strategy in men. *European Journal of Personality, 23*, 5–18.

Jones, D. N., & Figueredo, A. J. (2013). The core of darkness: Uncovering the heart of the dark triad. *European Journal of Personality, 27*, 521–531.

Kajonius, P. J., & Björkman, T. (2019). Individuals with dark traits have the ability but not the disposition to empathize. *Personality and Individual Differences, 109716*. https://doi.org/10.1016/j.paid.2019.109716.

Krueger, R. F., Hicks, B. M., & McGue, M. (2001). Altruism and antisocial behavior: Independent tendencies, unique personality correlates, distinct etiologies. *Psychological Science, 12*(5), 397–402. https://doi.org/10.1111/1467-9280.00373.

Lee, K., & Ashton, M. C. (2018). Psychometric properties of the HEXACO-100. *Assessment, 25*, 543–556.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130–149.

Marcus, D. K., Zeigler-Hill, V., Mercer, S. H., & Norris, A. L. (2014). The psychology of spite and the measurement of spitefulness. *Psychological Assessment, 26*, 563–574.

McCrae, R. R., & Costa, P. T. (2003). *Personality in Adulthood: A Five-Factor Theory Perspective* (2nd ed.). New York: Guilford.

McCrae, R. R., & Costa, P. T. (2004). A contemplated revision of the NEO Five-Factor Inventory. *Personality and Individual Differences, 36*, 587–596. https://doi.org/10.1016/S0191-8869(03)00118-1.

McCrae, R. R., & Costa, P. T. (2008). A five-factor theory of personality. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality psychology: Theory and research* (3rd ed., pp. 159–181). New York: Guilford.

Moshagen, M., & Erdfelder, E. (2016). A new strategy for testing structural equation models. *Structural Equation Modeling, 23*, 54–60.

Moshagen, M., & Hilbig, B. E. (2017). The statistical analysis of cheating paradigms. *Behavior Research Methods, 49*, 724–732. https://doi.org/10.3758/s13428-016-0729-x.

Moshagen, M., Hilbig, B. E., & Zettler, I. (2018). The dark core of personality. *Psychological Review, 125*, 656–688. https://doi.org/10.1037/rev0000111.

Moshagen, M., Zettler, I., & Hilbig, B. E. (2020). Measuring the dark core of personality. *Psychological Assessment, 32*, 182–196. https://doi.org/10.1037/pas0000778.

Muris, P., Merckelbach, H., Otgaar, H., & Meijer, E. (2017). The malevolent side of human nature: A meta-analysis and critical review of the literature on the Dark Triad (Narcissism, Machiavellianism, and Psychopathy). *Perspectives on Psychological Science, 12*, 183–204.

Muthén, B., duToit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continous outcomes. Retrieved from http://statmodel.com/download/Article_075.pdf.

Muthén, L. K., & Muthén, B. (2015). *Mplus user's guide* (7th ed.). Los Angeles: Muthén & Muthén.

O'Boyle, E. H., Forsyth, D. R., Banks, G. C., & McDaniel, M. A. (2012). A meta-analysis of the dark triad and work behavior: A social exchange perspective. *Journal of Applied Psychology, 97*, 557–579.

O'Boyle, E. H., Forsyth, D. R., Banks, G. C., Story, P. A., & White, C. D. (2015). A meta-analytic test of redundancy and relative importance of the dark triad and five-factor model of personality. *Journal of Personality, 83*, 644–664.

O'Meara, A., Davies, J., & Hammond, S. (2011). The psychometric properties and utility of the Short Sadistic Impulse Scale (SSIS). *Psychological Assessment, 23*, 523–531.

Pajevic, M., Vukosavljevic-Gvozden, T., Stevanovic, N., & Neumann, C. S. (2018). The relationship between the Dark Tetrad and a two-dimensional view of empathy. *Personality and Individual Differences, 123*, 125–130. https://doi.org/10.1016/j.paid.2017.11.009.

Park, H. (Hailey), Wiernik, B. M., Oh, I.-S., Gonzalez-Mulé, E., Ones, D. S., & Lee, Y. (2020). Meta-analytic five-factor model personality intercorrelations: Eeny, meeny, miney, moe, how, which, why, and where to go. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0000476.

Paulhus, D. L. (2014). Toward a taxonomy of dark personalities. *Current Directions in Psychological Science, 23*, 421–426.

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

Pletzer, J. L., Bentvelzen, M., Oostrom, J. K., & de Vries, R. E. (2019). A meta-analysis of the relations between personality and workplace deviance: Big Five versus HEXACO. *Journal of Vocational Behavior, 112*, 369–383. https://doi.org/10.1016/j.jvb.2019.04.004.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*, 667–696.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137–150. https://doi.org/10.1037/met0000045.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika, 75*, 243–248.

Saucier, G. (2002). Orthogonal markers for orthogonal factors: The case of the Big Five. *Journal of Research in Personality, 36*(1), 1–31. https://doi.org/10.1006/jrpe.2001.2335.

Schild, C., Heck, D. W., Ścigała, K. A., & Zettler, I. (2019). Revisiting REVISE:(Re) Testing unique and combined effects of REminding, VIsibility, and SElf-engagement manipulations on cheating behavior. *Journal of Economic Psychology, 75* 102161.

Sellbom, M., & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological Assessment, 31*, 1428–1441. https://doi.org/10.1037/pas0000623.

Seuntjens, T. G., Zeelenberg, M., van de Ven, N., & Breugelmans, S. M. (2015). Dispositional greed. *Journal of Personality and Social Psychology, 108*, 917–933. https://doi.org/10.1037/pspp0000031.

Sibley, C. G., & Duckitt, J. (2008). Personality and prejudice: A meta-analysis and theoretical review. *Personality and Social Psychology Review, 12*, 248–279.

Sibley, C. G., & Duckitt, J. (2009). Big-Five personality, social worldviews, and ideological attitudes: Further tests of a dual process cognitive-motivational model. *Journal of Social Psychology, 149*, 545–561. https://doi.org/10.1080/00224540903232308.

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*, 117–143.

Stead, R., & Fekken, G. C. (2014). Agreeableness at the core of the dark triad of personality. *Individual Differences Research, 12*, 131–141.

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69*, 613–625.

Thielmann, I., & Hilbig, B. E. (2019). Nomological consistency: A comprehensive test of the equivalence of different trait indicators for the same constructs. *Journal of Personality, 87*, 715–730. https://doi.org/10.1111/jopy.12428.

Thielmann, I., Spadaro, G., & Balliet, D. (2020). Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin, 30–90*. https://doi.org/10.1037/bul0000217.

Vize, C. E., Lynam, D. R., Collison, K. L., & Miller, J. D. (2018). Differences among dark triad components: A meta-analytic investigation. *Personality Disorders: Theory, Research, and Treatment, 9*, 101–111. https://doi.org/10.1037/per0000222.

Vize, C. E., Miller, J. D., & Lynam, D. R. (2019). Antagonism in the Dark Triad. In J. D. Miller & D. R. Lynam (Eds.), The Handbook of Antagonism (pp. 253–267). 10.1016/B978-0-12-814627-9.00017-7.

Vize, C., Lynam, D., Collison, K., & Miller, J. (in press). The" Core" of the Dark Triad: A test of competing hypotheses. Personality Disorders: Theory, Research, and Treatment.

Vize, C., Miller, J., & Lynam, D. (2020). Examining the Conceptual and Empirical Distinctiveness of Agreeableness and "Dark" Personality Items.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin and Review, 11*, 192–196.

Williams, E. J. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society: Series B, 21*, 396–399. https://doi.org/10.1111/j.2517-6161.1959.tb00346.x.

Wilt, J., & Revelle, W. (2009). Extraversion. In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior* (pp. 27–45). New York: Guilford.

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 30*, 165–200.