Manoeuvering through the jangle jungle.

What is the common core of aversive personality traits (not)?

Vom Promotionsausschuss des Fachbereichs Psychologie
der RPTU Kaiserslautern-Landau
zur Verleihung des akademisches Grades Doktor der Philosophie (Dr. phil.)
genehmigte Dissertation

vorgelegt von
Luisa K. Horsten

Vorsitzende des Promotionsausschusses: Prof. Dr. Tanja Lischetzke
Berichterstatterin: Prof. Dr. Eunike Wetzel
Berichterstatter: Prof. Dr. Edgar Erdfelder

Datum der wissenschaftlichen Aussprache: 23. Mai 2023
Numquam ponenda est pluralitas sine necessitate
[Plurality must never be posited without necessity].

William of Ockham
This thesis is based on five articles that have been published in peer-reviewed journals. In the following, these articles are listed in the order in which they are discussed in and appended to this thesis.


A corrigendum for the third article was published due to minor errors in the results section. This corrigendum will not be mentioned separately throughout this thesis but is meant to be included whenever the corresponding article is addressed.

Research across virtually all subfields of psychology has suffered from construct proliferation, often resulting in redundant constructs that strongly overlap conceptually and/or empirically. Such cases of old wine in new bottles, i.e., established constructs with new labels, are instances of the jangle fallacy and are problematic because they lead to fragmented literatures and thereby considerably impede the accumulation of knowledge.

The present thesis aims at demonstrating how to scrutinize potential jangle fallacies in a theory-driven, deductive, and falsificationist way. Using the example of the common core of aversive traits, D, I discuss the ways one can find and test differences between more or less overlapping, competing constructs. Specifically, the first paper tests the plausibility of a potential jangle fallacy with respect to D and a Fast Life History Strategy, concluding that the latter is unlikely to represent the common core of aversive traits at all. The remaining three papers test the distinctness of D from FFM Agreeableness, HEXACO Honesty–Humility, and a blend of the two, AG+, all of which are conceptually and empirically remarkably similar to, but could nevertheless be dissociated from D, thereby also refuting an instance of the jangle fallacy.

Although research often places emphasis on similarities, it is impossible to conclusively prove the equivalence of constructs. I therefore conclude that a falsificationist approach is more informative in that it allows to test whether any differences identified on a conceptual level can be confirmed empirically. Stated differently, if a new construct is dissociable both theoretically and empirically, one may assume that it is functionally distinct and no instance of the jangle fallacy.
ZUSAMMENFASSUNG


Auch wenn Forschung sich oft auf Ähnlichkeiten fokussiert hat, ist es nicht möglich, die Äquivalenz von Konstrukten endgültig und umfassend zu beweisen. Daher ist der falsifizierende Ansatz informativer, anhand dessen sich zeigen lässt, wie und wie stark sich Konstrukte voneinander unterscheiden und dass keine jangle fallacy vorliegt.
Content

Preface v
Abstract vi
Zusammenfassung vii
1 Introduction and theoretical background 1
  1.1 Aims of the present thesis 6
2 D vs FLHS 9
3 D vs AG 12
4 D vs HH 16
5 D vs AG+ 20
6 General Discussion and conclusions 23
  6.1 Conclusion 27
References 28
Acknowledgments 40
Declaration of Originality 42
Curriculum Vitae 44
Appendix: Copies of Articles 48
1 Introduction and Theoretical Background

“The first step in wisdom is to know the things themselves; this consists in having a true idea of the objects; objects are distinguished and known by classifying them methodically and giving them appropriate names. Therefore, classification and naming will be the foundation of our science.” (Carolus Linnaeus, Systema Naturae, 1738)

Why are some people generally more talkative than others? Why is a person who acts impulsively in one situation more likely to act impulsively in another situation? Why does a person who tends to trust others easily also tend to cooperate rather than argue with others? And why is a person who is highly interested in art as a teen likely to still be interested in art as an adult? From the viewpoint of personality psychology, the answers to these questions lie in personality traits. Traits are “relatively enduring patterns of thoughts, feelings and behaviors that reflect the tendency to respond in certain ways under certain circumstances” (Roberts, 2009, p. 7). In other words, a trait is manifested in several behaviors (and thoughts and feelings) that typically co-occur in individuals and that occur over a longer period of time and across different situations. As such, traits are used as relatively parsimonious descriptions of how individuals differ from each other in the ways they think, feel, and behave (Fleeson & Jayawickreme, 2018). Thus, a rather extraverted person is not only more talkative, but also tends to be more energetic and assertive than others, and likely shows these tendencies whenever she has the chance to (e.g., when she is with other people) across a good part of her lifespan.

However, traits are latent constructs. As such, they are merely “ideas that unite phenomena under a single term” (Bollen, 1989, p. 180, emphasis added). Thus, in contrast to observable entities studied in other sciences (such as matter in physics, or chemical substances in chemistry), the objects of investigation in personality psychology, traits, are not directly observable and need to be inferred from observable variables. To stick to the example above, extraversion can be observed in—and thus inferred from—its manifestations, e.g., being outgoing, lively, and talkative in social endeavors (Ashton & Lee,
2007). In other words, in order to study a personality trait it needs to be made observable through an operational definition.

However, before researchers can agree on an operational definition, they need to have a shared idea of the trait and it needs to be unambiguously clear at an abstract level which thoughts, feelings, and behaviors are within the scope of the trait (i.e., how the trait is defined formally). In other words, one needs to define exactly which behaviors, thoughts, and feelings are supposed to reflect the trait, so that one can decide through which instruments and in which contexts the trait can be observed (i.e., how the trait is defined operationally; Flake et al., 2017; Scheel et al., 2020). Indeed, proper definitions specifying the contents and limits of a research subject are foundational and ensure common ground on which to build and test theories, in psychology as much as in any other science (Wacker, 2004).

Equally important is a mutual and precise terminology. Not only do researchers need to agree on a definition for a given entity (e.g., a trait), but also on the label they assign this entity. Specifically, one trait should be given only one label and, in turn, the same label ought not to be used for another trait (Gonzalez et al., 2020; Leising et al., 2022). Otherwise, one may fall prey to what has become known as jingle and jangle fallacies, with jingle denoting the illusion that two traits must be equivalent due to similar names, despite them being distinct; and jangle, in turn, denoting the illusion that two traits must be distinct due to different names, despite them being essentially equivalent (Kelley, 1927).

Although both fallacies were already described almost a century ago, they can be found across virtually all subdisciplines of psychology to this day (Hodson, 2021). Instances of jingle, i.e., the same name given to distinct phenomena, have been discovered for example in research on depression (Fried, 2017), impulsivity (Whiteside & Lynam, 2001), emotion (Weidman et al., 2017), positive psychology (Hill et al., 2018), self-regulation (Eisenberg et al., 2019), and, not least, several of the Big Five basic personality dimensions (Block, 1995; Hilbig et al., 2016). That is, overlap of constructs in these fields is likely overestimated and therefore attempts of synthesizing research (such as meta-analyses) may entail
biased, seemingly inconsistent or even contradictory empirical evidence simply due to the fallacy of considering distinct entities the same (due to their shared labels).

In turn, instances of jangle, i.e., research on essentially the same construct under different labels, have been recognized in research on grit and conscientiousness (Credé et al., 2017), self-compassion and neuroticism (Pfattheicher et al., 2017), job-satisfaction and organizational commitment (Le et al., 2010), burnout and depression (Bianchi et al., 2021), several leader behaviors (Banks et al., 2018), several knowledge constructs (Alexander et al., 1991), and several health behaviors (Noar & Zimmerman, 2005), to name just a few. Hence, there may be a fragmented literature in these fields, leading to separate research programs that might as well be integrated if the overlap were recognized. Consequently, accumulation of knowledge is impeded because of non-matching terminology. Thus, both fallacies have problematic consequences and, ultimately, hinder scientific progress or at least waste time and resources (Block, 1995; Hagger, 2014).

One field which has especially suffered from fuzzy distinctions between constructs, definitions, and terminology is that of aversive personality traits (often denoted ‘dark traits’) which are used to describe and explain behaviors that are deliberately malevolent (Rogoza et al., 2022), socially inappropriate and prone to evoke interpersonal conflicts (Zeigler-Hill & Marcus, 2016a), such as cheating, violence, tax fraud, and many more. Besides the arguably most prominent examples of such traits, the Dark Triad, consisting of Narcissism, Machiavellianism, and Psychopathy (Furnham et al., 2013; Paulhus & Williams, 2002), more and more aversive traits have been introduced, e.g. Greed (Seuntjens et al., 2015), Amoralism (Knežević, 2003), or Sadism (O’Meara et al., 2011). Specifically, some of these are virtually indistinguishable from each other on a theoretical level (e.g., Sadism, Callousness, and Spitefulness, Buckels et al., 2013; Marcus et al., 2014) resulting in (often substantial) empirical overlap of aversive traits (latent correlations predominantly exceeding \( r = .50 \) and going as high as \( r = .84 \) in the case of Psychopathy and Spitefulness, Moshagen et al., 2018). In other words, there is a remarkable positive manifold among aversive traits.
On the one hand, this positive manifold may indicate that research on aversive personality traits, too, suffers from jangle fallacies and that some differently named traits are actually not distinct but should rather be considered equivalent. On the other hand, it may merely reflect the fact that aversive traits are, by definition, necessarily related given that they all represent socially aversive behavioral tendencies, affect, and thoughts. Consequently, they likely share a common underlying disposition, the “tendency toward ethically, morally, and/or socially questionable behavior” (Moshagen et al., 2018, p. 657). This common core of all aversive traits was recently termed the Dark Factor of Personality (D) and defined 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” (Moshagen et al., 2018, p. 657).

D has been shown to capture large variance portions of aversive traits ($0.18 < \text{ECV} < 0.74$, median $0.54$; Moshagen et al., 2018), thereby removing most of their aversive characteristics. Indeed, most residualized traits predicted no or only little variance in several self-reported aversive criteria (e.g., dominance, insensitivity, aggression; Moshagen et al., 2018) as well as aversive behaviors such as social value orientation (Hilbig, Thielmann, et al., 2022), cheating behavior and dictator game giving (Moshagen et al., 2018), allowing for the conclusion that aversive traits are manifestations of D that are flavored with unique, non-aversive aspects (Bader et al., 2022). Capturing the commonalities of aversive traits may thus be beneficial in some research contexts. Specifically, whenever one is interested in aversive personality in its entirety, there is henceforth no need to consider a wide array of aversive traits to capture aversiveness to its full extent, but one can simply measure the aversive core of personality.

Although the empirical evidence thus supports the conception of D as a common theoretical denominator among aversive traits that may limit further construct inflation, the introduction of D itself may have produced a new case of jangle. Indeed, previous research had already advanced the idea that some aversive traits have a common core and what exactly this may be. Specifically, research has drawn on at least six constructs to understand the commonalities of at least the “Dark Triad”: Callousness (i.e., lack in empathy; Jones & Paulhus, 2010; Paulhus, 2014), Factor 1 of Psychopathy (i.e., interpersonal
manipulation and callousness; Jones & Figueredo, 2013), a Fast Life History Strategy (FLHS; i.e., deficits in self-control; Jonason et al., 2009, 2010, 2012), low FFM Agreeableness (i.e., Jakobwitz & Egan, 2006; Paulhus & Williams, 2002; Stead & Fekken, 2014; Vize et al., 2020), low HEXACO Honesty-Humility (HH; i.e., the willingness to exploit others; Lee & Ashton, 2005), and a blend of the latter two, herein denoted AG+ (Vize et al., 2021).

However, all of these explanations seem incomplete. Specifically, although the array of personality traits linked to socially aversive features has expanded considerably beyond the Dark Triad traits (see further Zeigler-Hill & Marcus, 2016b), the explanations typically considered only the commonalities of the Dark Triad, or, at most, of the Dark Tetrad traits (which additionally comprise Sadism; Buckels et al., 2013). By contrast, D is defined as the core of all aversive traits. Consistently, all these suggested cores lack certain aversive aspects. This is perhaps most obvious for FFM Agreeableness and HEXACO Honesty-Humility. As orthogonal dimensions of basic models of personality structure, they cannot, by definition, cover aversive aspects that are subsumed by other dimensions, such as angry hostility (which is part of Neuroticism in the FFM; John et al., 2008), lack of adherence to moral rules and values (which is part of Conscientiousness; Muris et al., 2017) or empathy (which is part of Emotionality in the HEXACO; Ashton & Lee, 2007). Callousness, in turn, can hardly explain manipulative behaviors, and even in combination—forming Hare’s factor 1 of Psychopathy—they do not appear plausible explanations for sadistic and spiteful behavior (i.e., deriving utility from the very act of harming another, potentially at some own cost). Similarly, impulsivity and lack of self-control (key features identified to make FLHS a candidate for the common aversive core; Jonason & Tost, 2010) can arguably hardly be considered main drivers of sadistic or spiteful behaviors and malevolent behaviors that require strategic planning and the consideration of long-term over short-term benefits (e.g., tax fraud, deceit). In sum, neither of these constructs thus yield a satisfactory answer to the question which aspects aversive traits share at their core.

Although D is conceptualized to remedy these shortcomings and explicitly subsume all kinds of aversive behaviors and traits, there are now at least seven attempts to describe and explain the common
core of aversive traits. In fact, D has shown high correlations with some of the traits previously suggested as representations of the commonalities of aversive traits. For example, it correlates .80 with Honesty-Humility and .69 with FFM Agreeableness (Moshagen et al., 2018). This very picture has prompted other researchers to challenge the distinctness (and thereby novelty and usefulness) of D beyond the already existing constructs (Rose et al., 2022; Vize et al., 2021; Vize & Lynam, 2021). In essence, rather than alleviating some of the existing jangles, D was seen by some as yet another instance of jangle.

Ultimately, then, a formal dissocation between D and alternative notions on the aversive core is necessary. Otherwise, it remains unclear whether D is yet another instance of old wine in new bottles (i.e., jangle fallacy) and does not make a unique contribution to understanding aversive behavior and its dispositional basis or whether D is truly an extension of these previous approaches.

1.1 Aims of the present thesis

Using the specific example of the common core of aversive traits, this thesis thus aims to demonstrate how to scrutinize potential jangle fallacies to help limit further construct inflation in personality psychology. That is, in four empirical articles (and one rejoinder to a comment on one of these articles), which will be discussed in the following chapters, we demonstrate how to test the (non-)equivalence of constructs. Specifically, we test whether D is functionally distinct from other constructs used to describe the commonalities of aversive traits, specifically from a Fast Life History Strategy (Chapter 2), low FFM Agreeableness (Chapter 3), low HEXACO Honesty-Humility (Chapter 4), and finally a blend of the two latter, low AG+ (Chapter 5).

To account for the latent nature of the constructs investigated herein, all analyses were based on structural equation modeling. More specifically, given that traits underlying behaviors and attitudes cannot be directly observed themselves, all variables were estimated as latent factors from their respective indicators. Additionally, the primary interest in the relations between different constructs (rather than individual scores on specific measures) called for latent modeling so as to account for measurement error and thus yield more accurate estimations of the true relations between variables.
The starting points of these projects were slightly different. Specifically, especially the low poles of Agreeableness, Honesty-Humility, and AG+ have not only been shown to be strong predictors of aversive traits, with Honesty-Humility predicting up to 90% of their shared variance (Hodson et al., 2018), but have also showed considerable overlap with D directly (Agreeableness 48% shared variance, Honesty-Humility 64%, Moshagen et al., 2018; AG+ 90%, Vize et al., 2021). Thus, they are not only strong candidates to represent the common core of aversive traits, but also appear very similar to D.

Nonetheless, although their high covariances can be an indicator for potential equivalence, there is no fix lower boundary to gauge whether observed covariance is high enough or too low to indicate unity of two constructs. An additional—and arguably more compelling—criterion is thus nomological consistency (Hilbig et al., 2016; Thielmann & Hilbig, 2019) or external convergent validity (Gonzalez et al., 2020) of two constructs, i.e., “a high degree of convergence in correlational patterns with other variables” (Fiske, 1971, p. 245, as cited in Gonzalez et al., 2020, p. 3). In other words, only if two highly related constructs are also similar in their respective relations with external criterion variables can they be considered functionally equivalent in the sense of accounting for the same behaviorally relevant variance.

Of note, given that one cannot test the relations with all possible criterion variables, equivalence of two constructs can never be conclusively verified; nonetheless, it can be falsified and thus assumed to hold so long as serious deductive attempts at falsification fail (Meehl, 1967; Platt, 1964). Specifically, one needs to discern the constructs on a conceptual level to find criterion variables for which they ought to predict unique variance on theoretical grounds. If the expected unique predictive value is supported empirically, the constructs are functionally distinct. If, however, the expected unique predictive value is not supported empirically, one can more confidently assume functional equivalence of the constructs under consideration (Gonzalez et al., 2020; Hilbig et al., 2016; Thielmann & Hilbig, 2019).

In line with this reasoning, to test equivalence between D and FFM Agreeableness, HEXACO Honesty-Humility, and AG+, respectively, we sought criteria to which D and each given construct were a 1

---

1 It is, of course, possible to model unity and test this model against an unconstrained model. However, with sufficient statistical power even minuscule divergences from unity will appear meaningful (and vice versa).
priori likely to relate differently based on theoretical considerations. These conceptual differences were then tested through sequential latent regression analyses, expecting incremental variance prediction in the selected criteria by D or the respective competitor. Given their strong conceptual and empirical overlap with D, this approach was suitable for these three constructs.

FLHS, in contrast, has previously yielded rather low empirical overlap with aversive traits and, moreover, conceptually only weakly overlaps with D. It is thus a less strong candidate to represent the common core of aversive traits. Nevertheless, FLHS has previously been argued to reflect the underlying disposition of aversive traits. To conclusively avoid jangle, D thus also needs to prove its unique value over and above a FLHS. However, given the different starting point, we chose a different approach in this case. Specifically, we addressed the question to which extent FLHS reflects aversive traits (beyond the Dark Triad traits) and their common core at all. Arguably, if FLHS is not a good proxy for the common aversive core in the first place (and given that D is), there is no further need for a full dissociation analogous to the approach above.

Starting with FLHS, I will discuss these four projects in more detail in the following chapters. As the corresponding articles are appended to this thesis, I will merely give a broad overview rather than reiterate the specifics that can be found in the articles. In the final chapter, I will provide a general discussion and conclusion.
One prevailing take on the common core of the Dark Triad traits is rooted in evolutionary theories. Specifically, authors have argued that a Fast Life History Strategy is responsible for the emergence of aversive behavior (Figueroedo et al., 2006; Gladden et al., 2009; Jonason et al., 2012). Although humans as a species are generally characterized by a Slow Life History Strategy, individuals may still lean more toward the faster end of the continuum (Brumbach et al., 2009), at which an organism attempts to increase fitness by prioritizing the production of many offspring over taking care of one’s offspring (Pianka, 1970). Concomitantly, faster strategists typically have a shorter life expectancy and are thus assumed to prefer short-term over long-term benefits (Kaplan & Gangestad, 2005). As such, their behavior is expected to be driven by future discounting and low self-control, in turn leading to short-term cheating and exploitation. Thus, according to this view, impulsivity and the pursuit of immediate rewards reflect the core of aversive behavior (Jonason et al., 2012).

Indeed, some aversive behaviors have been linked to impulsivity and short-term orientation, e.g., reckless driving (Bader et al., 2022). However, there is no connection to other aversive behaviors/traits (e.g., Sadism, Spitefulness) and traits like Machiavellianism, which are characterized by strategic planning, are downright incompatible with impulsivity and short-term orientation. Moreover, impulsivity was neither related to pro- nor to antisocial behavior in a recent meta-analysis (Thielmann et al., 2020) and moderate to weak correlations of D with different subdimensions of self-reported impulsivity (and essentially no correlation with behavioral measures of impulsivity) indicated that impulsivity is not an integral part of aversive personality (Moshagen et al., submitted). Taken together, impulsivity and short-term orientation are neither necessary nor sufficient for aversive behavior.
Nonetheless, as FLHS has been argued to represent the common disposition underlying aversive traits (Buss, 2009; Jonason et al., 2012), it is just as important to test whether D can be distinguished from it as it is for conceptually closer constructs. Thus, in Horsten, Hilbig, et al. (2022) we aimed to provide more direct and conclusive evidence whether FLHS is indeed the dispositional basis of aversive traits.

However, given the weak conceptual and empirical similarities between D and FLHS, we did not directly test their equivalence but focused on the more fundamental question whether FLHS is a plausible candidate for the common core of aversive traits at all. If it is, we argued, it must show strong relations to a wide range of aversive traits and, crucially, must converge with their common core. To this end, we estimated the bivariate correlations of FLHS (assessed via the Mini-K; Figueredo et al., 2006; Hammerl, 2017) with aversive traits on the one hand, and with their common core on the other hand. The operationalization of the common core was twofold: first, we modeled it as a bi-factor (Reise, 2012) from the single specific aversive traits; second, we assessed it via the D70 which was designed to measure the common aversive core directly (Bader et al., 2021; Moshagen, Zettler, & Hilbig, 2020).

Results showed that correlations of FLHS both with single aversive traits (\(-0.06 < r < 0.46\), median \(r = 0.26\)) and with their common core (around \(r = 0.30\)) were relatively weak. Although there is no consensus on how strongly two constructs need to overlap to be to consider them identical, 9% of shared variance hardly constitute positive evidence. Likewise, the median shared variance of less than 10% with individual aversive traits (as compared to a median of 57% common variance explained in aversive traits by D, Moshagen et al., 2018) seems incompatible with the notion of FLHS explaining much of their shared aversive variance.

We concluded that FLHS is a relatively poor proxy for most aversive traits and does not represent their underlying shared disposition. Instead, FLHS seems to represent impulsivity and future discounting (neither of which is necessary or sufficient for aversive behavior), along with selfishness. By implication

---

2 Although the question of how many aversive traits there are is still not conclusively solved, those chosen herein were previously shown to load on a common core (Moshagen, Zettler, & Hilbig, 2020) and arguably cover a representative range.
FLHS is a poor candidate to reflect what aversive traits share and (thus) clearly distinct from D, in turn refuting the suspicion of a jangle fallacy.

However, as outlined above, FLHS is neither the only construct competing with D, nor a particularly strong contender. Other established traits have been shown to be stronger predictors of aversive traits and are conceptually closer to D than FLHS is. Specifically, previous research has sought the dispositional tendency underlying aversive traits among the major personality dimensions as defined within lexical models of personality structure: Agreeableness from the Five-Factor Model (FFM), and Honesty-Humility from the HEXACO model; the third broad personality dimension, AG+, is a blend of the former two.

Compared to FLHS, the theoretical and empirical overlap of Agreeableness, Honesty-Humility, and AG+, respectively, with D is much larger, as is their overlap with at least some aversive traits. Thus, as argued above, an alternative approach was necessary to test their distinctiveness from D. Starting with FFM Agreeableness, the following three chapters will thus describe how D was dissociated from each of these three dimensions by testing whether they empirically diverge in their respective relations to external criteria chosen a priori based on theoretical considerations.
To date, the Five-Factor Model of personality\(^3\) (FFM; McCrae & Costa, 2008) is still one of the most prominent descriptions of personality structure (John et al., 2008). It was obtained by organizing an exhaustive list of everyday English personality adjectives into categories. Specifically, through factor analysis, five largely unrelated categories (i.e., major personality dimensions)—summarizing those adjectives that showed the strongest empirical interrelations—were discovered: Neuroticism, Extraversion, Conscientiousness, Agreeableness, and Openness (McCrae & Costa, 1985). The idea of lexically derived models is that “just as any place on Earth can be specified by the three dimensions of latitude, longitude, and altitude, so anyone’s personality can be characterized in terms of the five dimensions of the FFM” (McCrae & Costa, 2008, p. 274). In other words, by assessing individuals’ scores on each dimension of the FFM, intraindividual differences in personality can be captured in a parsimonious, yet comprehensive way.

However, given that each dimension integrates more narrow traits that tend to co-occur empirically, any lexically derived model merely provides a descriptive trait taxonomy rather than theoretical definitions of the dimensions (Block, 1995; Briggs, 1989; Hilbig & Moshagen, 2020; John et al., 2008). In other words, as there are no a priori hypotheses about which factors should emerge from the factor analysis, the obtained factors are only interpreted and described post hoc. For instance, the factor subsuming cooperation, kindness, generousness, and warmth was labeled Agreeableness (Goldberg, 1992; Saucier, 2002). Thus, being derived in a bottom-up, inductive fashion, i.e., from trait

---

\(^3\) Although there are slight conceptual differences between the classic Big Five model (Goldberg, 1981) and the classic Five-Factor Model, they are commonly treated as interchangeable. Throughout this thesis I will subsume both under the term Five-Factor Model.
adjectives that cluster together, and lacking an a priori theoretical basis, the definition of each dimension of a lexically derived model such as the FFM is tied to its operationalization.

Notably, the Agreeableness dimension caught attention for subsuming several adjectives describing socially aversive aspects of behavior at its low pole, including coldness, selfishness, slyness and distrust (Goldberg, 1992; Saucier, 2002). Therefore, when the Dark Triad traits were first suspected to share a common core, their commonalities were initially ascribed to low FFM Agreeableness (Paulhus & Williams, 2002). In principle, locating the common core of aversive traits within a model of personality structure is appealing because, in case of a good match, it obviates the need for an entirely new construct and, thereby, avoids construct inflation.

However, ascribing the shared variance of aversive traits to one single dimension in the FFM is like trying to fit a square peg into a round hole. Specifically, considering Agreeableness as the sole factor explaining the shared variance of aversive traits ignores the fact that other FFM dimensions also subsume aversive characteristics. For example, this applies to angry hostility (which is part of Neuroticism; John et al., 2008) or lack of adherence to moral rules and values (which is part of Conscientiousness; Muris et al., 2017). Indeed, aversive traits have been shown to relate to other FFM dimensions (for a meta-analysis on the Dark Triad traits see O’Boyle et al., 2015). Given that Agreeableness is, by definition, orthogonal to the other FFM dimensions, it cannot account for these aspects and, thus, does not encompass all aversive dispositions.

Nevertheless, the low pole of FFM Agreeableness has repeatedly been argued to represent the shared variance of aversive traits (Jakobwitz & Egan, 2006; Paulhus & Williams, 2002; Stead & Fekken, 2014; Vize et al., 2020). Indeed, the adjectives comprised by low Agreeableness do reflect socially aversive tendencies and thereby make it a plausible candidate for the common core of aversive traits. It can thus be (and later has been) questioned whether D is more than (another name for) FFM Agreeableness (i.e., an instance of the jangle fallacy; Rose et al., 2022; Vize et al., 2021).
To compare Agreeableness and D on a conceptual level and to identify key differences between them, in Moshagen, Zettler, Horsten, et al. (2020) we relied on a comparatively broad description of Agreeableness put forward by Graziano & Tobin (2009). Based on those differences, we selected seven criteria (behavioral dishonesty, competitive and dangerous worldviews, empathic concern, guilt proneness, internet trolling, and stereotypical sexualized behaviors) that were particularly reflective of the core aspects contained in the definition of D but not (as clearly) of Agreeableness.

In order to represent Agreeableness adequately and not miss out on important content by relying on a single operationalization, we measured it using five established inventories and modeled it specifying a bi-factor structure with the general factor representing the gist of all included Agreeableness scales (accounting for 72% of the common variance among all included items) and the specific factors representing the variance specific to each inventory.

Despite considerable empirical overlap ($r = -.64$), D indeed predicted incremental variance ($0.04 < \Delta R^2 < .29$, median $\Delta R^2 = .07$) beyond Agreeableness in all but one of the criteria. Even though we had originally not intended a double dissociation, our results indicated that, in fact, Agreeableness explains incremental variance in empathic concern beyond D ($\Delta R^2 = .39$). However, our research was not designed to explore the extent to which Agreeableness carries behaviorally relevant meaning beyond D (only vice versa), so this particular finding must be considered exploratory.\(^4\) In sum, this study demonstrates that both D and Agreeableness carry meaning that is not covered by the respective other construct and can therefore not be considered equivalent.

Further support for their functional distinctness stems from evidence that D predicted incremental variance beyond Agreeableness in a diverse set of justifying beliefs (Hilbig, Moshagen, et al., 2022), in the weight individuals place on their own versus others’ utility (Hilbig, Thielmann, et al., 2022), as well as in cognitive and behavioral aspects related to aversive behavior (Scholz et al., 2022). Additionally, D has also been shown to correlate with other FFM dimensions (especially Conscientiousness; Moshagen et al., 2022).

\(^4\) A recent study provided further evidence corroborating that Agreeableness, too, carries behaviorally relevant meaning beyond D. Specifically, Agreeableness seems to be a stronger predictor of affective traits (i.e., of callousness and hostility; Scholz et al., 2022), which is in line with the cluster of emotion terms making up this dimension (Tobin et al., 2000).
On the whole, D seems to be theoretically and empirically dissociable from FFM Agreeableness and thus, does not seem to be forming a jangle with Agreeableness. Nevertheless, the FFM is only one of the two most widely used models of personality structure. The second, the HEXACO model of personality, also contains a dimension, Honesty-Humility, which strongly overlaps with D and has previously been claimed to account for the common core of the Dark Triad (Book et al., 2015). Thus, in the third article we aimed at dissociating D from Honesty-Humility. Specifically, overcoming a limitation of the dissociation from FFM Agreeableness, we aimed to demonstrate that both constructs in each comparison explains variance in certain criteria over the other, thereby attempting a double dissociation.
4 D vs HH


After the FFM had been the most prominent tool for personality research for several decades, the beginning of the 21st century witnessed the introduction of a new model of personality structure, the HEXACO (Ashton & Lee, 2007; Lee & Ashton, 2004), which was recovered through lexical analyses across several languages. The most obvious difference to the FFM was the emergence of a sixth major dimension, Honesty-Humility, which summarizes adjectives like fair-minded, faithful/loyal, honest, modest/unassuming, and sincere at the high pole, versus boastful, greedy, hypocritical, pompous, pretentious, and sly at the low pole (Ashton & Lee, 2007, p. 154). It was thus interpreted as "the tendency to be fair and genuine in dealing with others, in the sense of cooperating with others even when one might exploit them without suffering retaliation" (Ashton & Lee, 2007, p. 156). Additionally, Honesty-Humility has proven to be a valuable predictor for variance that is not accounted for (to the same extent) by either single or all FFM dimensions, e.g., egoism (de Vries et al., 2009), Psychopathy (Gaughan et al., 2012), Dictator Game giving (Hilbig et al., 2015), workplace delinquency (Lee, Ashton, et al., 2005; Pletzer et al., 2019), Supernumerary Personality (Lee, Ogunfowora, et al., 2005), trustworthiness (Thalmayer et al., 2011), and, not least, the Dark Triad traits (Book et al., 2015). In sum, the HEXACO

---

5 In the HEXACO, three dimensions (i.e., Extraversion, Openness to Experience, and Conscientiousness) are largely identical to, whereas the remaining two (i.e., Neuroticism and Agreeableness) are slightly rotated variants of their FFM counterparts. Resulting from that rotation, some of the aspects of FFM Agreeableness—specifically fairness, modesty, and sincerity/straightforwardness, the latter of which are part of FFM Agreeableness in some NEO-based operationalizations (Thielmann et al., 2021)—are shifted to the sixth dimension, Honesty-Humility.
seems to represent the structure of personality variation more comprehensively than the FFM (Feher & Vernon, 2021; Thielmann et al., 2021) and includes one dimension (Honesty-Humility) that is particularly close to D conceptually and empirically.

In fact, due to its high predictive validity for the Dark Triad traits—even outperforming FFM Agreeableness—, low Honesty-Humility has repeatedly been argued to be the most comprehensive representation of (and, indeed, redundant with) their empirical overlap, i.e., their common core (Book et al., 2015; Hodson et al., 2018; Lee et al., 2013; Muris et al., 2017). On closer inspection, this is a plausible assumption, too: Honesty-Humility is operationalized through the four facets sincerity, fairness, greed avoidance, and modesty. Accordingly, its low pole is operationally characterized by aversive tendencies like manipulation, cheating, greed, and feelings of superiority (Lee & Ashton, 2004). As such, it indeed bears notable similarities to D: manipulation, greed, and cheating mirror the aspect of utility maximization (with cheating also involving disutility for others), whereas feeling superior and entitled to special privileges is a belief that serves to justify socially aversive behavior. Consequently, the empirical overlap between Honesty-Humility and D is quite strong: In sum, Honesty-Humility seems to be a strong candidate to account for the common core of aversive traits, thus a potential jangle with D.

Although a strong overlap such as the one between Honesty-Humility and D does not necessarily imply unity of the two constructs, it does not leave much room for unique variance, either. Besides the previously described approach of extrinsic convergent validity, we thus pursued a second strategy to dissociate the two constructs in this article (Horsten et al., 2021; see also Horsten, Moshagen, et al., 2022). Specifically, we argued that, if the two constructs were indeed equivalent, they should equally determine the long-term development of aversive traits. Thus, in Study 1, we compared their longitudinal associations with various aversive traits. Results indicated that D longitudinally accounted for incremental variance beyond Honesty-Humility in all nine aversive traits studied ($0.08 < \Delta R^2 < 0.20$). In six of these, Honesty-Humility predicted no unique variance after D was accounted for, but in the remaining three—Agentic Narcissism, Psychological Entitlement, and Self-Interest, all of which are conceptually related to Honesty-Humility—, Honesty-Humility also predicted incremental variance beyond D ($0.10 < \Delta R^2 < 0.26$).
Thus, results of Study 1 provided initial evidence that D and Honesty-Humility are differentially related to aversive traits and thus not fully equivalent.

Nevertheless, we wanted to corroborate these findings by demonstrating how exactly D and Honesty-Humility differ. To this end, we identified conceptual differences between the two constructs that we tested empirically across three more pre-registered studies. Specifically, we argued that D should predict incremental variance beyond Honesty-Humility in justifying beliefs apart from superiority and entitlement (Hilbig, Moshagen, et al., 2022), in sadistic and spiteful behavior that involves utility maximization (in the sense of deriving pleasure from the pain of others) at a cost to oneself, and in callousness (which in the HEXACO model is part of Emotionality and thus, by definition, distinct from Honesty-Humility). In turn, we expected Honesty-Humility to predict incremental variance in the desire for social recognition and admiration (i.e., pretentiousness).

Taken together, results were mostly in line with our hypotheses, with the exception of spiteful behavior, for which only Honesty-Humility predicted unique variance portions, whereas D did not. For two reasons, this is a major limitation of our results. First, Spitefulness was the only trait for which we used a fully behavioral measure. Thus, we only found differences on self-reports, but were not able to demonstrate that D and Honesty-Humility differently affected actual behavior. Second, in the longitudinal design, D was a much stronger predictor of Spitefulness than Honesty-Humility, so it appears inconsistent that we were not able to demonstrate this difference using a behavioral measure of Spitefulness. Future research will need to place further emphasis on the question whether differences between Honesty-Humility and D are also behaviorally relevant and what their respective roles are in predicting Spitefulness in particular.

Nonetheless, although not all theoretically implied differences were supported empirically, our results in summary confirmed that both low Honesty-Humility and D carry psychologically relevant meaning beyond each other. On the whole, we thus concluded that, among basic personality dimensions,
Honesty-Humility may be the best approximation of the commonalities of aversive traits, but that D is nonetheless a distinct construct.

This conclusion has since been corroborated by further research. Specifically, D has been shown to predict unique variance beyond Honesty-Humility in affective and cognitive antagonistic traits (e.g., Hostility, Suspiciousness; Scholz et al., 2022), in further justifying beliefs (e.g., Authoritarianism, Social Dominance Orientation; Hilbig, Moshagen, et al., 2022), in political orientation and voting behavior (Moshagen et al., in prep) and in negative campaigning among candidates running for state elections in Germany (Dian et al., 2023).

After successfully dissociating D from both FFM Agreeableness and Honesty-Humility, the former dissociation was challenged on grounds of a purportedly insufficient representation of Agreeableness (Vize & Lynam, 2021). Thus, the question was whether D could also be dissociated from a broader conceptualization of Agreeableness (effectively resulting in a blend of FFM Agreeableness and HEXACO Honesty-Humility), which is the focus of the following chapter.
5 D vs AG+


Upon publication of our manuscript dissociating D from FFM Agreeableness (Moshagen, Zettler, Horsten, et al., 2020), Vize & Lynam (2021) published a comment raising several methodological concerns about our study. Amongst other points they claimed that our operationalization of Agreeableness fell short of content related to modesty and sincerity/straightforwardness and that our reported differences between Agreeableness and D merely resulted from an “overly narrow and idiosyncratic view of Agreeableness that differs sharply from how the field currently conceptualizes and assesses the domain” (2021, p. 3). Effectively, they argued that our definition and operationalization of Agreeableness resulted in an instance of the jingle fallacy and suggested that D would no longer be dissociable from Agreeableness if the latter were measured more comprehensively.

Specifically, they proposed to assess Agreeableness using a measure introduced by Crowe et al. (2018) and pointed to evidence indicating that, using this operationalization, Agreeableness was actually much more similar to D than we had reported (see also Vize et al., 2021). This measure was compiled of items from several Agreeableness scales and, importantly, from the Honesty-Humility, Agreeableness, and Altruism scales of the HEXACO. Thus, the construct operationalized through this measure is, per se,

---

6 As I outlined previously, there is no theoretical definition of Agreeableness and thus, disagreement may naturally arise as to which aspects this dimension should comprise. Although it is remarkable that personality psychology relies so heavily on the FFM despite it being largely atheoretical, one can assume that through the operational definitions of the dimensions there is a common understanding of the contents of each dimension. Accordingly, given that, for instance, modesty and sincerity/straightforwardness are included in only one measure of Agreeableness, there seems to be agreement that these aspects are at least not integral to this dimension.
broader than FFM Agreeableness in that it also comprises content from other personality dimensions. We therefore referred to it as AG+.

As a consequence of this extremely broad conceptualization, AG+ yields very substantial relations to other personality dimensions (correlating as highly as .57 with Conscientiousness), which we pointed out in a rejoinder to Vize & Lynam’s (2021) comment (Hilbig et al., 2021). Notwithstanding that their conceptualization is thus gravely different from FFM Agreeableness, they labeled the construct assessed by their scale Agreeableness. Thus, ironically, they themselves seem to have produced an instance of the jingle fallacy. In other words, in insisting on having created a measure that optimally represents the lower-order structure of Agreeableness, they actually created a representation of Agreeableness that is incompatible with the Agreeableness construct as it is traditionally understood (an orthogonal dimension within the FFM).

Nonetheless, both FFM Agreeableness and Honesty-Humility already strongly overlapped with D, so one may plausibly entertain the hypothesis that a blend of both, like AG+, shows even stronger overlap or indeed approaches equivalence with D. In fact, D is conceptualized to be fluid and should thus be reflected in any broad enough set of items measuring aversive personality. We thus decided to test whether AG+ is such an item set that is able to reflect D. As before, we argued that a strong overlap as implied by Vize et al. (2021) does not suffice as evidence for purported equivalence but rather that the equivalence assumption needs to hold up against falsification attempts.

To this end, in Horsten et al. (in press) we again derived and pre-registered criteria for which either D or AG+ should predict incremental variance beyond the other, once again aiming for a double dissociation. Based on theoretical considerations, we argued that D should better account for criteria representing the tendency to inflict disutility on others, as well as criteria representing justifying beliefs, and that AG+, in turn, should better account for criteria representing affiliative tendencies. Results strongly supported our hypotheses, with D and AG+ accounting for a median of 60% and 52% unique variance, respectively, in exactly those criteria we had hypothesized them to. Remarkably, shared variance
between D and AG⁺ (34%) was even lower than between D and FFM Agreeableness and thus provided further evidence against unity of the two.

Taken together, our results implied that adding breadth to FFM Agreeableness (resulting in AG⁺) still results in an item set that both misses aspects of and includes aspects beyond the scope of D—not to mention that it clearly does not represent an orthogonal dimension with the FFM and should thus not be labeled Agreeableness (which would itself constitute an instance of the jingle fallacy).
6 GENERAL DISCUSSION AND CONCLUSIONS

Research in personality psychology has witnessed a notable accumulation of new constructs. In many cases, however, purportedly new constructs are not new at all, but conceptually and/or empirically overlap strongly with already existing constructs. Such cases of old wine in new bottles, i.e., investigating known constructs under new labels, are instances of the jangle fallacy (Kelley, 1927). In effect, parallel research programs on the same constructs but under different labels, lead to fractured literatures, waste researchers’ time and efforts and, ultimately, impede the accumulation of knowledge. This arguably undesirable side effect should and can be avoided by scrutinizing the commonalities (vs. differences) between constructs. Unfortunately, this has often been neglected, in turn leading to the “déjà-variable phenomenon” (Hagger, 2014) across many subfields of personality psychology.

Thus, the present thesis aimed at providing an example of how to test the functional distinctness of two constructs. Specifically, in four empirical papers, I tested whether the recently introduced common core of aversive traits, D, can be dissociated from alternative explanations of the commonalities of aversive personality traits. Some research has tried to demonstrate that D is an instance of the jangle fallacy by proving the very equivalence of D with other notions on the common core of aversive traits (Rose et al., 2022; Vize et al., 2020, 2021). However, it is ultimately impossible to verify equivalence; one can only assume equivalence more confidently after critically testing it (Meehl, 1967; Platt, 1964).

To this end, the present thesis employed a variety of methodological approaches, although all followed the overarching goal to critically test the distinctness between D and competing constructs in a theory-driven, deductive, and falsificationist approach. Specifically, a first step was always to compare the two competing constructs’ conceptual definitions, origins, and locations within personality space. Concerning the latter, D can be seen as a blend of several aspects subsumed within several basic personality traits (especially FFM Agreeableness, Conscientiousness, and Neuroticism, as well as
HEXACO Honesty-Humility; Moshagen et al., 2018), whereas FFM Agreeableness and HEXACO Honesty-Humility are roughly orthogonal factors within their respective frameworks.

Concerning the conceptual comparisons, the deductive approach through which Moshagen et al. (2018) derived the notion of D proved very helpful because its definition allows for very clear predictions regarding the scope of D. In turn, the definition and the predictions can be tested quite straightforwardly. Inductively derived constructs, however, such as FFM Agreeableness, HEXACO Honesty-Humility, and AG+, are in large parts defined through their operationalizations, which have been shown to differ quite substantially from one another. Thus, not only does the exact meaning of a construct hinge on the instrument with which it is assessed, but also does the ambiguity regarding its scope make deriving predictions disproportionately less straightforward. Nevertheless, in all cases we were able to identify differences on the conceptual level which we then proceeded to test empirically.

Results indicated that Fast Life History Strategy is too weakly related to aversive traits and their common core to represent the common disposition underlying aversive traits. FFM Agreeableness and HEXACO Honesty-Humility, in turn, showed strong relations to D and may be good proxies for the common core of aversive traits. Nonetheless, they were functionally dissociable from D and, what is more, D accounted for incremental variance in aversive traits beyond them. Finally, AG+, a combination of FFM Agreeableness and HEXACO Honesty-Humility (and further items from the HEXACO model), was not closer to D than either of them separately and, importantly, was also dissociable from D. Again, D accounted for incremental variance in aversive traits beyond AG+. Altogether, i.e., also taking into account the fact that it distorts the FFM, AG+ thus seems an even less suitable proxy for the common core of aversive traits than FFM Agreeableness or HEXACO Honesty-Humility alone. All in all, D does not seem to form an instance of jangle with either of the tested constructs.

One limitation of this thesis is that the majority of our empirical findings are based on self-reports. Thus, results may have been affected by response biases and the similarities of two constructs may have been inflated by common-method variance (Chang et al., 2010). In fact, the similarity between Honesty-
Humility and D was much smaller when D was assessed using a Forced Choice format (Teves, unpublished). However, we have exclusively relied on latent variable modeling which accounts for measurement unreliability (Westfall & Yarkoni, 2016). Furthermore, the instruments used herein to measure D have been shown to be valid and reliable, and their self-observer agreement is comparable to that of other personality inventories (Bader et al., 2021), indicating that the data is likely not excessively distorted. Nevertheless, future research will benefit from explicitly considering actual behavior rather than merely internal processes. Specifically, a more convincing case against the suspicion that D jangles with other constructs would be if conceptual (and via self-report empirically supported) differences in D and the other constructs were also behaviorally relevant. In other words: do individuals who score high on D for example not only report to be more callous but do they actually behave more callously in certain situations than what would be expected for an individual who scores low on Honesty-Humility?

Undeniably, D overlaps strongly with HEXACO Honesty-Humility (about 66% shared variance), FFM Agreeableness (about 50% shared variance), and AG+ (about 34% shared variance). This strong overlap is to be expected as there is no denying that D and all these constructs are highly related. Thus, assuming that D indeed represents the common core of aversive traits, those alternative approaches are clearly quite useful approximations. In light of that, one may well question whether incremental variance prediction on a few selected criteria suffices to conclude that D does not form an instance of jangle with other constructs. Arguably, null vs existing relations or differential relations as indicated by opposite directions of relations would be more convincing evidence and may be aspired to in future research.

However, drawing on an old metaphor, apes and humans also share the majority of their genomes, yet we would not consider them identical. On the contrary, the many shared genomes are responsible for any similarities, but the few genomes on which they differ makes each unique. Similarly, the few criteria to which two otherwise highly overlapping constructs show differential relations are the most informative. Thus, it is crucial to test relations to criteria that are a priori expected to differ, rather than those to just any random criterion variables, possibly even without any theoretical link to the constructs in question (as recently suggested, see Rose et al., 2022).
Nevertheless, one may take the position that such differences are negligible in light of the overwhelming similarities and that the latter are stronger evidence for the constructs’ equivalence than the former are for their distinctness. In fact, some researchers have rather emphatically emphasized the similarities between D and the candidate they promote for the common core of Dark Triad traits (especially with respect to FFM Agreeableness and AG+; Rose et al., 2022; Vize & Lynam, 2021). They have underscored their position predominantly by referring to strong correlations between various measures or composite scores of the Dark Triad traits and measures of “their” candidate, respectively (Book et al., 2015, 2016; Hodson et al., 2018; Jonason & Tost, 2010; Lee et al., 2013; Muris et al., 2017; Rose et al., 2022; Vize et al., 2021), highly similar correlations with external (largely uninformative) criteria (Lee et al., 2013; Rose et al., 2022; Vize et al., 2020, 2021), high profile similarity (Rose et al., 2022; Vize et al., 2020, 2021), or exploratory factor analyses from which only a single factor instead of two separate factors emerged (Vize et al., 2021).

Given that there is a clear cutoff neither for when two constructs are similar enough to be considered identical, nor for when two constructs are different enough to be considered distinct, the two standpoints arguing based on similarities versus differences seem unreconcilable. Stated differently, disagreement on the amount of evidence required to falsify two (or more) constructs’ equivalence will likely lead to advocates of either standpoint declaring the other’s criterion, and consequently conclusions based thereupon, invalid. A potential solution in this conflict may be to regard similarity versus distinctness not as two separate categories but as a continuum (Lawson & Robins, 2021). Specifically, recognizing both similarities and differences offers the chance to understand the degree to which two constructs are similar and the ways in which they are different. However, if one only focuses on the similarities, as previous research has, one will only find similarities. In this regard, the present thesis complements previous research by illuminating the ways one can find and test differences, both on a conceptual and on an empirical level.
6.1 Conclusion

Virtually all sub-fields of psychological research have lately suffered from construct proliferation producing numerous instances of jangle fallacies. The introduction of allegedly new constructs that are essentially similar to well-established constructs, but are endowed with a new label, causes misunderstandings, wastes time and resources, and, ultimately, hinders scientific knowledge accumulation. Drawing on the specific example of the common core of aversive traits, the present thesis thus demonstrates how to test whether a newly introduced construct jangles with already existing, more or less similar, constructs. Though it is tempting to focus on the similarities and claim their equivalence, the more informative approach is to test whether any purported differences hold empirically. Put differently, if the new construct is dissociable both theoretically and empirically, one may assume that it is functionally distinct and no instance of the jangle fallacy.
REFERENCES


https://doi.org/10.1016/j.paid.2015.11.009

https://doi.org/10.1016/j.paid.2014.09.016


https://doi.org/10.1177/0956797613490749

https://doi.org/10.1111/j.1745-6924.2009.01138.x


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

https://doi.org/10.1002/per.733


https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00052


Horsten, L. K., Thielmann, I., Moshagen, M., Zettler, I., Scholz, D., & Hilbig, B. E. (in press). Testing the equivalence of the aversive core of personality and a blend of agreeableness-related items. *Journal of Personality, n/a(n/a).* https://doi.org/10.1111/jopy.12830


Welcome to the jangle: Comparing the empirical profiles of the “dark” factor and Antagonism. https://osf.io/vymbu


Agreeableness at the core of the dark triad of personality. Individual Differences Research, 12(4-A), 131–141.


Comparative validity of Brief to Medium-Length Big Five and Big Six Personality Questionnaires. Psychological Assessment, 23(4), 995–1009. https://doi.org/10.1037/a0024165


ACKNOWLEDGMENTS

This dissertation has been a jungle in which I would definitely have gotten lost without several people who helped me stay on course and find my way.

First of all, my heartfelt thanks go to Benni Hilbig for being a reliable mentor, role model (for better and for worse) and continual source of support and encouragement, for sharing your knowledge and experience, for always putting things in perspective, and for not allowing me to quit. I am beyond grateful to have spent the past years working in the exceptionally supportive, cooperative, inspiring, and, not least, fun atmosphere that you create in our lab.

On that note, I would also like to thank my colleagues of the Experimental Psychology & Personality Lab for stimulating discussions (sometimes even work related), sharing your expertise and being virtual office mates in pandemic times. Thank you to our Hiwis for helping out with all the things, including escalate lab festivities. Special thanks goes to Güli for keeping me sane, especially when it comes to having to deal with administrative stuff.

Also, thank you Ingo and Morten. Although you were not formally engaged as my supervisors, it certainly felt like I could always rely on you for methodological and mental support and I have learned a lot from working together with you. I would not want to miss the many times of returning to my inbox only to find a gazillion new emails going back and forth on some topic or other. Although they admittedly caused quite some anxiety at first, they never failed to cheer me up.

I am incredibly grateful to have been part of the SMiP group, from which I have benefited both on a professional and on a personal level. Thank you to all instructors of the qualification program for sharing your time, knowledge and experience with us. Thank you to my fellow SMiPsters, especially Eileen, Fabiola, Kilian, Lea, and Thomas, for making this journey a less lonesome one. I truly cherish the hotel breakfasts, late after-workshop hours spent in restaurants, pubs, and parks, the alternative retreat in Eußerthal, and countless (non-)academic discussions. Special thanks to Anke and Annette for pulling the strings in the background and for being so passionate about our well-being and success. Not least, thank you to my additional supervisors Tanja and Edgar for your support, helpful feedback as well as new ideas and perspectives.
I count myself incredibly lucky to have friends who were always there to take my mind off work. Anna, Saskia, and Bettina, I am so glad that the comparably short amount of time that we spent together in Landau built a solid enough foundation for our friendship to stay close despite some of us being physically distant. Thank you for endless walks, phone and jitsi calls, and, best of all, personal get-togethers, and for your shared passion for cake food. Cherisse and Jodessa, you have been even further away, but still feel so close after all these years. Thanks for keeping in touch, I miss you tons. Thank you, Steffi and Tamina, for all those afternoon playdates. Though our paths met only recently, you have been a constant source of joy this past year. Thank you, Kevin, for having been a companion and supporter literally all along, from day minus one of our bachelor studies up to today. Eileen, I am grateful for the truth in the saying that you always meet people twice. I am happy that we started the SMiP journey together and, more importantly, that you stayed close even after changing lanes. Thank you, Judith and Julian, for countless game and pizza nights, for all the nonsensical and wise discussions, and for inexhaustible supplies of chips. Thank you, Ina, former partner in crime, now long-distance soulmate, for being an adamant friend who sticks by me through thick and thin. You all are invaluable to me and it wouldn’t have been the same without you.

A final sincere thanks is due to my family. Olga, Jozef, Erika, Christoph, Esther, Marko, Lukas, and Lena, you are the best in-laws I could have gotten stuck with. I am grateful to be part of this family and to know that you are always there for me. Karl und Theresa, there are no words for how much I value having you in my life. Thank you for always taking an active interest in my life and for encouraging my curiosity from the get-go. Passo and Konstantin, you, too, have become an indispensable part of the Horstensens. Thanks for putting up with us. Mama und Papa, thank you for your unconditional love, trust, and support, for giving me the freedom to find out who I am, and for fueling my ambitions. (Papa, see? I made it!) Rabea, thank you for showing me what really matters in life, for making me go outside when all I want to do is curl up on the couch, for being hilarious, for helping me see the world from a new perspective and for giving me no choice but not think of work. Adrian, thank you for your fierce loyalty through thick and thin, for taking me as I am, for making me laugh, for believing in me when I don’t, for always being willing to chip in pizza vouchers as incentives, and simply for being my rock, my home, my best friend.
DECLARATION OF ORIGINALITY

I hereby declare that I have written the present dissertation independently, without assistance from external parties, and without use of other resources than those indicated. The ideas taken directly or indirectly from external sources (including electronic sources) are duly acknowledged in the text. This thesis has not been previously submitted by me (in full or in part) for grading at this or any other academic or governmental institution.

CRediT authorship contribution statements for the jointly authored manuscripts:

   LKH: Conceptualization, Methodology, Formal analysis, Data curation, Writing—original draft.
   BEH: Conceptualization, Methodology, Investigation, Writing—review & editing, Supervision, Funding acquisition. IT: Conceptualization, Resources, Software, Investigation, Data curation, Writing—review and editing. IZ: Conceptualization, Writing—review and editing. MM: Conceptualization, Writing—review and editing.

   MM: Conceptualization, Methodology, Investigation, Formal analysis, Writing—original draft, IZ Conceptualization, Methodology, Writing—review and editing, Funding acquisition. LKH: Conceptualization, Writing—review and editing. BEH: Conceptualization, Methodology, Software, Writing—review and editing.

LKH: Conceptualization, Methodology, Software, Investigation, Formal Analysis, Data curation, Writing—original draft. MM: Conceptualization, Methodology, Writing—review and editing. IZ: Conceptualization, Writing—review and editing. BEH: Conceptualization, Methodology, Writing—review and editing, Supervision.


BEH: Conceptualization, Writing—original draft. MM: Conceptualization, Formal analysis, Writing—review & editing. LKH: Conceptualization, Writing—review and editing. IZ: Conceptualization, Writing—review and editing.


LKH: Conceptualization, Methodology, Software, Investigation, Formal analysis, Data curation, Writing—original draft. IT: Conceptualization, Resources, Writing—review and editing. MM: Conceptualization, Writing—review and editing. IZ: Conceptualization, Writing—review and editing. DS: Conceptualization, Software, Investigation, Writing—review and editing. BEH: Conceptualization, Methodology, Writing—review and editing, Supervision.
CURRICULUM VITAE

LUIZA K. HORSTEN

Address
RPTU Kaiserslautern-Landau
Experimental Psychology & Personality Lab
Fortstr. 7
76829 Landau

e-mail
luisa.horsten@rptu.de

Education
10/2014 – 09/2018 Master of Science in Psychology, University of Koblenz-Landau (final grade: 1,4)
10/2011 – 03/2015 Bachelor of Science in Psychology, University of Koblenz-Landau (final grade: 1,7)
07/2011 Abitur, Städtisches Gymnasium Kreuztal (final grade: 1,5)

Research and academic experience
since 2021 Research Assistant in DFG funded project “Conflict of cooperation and pro-environmental behavior: Social and temporal determinants of choice behavior”
since 2018 Ph.D. Student at the Research Training Group "Statistical Modeling in Psychology" (Supervisor: Prof. Benjamin Hilbig, PhD)
2018 Research Internship, Supervisor: Prof. Benjamin Hilbig, PhD, Cognitive Psychology Lab, University of Koblenz-Landau
2015 – 2018 Student Research Assistant, Cognitive Psychology Lab, University of Koblenz-Landau
2015 – 2018 Student Teaching Assistant, Research Methods of Cognitive Psychology, University of Koblenz-Landau
2013 – 2015 Student Research Assistant, Institute for Communication Psychology and Media
2013 – 2014 Student Teaching Assistant, Empirical Laboratory Course, University of Koblenz-Landau
Awards, grants, and scholarships

2022  
Open Science Award (Gold) of the Open Science Commission, Department of Psychology, University of Koblenz-Landau

2021  
Open Science Award (Gold) of the Open Science Commission, Department of Psychology, University of Koblenz-Landau

2020  
Open Science Award (Gold) of the Open Science Commission, Department of Psychology, University of Koblenz-Landau

2019  
co-recipient of a SIPS Commendation for the Open Science Knowledge Base

2019  
Open Science Award (Gold) of the Open Science Commission, Department of Psychology, University of Koblenz-Landau

2019  
Travel Grant for the Pre-SIPS Replicats Workshop (400 USD), University of Melbourne

2014  
PROMOS Scholarship for an internship in Tangier, Morocco, DAAD (German Academic Exchange Service), University of Koblenz-Landau

Teaching

Winter 2019/2020  
The Dark Factor of Personality  
(lab-tutorial & methods course; level: B.Sc., language: German)

Winter 2020/2021  
The Dark Factor of Personality  
(lab-tutorial & methods course; level: B.Sc., language: German)

Winter 2022/2023  
Introduction to Web-based experimenting (seminar; level: M.Sc., language: German)

Winter 2022/2023  
Web-based programming (seminar; level: M.Sc., language: German)

Supervised Theses

- D versus Honesty-Humility: Gemeinsamkeiten und Unterschiede im Bestrafungsverhalten im uncostly retaliation game
- D-Faktor und Psychotizismus: Ein Vergleich in Bezug auf Risikobereitschaft
- Dark Factor of Personality – Ein Moderator zwischen Framing und Entscheidungsverhalten?
• Der Zusammenhang zwischen den Narzissmus-Dimensionen und dem Dunklen Persönlichkeitskern
• Welchen Einfluss hat die Empathie auf den Zusammenhang zwischen dem D-Faktor und aversiven Verhaltenstendenzen?
• „Alarmstufe rot“ versus „halb so wild“ – Wie die Darstellung des Klimawandels Climate Change Distress beeinflusst

Publications


Conference Contributions

Talks


Posters


Horsten, L., & Rasch, T. (2018). *I am proud of you. ‘Or should we say,’I am proud for you’?* 7th CERE Conference, 4-5 April in Glasgow, UK.


Other

APPENDIX: COPIES OF ARTICLES
Empirical

Fast, but not so Furious. On the Distinctiveness of a Fast Life History Strategy and the Common Core of Aversive Traits

Luisa K. Horsten, Benjamin E. Hilbig, Isabel Thielmann, Ingo Zettler, Morten Moshagen


Received: 2021-06-08 • Accepted: 2021-12-14 • Published (VoR): 2022-05-19

Handling Editor: Ryan Hong, National University of Singapore, Singapore, Singapore

Reviewing: Round 1 - Vlad Burtaverde; Colin Vize. Open reviews are available. [see Index of Supplementary Materials]

Corresponding Author: Luisa K. Horsten, University of Koblenz-Landau, Campus Landau, Department of Psychology, Cognitive Psychology, Fortstr. 7, 76829 Landau, Germany. E-mail: horsten@uni-landau.de

Supplementary Materials: Data, Materials [see Index of Supplementary Materials]

Abstract

Evolutionary Psychology has considered a Fast Life History Strategy (FLHS), denoting an individual’s tendency to invest more resources in proliferation than in child-rearing, to be responsible for the emergence of aversive traits. Empirical evidence for this notion has been inconsistent, however. Herein, we tested whether FLHS is an adequate representation of the underlying disposition of aversive traits (N = 869). To this end, we considered twelve specific aversive traits, and additionally measured and modeled the common core of these traits. We found only weak correlations of FLHS with individual aversive traits as well as with their common core. In sum, the results suggest that the common core of aversive traits is only marginally reflected in FLHS.

Keywords

life history theory, fast life history strategy, dark core of personality, aversive traits

This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International License, CC BY 4.0, which permits unrestricted use, distribution, and reproduction, provided the original work is properly cited.
Relevance Statement

Although a Fast Life History Strategy (FLHS) is related to some aversive traits to some extent, this study suggests that FLHS lacks relevant aspects of what is common to all aversive traits and thus does not adequately represent their underlying disposition.

Key Insights

- FLHS was only weakly related to some aversive traits.
- FLHS shared little variance with common core of aversive traits.
- FLHS was correlated most strongly with self-reported selfishness.
- Relevant aspects of aversive traits are hardly represented in FLHS.
- FLHS does not represent underlying disposition of aversive traits.

Over the last decades, research in personality psychology has increasingly directed attention to personality traits linked to socially aversive and ethically questionable attitudes and behaviors, often denoted as dark traits. Most prominent are arguably the components of the “Dark Triad”, i.e., Machiavellianism, Narcissism, and Psychopathy (Paulhus & Williams, 2002), although there are many other aversive traits such as Spitefulness (Marcus et al., 2014) or Greed (Seuntjens et al., 2015). In light of substantial theoretical and empirical overlap between aversive traits, there is now considerable agreement that socially aversive traits share a common dispositional core (Jonason et al., 2017; Moshagen et al., 2018; Muris et al., 2017; Schreiber & Marcus, 2020; Vize et al., 2020). This common core, termed the Dark Factor of Personality (D), was recently defined 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” (Moshagen et al., 2018, p. 657). Prior research has tried to approximate this common core by established personality constructs, one of which is the focus of the present work.

Specifically, one recurring theme invoked to account for the common basis of aversive traits is Life History Theory (LHT). Jonason et al. (2012), for instance, concluded that the “Dark Triad may indicate a fast life strategy based on immediate rewards and gratification” (p. 193). LHT is a framework originating in evolutionary biology which classifies organisms by how they spend their finite resources to enhance their reproductive fitness. Specifically, LHT locates organisms on a continuum from r- to K-selected (MacArthur & Wilson, 1967). Organisms closer to the r-endpoint mainly invest their resources in mating and producing as many offspring as possible, whereas organisms closer to the K-endpoint mainly invest their resources in their own continued survival as well as in parenting and survival of a small number of offspring (Pianka, 1970). These strategies are considered to be optimized to the circumstances and environment in which an organism lives: unstable environments with higher mortality rates cause more short-term oriented characteristics and behaviors located at the r-endpoint of the continuum (e.g., short gestation times, early reproduction, large litters, low parental investment; Kaplan &
Gangestad, 2005), whereas stable and more persisting environments with lower mortality rates cause more long-term oriented characteristics and behaviors located at the K-endpoint of the continuum (e.g., delayed sexual development, low fertility, high parental investment, high group cohesion; Kaplan & Gangestad, 2005). Accordingly, and referring to the average life spans and reproduction rates of organisms pursuing these strategies, r- and K-selection are often called Fast and Slow Life History Strategy, respectively.

Even though LHT originally referred to differences between species, it has been adapted by evolutionary psychologists to explain individual differences among humans (Figueroedo et al., 2005; Nettle & Frankenhuis, 2020). Like most mammals, humans are generally highly K-selected, but some individuals may nonetheless lean more towards the r-end of the continuum (Brumbach et al., 2009). Characteristic of human life strategy is a comparably long life span, the organization in small, mutually dependent and stable social groups, and a great investment of time and energy in child-rearing. Accordingly, human evolution strongly favored long-term oriented reciprocal altruistic behavior (Boyd & Richerson, 1988; Trivers, 1971). Put differently, humans tend to consider (both positive and negative) future consequences of their social interactions. Taking into account long-term benefits instead of purely pursuing short-term gains in turn leads to mostly cooperative behavior. Nonetheless, in such highly cooperative environments, short-term cheating and exploitation may yield benefits that outweigh the costs resulting from potential punishment and are therefore also expected to develop and persist (Troisi, 2005). Such behavior is facilitated by future discounting and low self-control and has therefore been interpreted as manifestations of a faster Life History Strategy (Jonason & Tost, 2010). In other words, “what is often disparaged as a maladjusted personality marked by impulsivity and lack of self-control instead can be conceptualized within LHT as an adaptive stable strategy” (Buss, 2009, p. 361; see also Dunkel et al., 2013).

Although impulsivity and behaviors directed at instant gratification do represent aspects of some aversive traits, most notably of Psychopathy (Hart et al., 1992; Paulhus & Williams, 2002), these attributes are neither sufficient nor necessary to explain aversive behavior in general. First, whereas impulsivity and the pursuit of immediate rewards may in some cases incur externalities or interfere with others’ needs and may thus be perceived as aversive, they are not socially or ethically aversive attributes per se. For example, impulsively buying an item at the grocery store that was not on the shopping list rarely causes anybody harm and can thus hardly be considered socially aversive. Second, impulsivity and a focus on instant gratification are not universally featured in aversive traits. In fact, short-term thinking is conceptually unrelated to traits such as Sadism (deriving pleasure from the suffering of others; O’Meara et al., 2011) or Moral Disengagement (dismissing ethical standards for oneself; Moore et al., 2012), and is even partially incompatible with Machiavellianism (planful and strategic manipulative behavior; Jones & Paulhus, 2011). Thus, whereas impulsive and short-term oriented behavior may be considered manifestations of a Fast Life History Strategy, it appears
oversimplified to assume that they reflect the common core of (all) aversive traits, in turn questioning whether a Fast Life History Strategy is an adequate approximation of the underlying disposition of aversive traits.

Indeed, the empirical picture does not unanimously support a link between a Fast Life History Strategy and single specific aversive traits and outcomes. On the one hand, a link between aversive traits and Life History traits has been demonstrated by positive correlations between the Dark Triad components and measures of short-term mating (.22 < r < .50; Jonason et al., 2009), by Machiavellianism loading negatively on a latent K-factor (combining several indicators of Life History traits, with low levels indicating a Fast Life History Strategy; Figueredo et al., 2005), or by criminal offending being associated with increased reproductive success and less committed pair bonding compared to a non-criminal lifestyle (Yao et al., 2014). On the other hand, social deviance was not part of a higher-order Life History factor (Brumbach et al., 2009), faster strategists were no less cooperative and did not relate to a more selfish orientation than slower strategists (Wu et al., 2017), and the K-factor correlated more strongly with HEXACO Extraversion and Conscientiousness (rs = .53 and .46, respectively) than with Agreeableness or Honesty-Humility (rs = .24 and .21, respectively; Strouts et al., 2017) — although the latter are known to be of primary and even exclusive relevance for prosocial and ethical behavior (Heck et al., 2018; Thielmann et al., 2020; Zettler et al., 2021). Additionally, there is evidence indicating that different aspects of socially aversive traits are differentially related to Life History Strategy. Specifically, whereas the Impulsive Antisociality facet of Psychopathy and the Entitlement/Exploitativeness facet of Narcissism were negatively correlated with a Slow Life History Strategy, the Fearless Dominance facet of Psychopathy and the Leadership and Grandiose Exhibitionism facets of Narcissism were positively correlated with a Slow Life History Strategy (McDonald et al., 2012). Thus, despite associations between a Fast Life History Strategy and some aversive traits and outcomes, extant evidence is mixed, at best. By implication, there is even less evidence to support the more wide-ranging conclusion that a Fast Life History Strategy may indeed represent the commonalities of all aversive traits.

The present study sought to provide more direct and conclusive evidence on this question, that is, whether a Fast Life History Strategy adequately represents the common dispositional basis of aversive traits. To this end, we considered not only a wide range of (twelve) specific aversive traits, but additionally measured and modeled the common core of these traits and related them to a measure of life history strategy. Specifically, we first approximated the common core of the measured aversive traits via bifactor modeling in which the general factor captures the commonalities among all items used to measure aversive traits (Reise, 2012; see also Moshagen et al., 2018). Secondly, we measured the common core of dark traits directly through a corresponding item set designed specifically to operationalize the underlying dispositional tendency of which all aversive traits are specific manifestations (Bader, Hartung, et al., 2021; Moshagen,
Zettler, & Hilbig, 2020). If a Fast Life History Strategy indeed represents the underlying disposition of aversive traits, it must be substantially related to most, if not all, specific aversive traits and—arguably even more strongly so—to their common core, both when modeled via the single specific aversive traits and when operationalized via an item set designed to measure the common core of these traits directly.

**Method**

The study was not preregistered. Data and analysis scripts are available in the Supplementary Materials. The study was run based on approval by the ethics committee of the University of Koblenz-Landau (#154_2018).

**Measures**

Fast Life History Strategy was assessed using the German translation of the Mini-K (Hammerl, 2017). The 20-item scale covers six dimensions of Life History Strategy (insight, planning, and control; mother/father relationship quality; friend social contact/support; family social contact/support; harm avoidance; community involvement; Figueredo et al., 2006), with lower scores indicating a faster Life History Strategy. Additionally, we measured a total of twelve aversive traits as summarized in Table 1.\(^1\) Finally, as a direct measure of the common core of aversive traits, we used the German D70 (Bader, Horsten, et al., 2021; Moshagen, Zettler, & Hilbig, 2020). All items were rated on a 5-point Likert Scale (1 = “strongly disagree” to 5 = “strongly agree”).

\(^1\) Although there is no consensus on which traits ought to be considered ‘aversive’, we relied on these twelve traits because they arguably represent a comprehensive array of aversive traits (as compared to the so-called Dark Triad or Dark Tetrad most commonly considered in this context) and have been shown to load on a common aversive core (Moshagen, Zettler, & Hilbig, 2020).
### Table 1
Overview of Included Aversive Traits and Corresponding Inventories

<table>
<thead>
<tr>
<th>Trait</th>
<th>Scale</th>
<th>Number of items</th>
<th>Definition</th>
<th>Sample item</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greed</td>
<td>Dispositional Greed Scale</td>
<td>7</td>
<td>“the desire to acquire more and the dissatisfaction of never having enough” (p. 522)</td>
<td>One can never have too much money.</td>
<td>Seuntjens et al., 2015</td>
</tr>
<tr>
<td>Machiavellianism</td>
<td>German Short Dark Triad</td>
<td>9</td>
<td>“self-interest and tendencies toward deceptiveness, exploitation and manipulation of others, cynical perspective on life and interpersonal relationships” (p. 855)</td>
<td>I like to use clever manipulation to get my way.</td>
<td>Malesza et al., 2019</td>
</tr>
<tr>
<td>Narcissism</td>
<td>German Short Dark Triad</td>
<td>9</td>
<td>“self-absorption, dominance, and feelings of entitlement and grandiosity, as well as devaluation of others” (p. 855)</td>
<td>I know that I am special because everyone keeps telling me so.</td>
<td>Malesza et al., 2019</td>
</tr>
<tr>
<td>Psychopathy</td>
<td>German Short Dark Triad</td>
<td>9</td>
<td>“high thrill-seeking, callousness, interpersonal antagonism, manipulation, and anti-social behavioral style” (p. 855)</td>
<td>It’s true that I can be mean to others.</td>
<td>Malesza et al., 2019</td>
</tr>
<tr>
<td>Sadism</td>
<td>Short Sadistic Impulse Scale</td>
<td>10</td>
<td>“a person who humiliates others, shows a longstanding pattern of cruel or demeaning behavior to others, or intentionally inflicts physical, sexual, or psychological pain or suffering on others in order to assert power and dominance or for pleasure and enjoyment” (p. 523)</td>
<td>Hurting people would be exciting.</td>
<td>O’Meara et al., 2011</td>
</tr>
<tr>
<td>Amoralism</td>
<td>AMR40⁶</td>
<td>13</td>
<td>“amoralism involving brutality” (Knežević, 2008, as cited in Paulhus &amp; Jones, 2015, p. 587)</td>
<td>It is all the same to me how people around me feel, If I am enjoying myself.</td>
<td>Knežević, 2003</td>
</tr>
<tr>
<td>Crudelia</td>
<td>AMR40⁶</td>
<td>14</td>
<td>“amoralism caused by frustration” (Knežević, 2008, as cited in Paulhus &amp; Jones, 2015, p. 587)</td>
<td>All is fair in love and war.</td>
<td>Knežević, 2003</td>
</tr>
<tr>
<td>Egoism</td>
<td>Egoism Scale</td>
<td>12</td>
<td>“the excessive concern with one’s own pleasure or advantage at the expense of community well-being” (p. 349)</td>
<td>It is hard to get ahead without cutting corners here and there.</td>
<td>Weigel et al., 1999</td>
</tr>
<tr>
<td>Moral Disengagement</td>
<td>Propensity to Morally Disengage Scale</td>
<td>8</td>
<td>cognitively processing decisions and behavior with ethical import in a way that</td>
<td>Considering the way people grossly misrepresent themselves, it’s hardly a</td>
<td>Moore et al., 2012</td>
</tr>
<tr>
<td>Trait</td>
<td>Scale</td>
<td>Number of items</td>
<td>Definition</td>
<td>Sample item</td>
<td>Source</td>
</tr>
<tr>
<td>-----------------------</td>
<td>--------------------------------</td>
<td>-----------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>Psychological Entitlement</td>
<td>Psychological Entitlement Scale</td>
<td>9</td>
<td>allows to behave unethically without feeling distress</td>
<td>sin to inflate your own credentials a bit.</td>
<td>Campbell et al., 2004</td>
</tr>
<tr>
<td>Self-Centeredness</td>
<td>Self-Control Scale, Self-Centeredness Subscale</td>
<td>4</td>
<td>“indifferent, or insensitive to the suffering and needs of others” (Gottfredson &amp; Hirschi, 1990, p.89, as cited in Grasmick et al., 1993)</td>
<td>If things I do upset people, it’s their problem not mine.</td>
<td>Grasmick et al., 1993</td>
</tr>
<tr>
<td>Spitefulness</td>
<td>Spitefulness Scale</td>
<td>17</td>
<td>“a behavior or preference that would harm another but that would also entail harm to oneself. This harm could be social, financial, physical, or an inconvenience” (p. 566)</td>
<td>It is sometimes worth a little suffering on my part to see others receive the punishment they deserve.</td>
<td>Marcus et al., 2014</td>
</tr>
</tbody>
</table>

*An ad-hoc translation was used.*

### Participants and Procedures

Data for this study were collected as part of the Prosocial Personality Project (PPP), a large-scale web-based study involving six measurement occasions for the base project and several follow-up assessments. All data was collected via a German online panel provider (Respondi). A detailed documentation of the project including verbatim items of all constructs assessed as well as more detailed information on sample composition, sample sizes at each measurement occasion and exclusion criteria is available on the OSF (https://osf.io/m2abp/).

D70 was assessed at T1; Greed, Machiavellianism, Narcissism, Psychopathy, and Sadism, were assessed at T3 (61 days after T1 on average) of the base project. Life History Strategy (Mini-K), in turn, was assessed at follow-up 2020-05a (171 days after T1 on average); Crudelia, Frustralia, Egoism, Moral Disengagement, Psychological Entitlement, Self-Centeredness, and Spitefulness were assessed at follow-up 2020-05b (167 days after T1 on average). The order of scales was randomized within each measurement occasion. Moreover, at each measurement occasion, two attention check items were embedded within the scales (e.g., “Please select ‘strongly disagree’ here. This serves to check your attention.”).

The final sample for this study consisted of 869 participants (46% female, aged 18 to 66 years, $M = 44.0$, $SD = 12.5$; all demographics measured at T1) who provided valid

---

2) Besides the D70, the data reported herein have not been published before. For other publications that were based on data from the PPP, please see the project’s documentation on the OSF at https://osf.io/m2abp/.
answers for the D70 and the Mini-K, passed both attention checks at each measurement occasion, and did not provide invalid answers to more than 50% of the scales at a respective measurement occasion. According to the general a priori exclusion criteria defined for the PPP, responses to a given scale were defined as invalid if we suspected inattentive response behavior on that scale (based on response times of less than 2 seconds per item on average and/or very low variation, i.e., SD < 0.2 at T1 and SD = 0 at all further measurement occasions). Calculations for bivariate correlations are based only on those participants who provided complete and valid answers on the respective criterion scale (see Table 2 for specific ns).

Table 2

<table>
<thead>
<tr>
<th>Trait</th>
<th>n</th>
<th>M</th>
<th>SD</th>
<th>α</th>
<th>ω_u</th>
<th>K [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>869</td>
<td>3.53</td>
<td>0.47</td>
<td>.89</td>
<td>.89</td>
<td>-.24 [-.33; -.15]</td>
</tr>
<tr>
<td>Spitefulness</td>
<td>866</td>
<td>1.79</td>
<td>0.54</td>
<td>.89</td>
<td>.89</td>
<td>-.32 [-.40; -.23]</td>
</tr>
<tr>
<td>Egoism</td>
<td>862</td>
<td>2.46</td>
<td>0.62</td>
<td>.85</td>
<td>.85</td>
<td>-.13 [-.23; -.03]</td>
</tr>
<tr>
<td>Psychological Entitlement</td>
<td>865</td>
<td>2.66</td>
<td>0.66</td>
<td>.87</td>
<td>.87</td>
<td>-.13 [-.23; -.04]</td>
</tr>
<tr>
<td>Moral Disengagement</td>
<td>866</td>
<td>1.95</td>
<td>0.60</td>
<td>.81</td>
<td>.81</td>
<td>-.30 [-.40; -.21]</td>
</tr>
<tr>
<td>Self-centeredness</td>
<td>867</td>
<td>2.33</td>
<td>0.76</td>
<td>.73</td>
<td>.74</td>
<td>-.13 [-.23; -.04]</td>
</tr>
<tr>
<td>Cruelty</td>
<td>866</td>
<td>1.97</td>
<td>0.57</td>
<td>.88</td>
<td>.90</td>
<td>-.46 [-.55; -.37]</td>
</tr>
<tr>
<td>Frusultralia</td>
<td>865</td>
<td>2.44</td>
<td>0.50</td>
<td>.74</td>
<td>.74</td>
<td>-.30 [-.39; -.20]</td>
</tr>
<tr>
<td>Psychopathy</td>
<td>779</td>
<td>2.06</td>
<td>0.59</td>
<td>.75</td>
<td>.72</td>
<td>-.27 [-.36; -.18]</td>
</tr>
<tr>
<td>Sadism</td>
<td>768</td>
<td>1.47</td>
<td>0.53</td>
<td>.87</td>
<td>.84</td>
<td>-.25 [-.35; -.15]</td>
</tr>
<tr>
<td>Machiavellianism</td>
<td>779</td>
<td>2.93</td>
<td>0.67</td>
<td>.83</td>
<td>.84</td>
<td>-.26 [-.35; -.17]</td>
</tr>
<tr>
<td>Narcissism</td>
<td>778</td>
<td>2.42</td>
<td>0.59</td>
<td>.76</td>
<td>.75</td>
<td>.06 [-.04; .16]</td>
</tr>
<tr>
<td>Greed</td>
<td>770</td>
<td>2.36</td>
<td>0.81</td>
<td>.86</td>
<td>.86</td>
<td>-.14 [-.24; -.05]</td>
</tr>
<tr>
<td>Common core (across 12 specific aversive traits)</td>
<td></td>
<td>.97</td>
<td>.91</td>
<td>.91</td>
<td>.91</td>
<td>-.30 [-.41; -.19]</td>
</tr>
<tr>
<td>Common core (D70)</td>
<td></td>
<td>.95</td>
<td>.91</td>
<td>.91</td>
<td>.91</td>
<td>-.28 [-.37; -.18]</td>
</tr>
</tbody>
</table>

Note. n = number of participants with valid answers on the criterion; α = Cronbach’s alpha; ω_u = unidimensional omega.

Analyses and Results

Hypotheses were tested estimating confirmatory factor analyses with the lavaan package (Rosseel et al., 2019) in R (R Core Team, 2020). Non-normality in the data was accounted for by employing maximum likelihood estimation with robust standard errors and scaled test statistics (as provided by the lavaan package when specifying “MLM” estimation; Satorra & Bentler, 2001). In assessing model fits, we considered the robust root mean squared error of approximation (RMSEA) and the standardized root mean
residual (SRMR) in addition to the chi-square model test, for which the statistical power was so high that even minor misspecifications would lead to a rejection of a model (see Jobst et al., 2021). For transparency, we further report the robust comparative fit index (CFI), although its utility to evaluate the fit of a single model is questionable given its dependence on loading magnitude (Moshagen & Auerswald, 2018).

Following the commonly used approach, Life History Strategy was modeled by specifying a higher-order structure (Richardson et al., 2017). More precisely, we specified six lower-order factors representing the six dimensions of the Mini-K from the respective items. Additionally, we specified a higher-order factor representing K, on which the six lower-order factors loaded. Each factor was assigned a scale by fixing its variance to 1 (which also applies for all other factors). The model fit the data well (according to conventional guidelines; Browne & Cudeck, 1992), $\chi^2(164) = 497, p < .001; \text{RMSEA} = .053, 90\% \text{ CI} [.047; .059], \text{SRMR} = .054, \text{CFI} = .93$. The higher-order omega of the K-factor indicated an acceptable reliability ($\omega_{HO} = .66$), whereas the unidimensional omegas and Cronbach’s alphas for the lower-order factors provided a relatively varied picture, ranging from poor to high reliabilities (.50 < $\omega_u < .93$ and .41 < $\alpha < .93$, respectively).

To estimate the bivariate correlations between Life History Strategy and the individual aversive traits, we specified separate models containing a factor for one of the aversive traits along with the latent K-factor. The reliabilities of all aversive traits were acceptable to high both in terms of Cronbach’s alpha (.73 < $\alpha < .89$) and unidimensional omega (.72 < $\omega_U < .89$; see Table 1). As can be seen in Table 2, the correlations of single aversive traits with K varied greatly, yielding a medium-sized effect on average (median $|r| = .26$). Indeed, K was unrelated to Narcissism and only barely related to three other traits (Psychological Entitlement, Moral Disengagement, and Greed). A strong negative association was only found for one single aversive trait (Crudelia). In other words, individuals with a faster Life History Strategy tended to have higher scores on some traits like Crudelia, Self-Centeredness, and Egoism, whereas their scores on other aversive traits, such as Psychological Entitlement, Moral Disengagement, and Greed, were hardly higher than those of individuals with a slower Life History Strategy.

To further test whether K can approximate the latent common core of all aversive traits, we specified a bifactor model with all aversive trait indicators loading on a general factor and on a specific factor for the individual aversive trait. The general factor in a bifactor model captures the variance shared among all items and thus represents their common core, whereas the specific factors capture the remaining variance shared among the items of a given trait that is not shared with the other traits. In this case, the general factor captures the aversive content shared by the trait indicators and can thus be interpreted as the latent disposition that accounts for individual differences in aversive traits.

3) A single-factor model fit the data considerably worse, $\chi^2(170) = 2,304, p < .001; \text{RMSEA} = .133, 90\% \text{ CI} [.129, .138], \text{SRMR} = .111, \text{CFI} = .53; \Delta \chi^2 = 1156.6, p < .001$, suggesting to retain the higher-order model.
and behavioral tendencies. By contrast, the specific factors capture only the remaining, non-aversive characteristics of the respective traits. As a consequence, they do not represent the original constructs anymore and will hence not be further considered substantively. For identification purposes, the general and specific factors were constrained to mutual orthogonality, which also reflects the fact that they account for non-overlapping portions of variance. This model structure fit the data well, $\chi^2(7,018) = 14,368, p < .001$; RMSEA = .039, 90% CI [.039; .040]; SRMR = .054, CFI = .79. The reliability of the general factor was excellent both in terms of hierarchical omega ($\omega = .90$) and Cronbach’s alpha ($\alpha = .97$). Including the higher-order structure for K resulted in a medium-sized latent negative correlation between K and the general factor of aversive traits, $r = -.30$.

Finally, we considered the association between the K-factor and D as a direct measure of the common core of aversive traits. Following Bader, Hartung, et al. (2021), D was also modeled by specifying a bifactor structure such that all items loaded both on the general factor representing D (i.e., the shared variance among all items) and on one of five specific factors or themes (representing the shared variance among subsets of items that is independent from D). Again, the general and specific factors were constrained to mutual orthogonality. The bifactor model yielded a good fit to the data, $\chi^2(2,275) = 5,907, p < .001$; RMSEA = .047, 90% CI [.046; .049], SRMR = .054, CFI = .79. The reliability of D was excellent both in terms of hierarchical omega ($\omega = .91$) and Cronbach’s alpha ($\alpha = .95$). Almost perfectly in line with the previous findings, the latent bivariate correlation between K and D was negative and medium-sized ($r = -.28$).

**Discussion**

Recent research in personality psychology has come to agree that socially aversive traits share a common dispositional core (Jonason et al., 2017; Moshagen et al., 2018; Muris et al., 2017; Schreiber & Marcus, 2020; Vize et al., 2020). Among other suggestions, it has been presumed that aversive traits signify a Fast Life History Strategy (Buss, 2009; Jonason et al., 2012). According to Life History Theory (LHT), this strategy describes species that maximize their reproductive fitness by high proliferation and little parental efforts (Pianka, 1970). In explaining individual differences within the human species, such a strategy is thought to reflect in the general preference for immediate rewards over long-term benefits or, more broadly speaking, impulsivity, in turn leading to exploitative and otherwise aversive behavior (Buss, 2009; Jonason et al., 2012). Empirical evidence, however, has been inconsistent on the potential link between a Fast Life History Strategy and socially aversive traits, let alone their common core. Thus, the present study strictly

4) The latent correlation between D and the general factor estimated across the aversive traits was $r = .87$. 
tested whether a Fast Life History Strategy indeed reflects aversive traits and strongly represents their common core.

In a large, heterogeneous sample, we found that K was related only to some individual aversive traits, with a maximum of only 22% shared variance (with Crudelia) and a median of 7% across all aversive traits, which is notably less than the shared variance among the latter (median 34%, see Table A1 in the Supplementary Materials). Similarly, shared variance between K and the common core of all aversive traits—both modelled via the individual aversive traits and measured directly—only amounted to around 10%. Thus, whereas individuals characterized by a faster Life History Strategy also tend to be higher on some aversive traits, this association is arguably too weak for K to be an adequate representation of the common underlying disposition of all aversive traits.\(^5\) In fact, any single aversive trait alone constituted a better proxy for their common core than did K (.28 < r < .91, median \(r = .75\); see Table A1 in the Supplementary Materials).\(^6\)

K shared the smallest portion of variance with Narcissism, Moral Disengagement, and Psychological Entitlement, and the largest with Crudelia and Egoism. Although Crudelia is supposed to manifest in sadistic, brutal, and destructive behaviors (Vukosavljevic-Gvozden, Opacic, & Perunicic-Mladenovic, 2015), the items of the respective scale arguably also reflect egoistic as opposed to big-hearted tendencies. Thus, whereas the relation of a Fast Life History Strategy with socially aversive behavior seems to be largely driven by selfishness, other relevant aspects of aversive traits are poorly represented. Most notably, individuals with a faster Life History Strategy neither seem to be driven by convictions regarding their superiority and privileges as motives for exploitative behaviors (as reflected in Psychological Entitlement and Narcissism), nor do they derive utility from the disutility of others (as reflected in Sadism and Spitefulness). Indeed, it is entirely plausible that a Fast Life History Strategy cannot represent these aversive traits, as they are neither driven by impulsiveness—which is suggested to be the main aspect linking Fast Life History Strategy to aversive behaviors—, nor is there an obvious evolutionary advantage to hurting others for mere enjoyment.

Taken together, the findings are compatible with the fact that LHT primarily predicts how a species maximizes its reproductive fitness in light of evolutionary trade-offs. According to this theory, individuals characterized by a Fast Life History Strategy exhibit

---

5) We also verified our results by modeling K—analogously to the common aversive core—as a bifactor structure. The analysis script and results are provided in the Supplementary Materials. In short, although single correlations between K and aversive traits slightly differed from those reported herein, the correlations with both the aversive traits (median \(|r| = .24\), see Table A2 in the Supplementary Materials) and the common core modeled from all aversive traits were of equal magnitude overall. Thus, the substantive conclusion that K is not an adequate approximation of the common aversive core also holds for this modeling approach.

6) Likewise, basic personality dimensions such as Honesty-Humility and Agreeableness have been shown to share substantially more overlap with and thus outperform Life History Strategy in accounting for the commonalities of aversive traits (Hodson et al., 2018; Horsten et al., 2021; Moshagen, Zettler, Horsten, et al., 2020; Vize et al., 2020).
various behaviors which are not commonly regarded as “dark” in the sense of ethically or morally aversive (e.g., early sexual intercourse, non-use of birth control, having multiple sexual partners or being an absent parent) and would thus be beyond the scope of a common aversive core. Moreover, to explain why a faster Life History Strategy would lead to aversive personality traits and behaviors, auxiliary assumptions about co-occurring traits are necessary (e.g., that a behavioral strategy optimized for short-term relations—for instance, cheating—is caused by absent parents; Gladden et al., 2009).

It should be noted that our conclusions are limited by the specific operationalization of LHT which is purely psychometric in nature and does not assess actual life history traits or the timing of life history events (Copping et al., 2014, 2017; Sear, 2020). As has been argued before, however, organisms are “adaptation executers”, not “fitness maximizers”, meaning that the execution of predicted adaptations (e.g., amount of resources invested in child rearing or own survival) is deemed at least as or even more indicative of a Fast or Slow Life History Strategy than their outcomes (e.g., number of sexual partners and offspring, life expectancy), which are not only influenced by Life History Strategy, but also by environmental conditions (Figueroedo et al., 2014). The Mini-K has been shown to assess such adaptation executions in terms of patterns of resource investment in the major psychosocial areas associated with a slower Life History Strategy (Figueroedo et al., 2017).

A further limitation pertaining to the operationalization of Life History Strategy is that the Mini-K (containing 20 items) is a short-form of the much longer 199-item Arizona Life History Battery (ALHB; Figueredo, 2007). The Mini-K might thus not fully represent the full breadth of Life History Strategy. However, the items of the Mini-K were designed to summarize the content covered by all six dimensions of the ALHB (Figueroedo et al., 2006) and it has been shown to closely converge with the ALHB ($r = .80$, Olderbak et al., 2014; $\rho = .91$, Figueredo et al., 2014), thereby proving an efficient and practical measure of Life History Strategy. Furthermore, given that the Mini-K is nomologically validated and has been widely used to measure Life History Strategy in personality psychology (Figueroedo et al., 2014), practically all prior work regarding a Fast Life History Strategy as the basis of aversive traits was based on this operationalization (or the ALHB, respectively; Figueredo et al., 2006). Thus, at the very least, the present findings imply that a Fast Life History Strategy as measured by the Mini-K is not equivalent to the dispositional basis of aversive traits.

Nonetheless, it has been suggested that the assessment of Life History Strategy should not solely rely on a psychometric approach but also take into account biographic data (Black et al., 2017; Nettle & Frankenhuys, 2020; Sear, 2020). Thus, future research may need to be grounded on a combination of psychometric and biometric data for more conclusive insights on the link between a Fast Life History Strategy and aversive traits and behaviors.
In sum, whereas a Fast Life History Strategy (as measured by the Mini-K) is to some extent related to and thus may constitute a distal antecedent of at least a few specific aversive traits—most likely Crudelia and Egoism—it is a relatively poor proxy for most aversive traits. Correspondingly, it shares only limited variance with the common core of these traits and does not, per se, represent the underlying disposition of all aversive traits.

**Funding:** This research was funded by grants 2277, Research Training Group „Statistical Modeling in Psychology“ (SMiP) and HI 1600/1-2 by the German Research Foundation (DFG).

**Acknowledgments:** The authors have no additional (i.e., non-financial) support to report.

**Competing Interests:** Isabel Thielmann is an editorial board member of the journal.

**Author Contributions:** Luisa K. Horsten—Idea, conceptualization | Design planning | Resource provision (materials, participants, etc.) | Research implementation (software, hardware, etc.) | Data collection | Data management (storage, curation, processing, etc.) | Data analysis | Writing | Funding to conduct the work. Benjamin E. Hilbig—Idea, conceptualization | Design planning | Resource provision (materials, participants, etc.) | Data collection | Writing | Supervision, mentoring | Funding to conduct the work. Isabel Thielmann—Idea, conceptualization | Resource provision (materials, participants, etc.) | Research implementation (software, hardware, etc.) | Data collection | Data management (storage, curation, processing, etc.) | Feedback, revisions. Ingo Zettler—Idea, conceptualization | Feedback, revisions. Morten Moshagen—Idea, conceptualization | Feedback, revisions.

**Ethics Approval:** The study was approved by the local ethics committee of the University of Koblenz-Landau (#154_2018).

**Data Availability:** For this article, data is freely available (for access, see Index of Supplementary Materials below).

**Supplementary Materials**

For this article the following Supplementary Materials are available via the PsychArchives repository (for access see Index of Supplementary Materials below):

- Dataset including demographics and all reported variables
- Data from the Prosocial Personality Project (PPP)
- Codebook to map variables in dataset with items and constructs
- Code to reproduce all results reported in the manuscript
- Code to reproduce additional results modeling K as bifactor
- Verbatim items and instructions as used in the study
- Additional tables A1 and A2
- Open peer-reviews
Index of Supplementary Materials

Horsten, L. K., Hilbig, B. E., Thielmann, I., Zettler, I., & Moshagen, M. (2022). Supplementary materials to "Fast, but not so furious. on the distinctiveness of a fast life history strategy and the common core of aversive traits" [Data, codebook]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.5649


Personality Science. (Ed.). (2022). Supplementary materials to "Fast, but not so furious. on the distinctiveness of a fast life history strategy and the common core of aversive traits" [Open peer-review]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.5640


References


FLHS and the Common Core of Averse Traits 16

Gesellschaft für Soziologie, Bamberg, Germany.


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

Morten Moshagen a,⇑, Ingo Zettler b, Luisa K. Horstenc, 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.

© 2020 Elsevier Inc. All rights reserved.

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
components (Furnham, Richards, Rangel, & Jones, 2014; Muris, correspondingly, empirical evidence concerning both, the Dark Triad et al., 2017) 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 “dismissing” 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 Fuyt, De Bolle, & De Ciercq, 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. 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 rein- 

vited to complete the second part of the study, which yielded N = 958 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).
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 IPPT (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, Ścigala, & 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, du Toit, & 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 − β > 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.
Table 1
Latent correlations and internal consistencies.

<table>
<thead>
<tr>
<th>Variable</th>
<th>α</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Agreeableness</td>
<td>0.96</td>
<td>-0.64</td>
<td>-0.64</td>
<td>-0.68</td>
<td>0.87</td>
<td>0.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 D</td>
<td>0.95</td>
<td>-0.56</td>
<td>0.89</td>
<td>0.93</td>
<td>0.77</td>
<td>0.72</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>3 BFI-A</td>
<td>0.84</td>
<td>-0.74</td>
<td>0.93</td>
<td>0.71</td>
<td>0.44</td>
<td>0.45</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>4 BFAS-A</td>
<td>0.88</td>
<td>-0.39</td>
<td>0.71</td>
<td>0.44</td>
<td>0.45</td>
<td>0.39</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td>5 IPIP-A</td>
<td>0.92</td>
<td>-0.39</td>
<td>0.71</td>
<td>0.44</td>
<td>0.45</td>
<td>0.39</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td>6 NEO-A</td>
<td>0.81</td>
<td>-0.39</td>
<td>0.71</td>
<td>0.44</td>
<td>0.45</td>
<td>0.39</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td>7 HEXACO-A</td>
<td>0.82</td>
<td>-0.17</td>
<td>0.26</td>
<td>-0.12</td>
<td>-0.18</td>
<td>-0.14</td>
<td>-0.16</td>
<td>-0.09</td>
</tr>
<tr>
<td>Behavioral Dishonesty</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Competitive Jungle Beliefs</td>
<td>0.75</td>
<td>-0.62</td>
<td>0.84</td>
<td>-0.60</td>
<td>-0.65</td>
<td>-0.57</td>
<td>-0.62</td>
<td>-0.39</td>
</tr>
<tr>
<td>Dangerous World Beliefs</td>
<td>0.82</td>
<td>-0.09</td>
<td>0.30</td>
<td>-0.19</td>
<td>-0.08</td>
<td>-0.09</td>
<td>-0.21</td>
<td>-0.16</td>
</tr>
<tr>
<td>(Lack of) Empathic Concern</td>
<td>0.85</td>
<td>-0.82</td>
<td>0.55</td>
<td>-0.75</td>
<td>-0.81</td>
<td>-0.81</td>
<td>-0.61</td>
<td>-0.38</td>
</tr>
<tr>
<td>(Lack of) Guilt Proneness</td>
<td>0.78</td>
<td>-0.50</td>
<td>0.62</td>
<td>-0.47</td>
<td>-0.54</td>
<td>-0.47</td>
<td>-0.48</td>
<td>-0.32</td>
</tr>
<tr>
<td>Internet Trolling</td>
<td>0.70</td>
<td>-0.44</td>
<td>0.54</td>
<td>-0.45</td>
<td>-0.46</td>
<td>-0.41</td>
<td>-0.47</td>
<td>-0.25</td>
</tr>
<tr>
<td>Stereotypical Sexualized Behaviors</td>
<td>0.85</td>
<td>-0.02</td>
<td>0.18</td>
<td>-0.11</td>
<td>-0.06</td>
<td>0.03</td>
<td>-0.19</td>
<td>-0.02</td>
</tr>
</tbody>
</table>

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| ≥ 0.09 (and |r| ≥ 0.12 concerning the polychoric estimates, respectively) significantly differ from zero at p < .05.

1 Polychoric correlations estimates.

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.

& 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, χ²(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 (ωₐ = 0.91 and ωᵦ = 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 (r² = 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
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; Thiellmann & 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 ($r = −.17$) and D ($r = .26$), 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 ($r = .82$ vs. $r = −.55$), 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 ≤ $\Delta R^2 ≤ 0.29$), 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 ($RMSE^4$) 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.\(^4\) 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.

\(^a\) The $RMSE$ is the root of the mean squared difference, $RMSE = \sqrt{\sum_k (r_{A k} - r_{D k})^2}$, where $k$ denotes the number of criteria.

\(^b\) 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 $RMSE = 0.069$. The RMSE is the root of the mean squared difference, $RMSE = \sqrt{\sum_k (r_{A k} - r_{D k})^2}$, where $k$ denotes the number of criteria.
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 ≤ ΔR² ≤ 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 (ΔR² = 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) 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/.
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


Theoretical and empirical dissociations between the Dark Factor of Personality and low Honesty-Humility

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

a University of Koblenz-Landau, Forstr. 7, 76829 Landau, Germany
b Ulm University, Albert-Einstein-Allee 47, 89081 Ulm, Germany
c University of Copenhagen, Øster Farimagsgade 2A, 1353 Copenhagen, Denmark

ARTICLE INFO

Keywords:
- Honesty-Humility
- D factor
- Dark Factor of Personality
- Dark traits

ABSTRACT

Recent research suggests that the common core of all aversive traits can be understood through the Dark Factor of Personality (D). Previously, the overlap among aversive traits has also been described as the low pole of HEXACO Honesty-Humility. Relying on longitudinal data and a range of theoretically derived outcome criteria, we test in four studies (total N > 2,500) whether and how D and low Honesty-Humility differ. Although the constructs shared around 66% of variance (meta-analytically aggregated across all studies), they longitudinally differently accounted for diverse aversive traits and showed theoretically meaningful and distinct associations to pretentiousness, distrust-related beliefs, and empathy. These results suggest that D and low Honesty-Humility are best understood as strongly overlapping, yet functionally different and nomologically distinct constructs.

1 Introduction

People sometimes engage in socially and morally questionable or downright malevolent behavior. From the viewpoint of Personality Psychology, this is attributed to socially and ethically aversive (“dark”) traits, with the Dark Triad components Machiavellianism, Narcissism, and Psychopathy being particularly prominent (Furnham, Richards, & Paulhus, 2013; Muris, Merckelbach, Otgaar, & Meijer, 2017; Paulhus & Williams, 2002). Given that aversive traits show substantial theoretical and empirical overlap, consensus has emerged that they share a common core (Jonason, Zeigler-Hill, & Okan, 2017; Muris, Merckelbach, Otgaar, & Meijer, 2017; Schreiber & Marcus, 2020; Vize, Collison, Miller, & Lynam, 2020). Next to other suggestions, this common core has been suggested to reflect the low pole of Honesty-Humility from the HEXACO model of personality structure (Hodson et al., 2018; Lee et al., 2013; Muris et al., 2017).

Honesty-Humility is one of the six basic personality dimensions in the lexically derived HEXACO model of personality structure (Lee & Ashton, 2008). Like any such lexically and thus inductively derived trait dimension, it is defined by the trait-descriptive adjectives that show particularly high loadings on the corresponding factor—adjectives like trustworthy, loyal, and humble versus deceitful, selfish, and pretentious.

in the case of Honesty-Humility. Within each of the broader HEXACO dimensions, the defining adjectives are further subsumed in narrower facets, which in the case of Honesty-Humility are Modesty, Sincerity, Fairness, and Greed Avoidance (Lee & Ashton, 2006). As a consequence of being defined by the co-occurrence of certain adjectives or trait aspects, the verbal definition of the construct itself is essentially a generic summary of these defining adjectives, namely, that Honesty-Humility represents “the tendency to be fair and genuine in dealing with others, in the sense of cooperating with others even when one might exploit them without suffering retaliation” (Ashton & Lee, 2007, p. 156). Accordingly, individuals at the low pole of Honesty-Humility “will flatter others or pretend to like them to obtain favors, […] are willing to gain by cheating or stealing, […] enjoy and display wealth and privilege, […] and consider themselves as superior and entitled to privileges that others do not have” (Ashton & Lee, 2005, p. 1331). Clearly, both the defining aspects of Honesty-Humility and its definition are compatible with the notion that (low) Honesty-Humility overlaps with (the common core of) aversive traits.

Correspondingly, (low) Honesty-Humility was repeatedly shown to be the strongest predictor of the Dark Triad traits out of all basic personality dimensions (with up to 90% shared variance, Hodson et al., 2018), including stronger relations than Agreeableness from the Five-
Factor Personality Model (Book et al., 2016; Howard & Van Zandt, 2020; Moshagen, Hilbig, & Zettler, 2018). Such findings have nourished conclusions that “the Dark Triad latent covariance almost fully overlaps with the low pole of Honesty-Humility” (Hodson et al., 2018, p. 128) and that “the dark triad concept largely is redundant and has little to add to traditional personality models” (Muri et al., 2017, p. 196).

Whereas previous research focused on testing the overlap of the Dark Triad variables with basic personality dimensions, recent research has conceptualized another factor as the common core of—expressis verbis—all aversive traits. That is, Moshagen et al. (2018) introduced the Dark Factor of Personality (D) as the basic disposition underlying all aversive traits, defined 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” (Moshagen et al., 2018, p. 656).

Notably, a central idea underlying the conceptualization of D is that any aversive trait can be understood as a specific, flavored manifestation of D which, in turn, subsumes the malevolent aspects of all aversive traits. More generally speaking, the theoretical conceptualization of D is akin to the g-factor of intelligence and implies that “D is responsible for the commonalities between various traits and thereby represents their common core” (Moshagen et al., 2018, p. 658). Accordingly, the internal structure of D is best represented by a bifactor model (Moshagen et al., 2018) in which D is represented by a general factor on which all observed (aversive) items load and which thus captures their shared variance. Additionally, each item loads on one of five orthogonal specific factors, or themes—Callousness, Deceitfulness, Narcissistic Entitlement, Sadism, and Individuation—which capture the remaining common variance among subsets of items that is not accounted for by D (Bader, Hartung, & et al., 2021). This modeling approach closely maps onto the theoretical conceptualization of D in that the general factor in a bifactor model functions as the prime and direct source of individual differences on the indicator level and represents their common underlying disposition. The specific factors, in turn, reflect themes within D as well as sets of unique aspects beyond the scope of D (Moshagen, Zettler, & Hilbig, 2020).

Clearly, as is apparent from the theoretical definitions, D and low Honesty-Humility share various similarities. On the theoretical level, the aspect of utility maximization in the definition of D is mirrored in the aspects of greed and lack of sincerity in Honesty-Humility. Moreover, the aspect of justifying beliefs in the definition of D is—in part—mirrored in the aspect of lack of modesty in Honesty-Humility. Correspondingly, it is unsurprising that D and Honesty-Humility show substantial empirical associations, sharing up to 64% of their variance (Moshagen et al., 2018).

However, neither the correspondence in some defining aspects nor about two thirds of shared variance are sufficient to conclude that D and low Honesty-Humility essentially represent the same construct.1 Indeed, their respective origins and conceptualizations differ fundamentally with regard to several aspects. First and foremost, Honesty-Humility was inductively derived from lexical studies and is thus tied to a model of basic personality structure. D, by contrast, was deductively derived from the theoretical definitions of aversive trait constructs, disregarding whether and where such aspects are located in models of personality structure. Correspondingly, D is also substantially related to other basic personality dimensions in the HEXACO model from which Honesty-Humility is, by definition, expected to be independent. This holds in particular to HEXACO Agreeableness (r = −0.45), but also Conscientiousness (r = −0.32; Moshagen et al., 2018). Similarly, loadings of Honesty-Humility (λ = −0.69), Agreeableness (λ = −0.39), and Conscientiousness (λ = −0.17) on the common core of the Dark Triad (as an approximation of D) were recently confirmed meta-analytically (Schreiber & Marcus, 2020). These associations are to be expected given that low HEXACO-Agreeableness involves aspects clearly aligned with the definition of D (being ill-tempered, quarrelsome, and vengeful) as does low Conscientiousness, though less prominently so (being irresponsible, delinquent, or disobedient, see Lee & Ashton, 2008). As such, D involves aspects that are defined to be subsumed across basic (HEXACO) dimensions, including dimensions other than Honesty-Humility.

Second, the conceptual differences between D and Honesty-Humility imply several important differences in their respective content: Whereas D explicitly includes all beliefs and attitudes that may serve as justifications for malevolent behaviors (as reflected in items such as “Doing good deeds serves no purpose; it only makes people poor and lazy.”, or “People who get mistreated have usually done something to bring it on themselves.”), low Honesty-Humility is limited to beliefs that express a sense of superiority or entitlement (“I deserve more influence and authority than most other people do.”, “I am special and superior in many ways.”). Although Honesty-Humility may also empirically relate to other beliefs, these are neither included in its theoretical conceptualization, nor in items used to indicate Honesty-Humility, nor in adjective lists Honesty-Humility was derived from (Ashton, Lee, & Boies, 2015). Moreover, low Honesty-Humility places a focus on pursuing and displaying materialistic gains or high social status (“If I knew that I could never get caught, I would be willing to steal a million dollars.”, “I would enjoy being a member of a fancy, high-class casino.”). By comparison, D explicitly extends to any type of utility, such as joy or self-enhancement—even to the extent that such utility may involve costs (“I think about harassing others for enjoyment.”, “If I had the opportunity, then I would gladly pay a small sum of money to see a classmate who I do not like fail his or her final exam.”). Specifically, D covers behavior characterized by deriving utility from the very act of harming others (e.g., sadistic and spiteful behavior) which may actually cost money or reputation and thus seems incompatible with low Honesty-Humility.

In addition, it has been argued that callousness is a prerequisite for aversive traits to emerge or manifest themselves (Jones & Figueredo, 2013; Paulhus, 2014). Indeed, a lack of empathy plays an important role within D, as disregarding potential disutility for others in pursuing one’s own utility is part of its definition (“I feel sorry if things I do upset people”, reverse coded). Correspondingly, previous studies have reported substantial correlations between D and lack of empathic concern (Moshagen et al., 2020; Moshagen, Zettler, Horsten, & Hilbig, 2020). In the HEXACO model, however, empathy is comprised in the sentimentality facet of Emotionalty (e.g., “I feel like crying when I see other people crying”) and is thus theoretically independent of Honesty-Humility (Ashton, Lee, & De Vries, 2014). In conclusion, (lack of) empathy is vital to the concept of D, whereas it lies outside of the theoretical scope of Honesty-Humility.

As implied by these conceptual differences, D and low Honesty-Humility—despite their overlap—may well constitute functionally distinct constructs in the sense that they comprise different behaviorally relevant variance components; each may carry meaning not carried by the other. To some extent, corresponding evidence is already available, given that D explained incremental variance beyond low Honesty-Humility in several aversive outcomes (Hilbig, Thiellmann, Klein, Moshagen, & Zettler, 2021; Moshagen et al., 2018) and vice versa.2 Thus, there are hints that—despite notable overlap—D and low Honesty-Humility are functionally different and comprise meaning not carried by the other.

The criteria to which D and Honesty-Humility were differentially

---

1 Consider, for example, foot length and body weight. They, too, are strongly associated (r = −0.82 for males, r = −0.76 for females; Green, 1961; Grivas, Milas, Arapaki, & Vasilides, 2008) despite obviously measuring different physical entities.

2 We performed a re-analysis of the Moshagen et al. (2018, Study 3) data showing that Honesty-Humility also explains incremental variance beyond D in some criteria (see additional material at https://osf.io/35dh/?view_only=9cf84c5b2934c9351ace6e301263b).
related were, however, only selected to inspect the relation between D and socially aversive outcomes, rather than to explicitly test the distinctiveness between D and low Honesty-Humility. For a more conclusive test of the functional equivalence of D and low Honesty-Humility, it is thus necessary to put forward and test a priori hypotheses about theoretically-implied differences as sketched above. If D and low Honesty-Humility are indeed functionally equivalent, neither will account for unique variance in thus selected criteria.

Based on the above, two steps are needed. As low Honesty-Humility has only been suggested and tested to represent the common core of the Dark Triad so far, it is first necessary to test whether it also accounts for the commonalities of all aversive traits to an equivalent extent as D does. Arguably, if the core of all aversive traits was captured in an already established basic personality dimension (like Honesty-Humility), one ought not to add a novel construct (such as D)—for parsimony and to avoid jangle fallacies. Thus, the first goal of this paper is to test whether D and low Honesty-Humility predict the same aversive traits to a comparable extent: a necessary condition for the assumption that both D and low Honesty-Humility are equivalent representations of the common core of all aversive traits. If, by contrast, the predictions differ, the second necessary step is to test whether D and (low) Honesty-Humility can be empirically dissociated by their theoretically implied differences.

2 Study 1

We re-analyzed data from a previous study (Moshagen et al., 2018, Study 3; Zettler, Moshagen, & Hilbig, 2020) to investigate whether D and (low) Honesty-Humility differently predict aversive traits on a longitudinal basis. If D and low Honesty-Humility are equivalent representations of the common core of all aversive traits, they will equally determine the development of these traits and neither will predict incremental variance over the other in longitudinally accounting for these traits. In a first step, we tested whether D and low Honesty-Humility are correlated with aversive traits to a comparable extent. Importantly, comparing the size of correlations of D and low Honesty-Humility with criteria is only a weak indicator of whether the two constructs are functionally different: Even if the correlations were equal, D and low Honesty-Humility could account for different, non-overlapping parts of variance and thus explain incremental variance. Consequently, after comparing correlations, we conducted sequential latent multiple regression analyses to test whether D and low Honesty-Humility predict incremental variance over the other.

2.1 Methods

2.1.1 Procedure

The study was not pre-registered. We re-analyzed data that was collected for two previous studies investigating D (Moshagen et al., 2018, Study 3; Zettler et al., 2021). More detailed descriptions of measures and procedures can be found in the corresponding publications. Participants were recruited and compensated through a German professional panel provider. Two measurement occasions were realized (in 2014 and in 2018; interval $M = 46.7$, $SD = 0.1$ months). At the first measurement occasion, participants completed nine self-report scales measuring aversive traits and a measure assessing the HEXACO traits, and at the second measurement occasion participants completed the nine self-report scales measuring aversive traits again. Each measurement occasion started with asking participants for informed consent and demographics, followed by the self-report scales, and ended with debriefing.

2.1.2 Measures

Honesty-Humility was assessed using the German 60-item version of the HEXACO Personality Inventory-Revised (HEXACO-60; Moshagen, Hilbig, & Zettler, 2014), consisting of 10 items per dimension. D was assessed using nine scales that measure aversive traits as specified in Moshagen et al. (2018). A short overview is given in Table 1; a more detailed description is available in the aforementioned publication. Note that the results in the original publication showed that two of these nine aversive traits were operationalized such that their aversive components were not fully represented (i.e., Self-Interest and agentic Narcissism’). They will thus necessarily be accounted for by D to a lower extent than the other aversive traits. Both the HEXACO and all trait scales were answered on a five-point Likert scale ranging from 1 = “strongly disagree” to 5 = “strongly agree”.

2.1.3 Participants

The final sample was the same as described in more detail in Zettler et al. (2020). Out of initially 1,261 participants (48% female) in 2014, a final sample of $N = 470$ completed both measurement occasions. At the first measurement occasion, participants were aged 18–65 ($M = 41.6$, $SD = 13.2$) years.

2.2 Data analysis

We tested our hypotheses in R (Version 3.6.3; R Core Team, 2020) based on structural equation modeling using lavaan (Version 0.6.5; Rosseel, Jorgensen, Oberski, Byrnes, Vanbrabant, Savalei, Merkle, Hallquist, Rhemtulla, Katsikatsou, Barendse, & Scharf, 2019). Assuming data are missing at random, we addressed incomplete data at the second measurement occasion by employing full information maximum likelihood estimation. To account for non-normality in the data, we used maximum likelihood estimation with robust Huber-White standard errors and a scaled test statistic that is asymptotically equivalent to the Yuan-Bentler test statistic.

D was estimated using bifactor modeling as described in more detail in Moshagen et al. (2018). That is, D was modeled as the general factor on which each observed item of the nine aversive trait scales loaded. This general factor thus captures the commonalities among all included aversive trait items. Further, we modeled one specific factor for each aversive trait on which each item of the measure of that particular trait loaded. These nine specific factors capture only the remaining covariance among their respective indicators which is not absorbed by D and thus typically yield little variance which does not represent the original construct. They are thus not considered substantively but must nonetheless be included in the measurement model in order to avoid biased estimates of the correlations between D and covariates (Moshagen, 2021). For identification, all correlations among the specific factors as well as between the specific factors and D were fixed to zero.

Additionally, we modeled one latent factor indicating low Honesty-Humility (by reversing the item coding). The basic model thus consisted of eleven latent factors, representing D, low Honesty-Humility, and nine aversive traits residualized for D in 2014, using the item responses at the first measurement occasion. Detailed estimates of factor loadings on the general and specific factors (for this and all further studies reported herein) are provided in the additional materials on the OSF (https://osf.io/35sdh).

3 The Narcissism subscale of the Short Dark Triad predominantly measures grandiosity and authoritativeness, which are core features of Agentic Narcissism (Back et al., 2013; Miller et al., 2016) and as such related to the modesty aspect of Honesty-Humility. It does not, however, measure the aggressive and exploitative Antagonistic Narcissism, which bears the stronger theoretical overlap with D (Moshagen et al., 2018). Analogously, Self-Interest describes utility maximization in socially valued domains, but does not imply causing dissility for another person (Gerbasi & Prentice, 2013). As such, it is beyond the scope of D, but actually well aligned with Honesty-Humility, which includes seeking wealth and status at the lower pole.
Overview of included dark traits and corresponding inventories (Study 1).

For the longitudinal predictions, we altered this basic model for each of the nine aversive traits. Specifically, we added a latent factor for the unresidualized particular trait in 2018 and, crucially, omitted the indicators for that trait from the general factor representing D in 2014 to avoid predictor-criterion contamination. Thus, in the prediction of a particular aversive trait in 2018 by D in 2014, D was modeled without the items of said trait.

We first tested whether D and low Honesty-Humility were correlated with the aversive traits to a comparable extent. To this end, we conducted nested model comparison based on the scaled $\chi^2$-difference (Gonzalez, MacKinnon, & Muniz, 2021) and normalized evidence ratios (ER) computed from weighted BICs (Wagenmakers & Farrell, 2004; Wu, Cheung, & Leung, 2020). ERs quantify the support in favor of the less parsimonious model over the more parsimonious model. The ER ranges from 0 to 1, with $ER = 1$ representing perfect evidence for the less parsimonious model, whereas $ER = 0$ represents no evidence for the less parsimonious model. As an effect size measure for the difference between the correlations, we provide Cohen’s $q$ and the associated Holm-Bonferroni corrected $p$-values according to Williams (1959).

2.3 Results and discussion

Descriptive statistics, internal consistencies of the scales, and inter-correlations are summarized in Tables A1 and A2 on the OSF. Model fit statistics for the base-models were $\chi^2(4,092) = 11,780$, $p < .01$, $RMSEA = 0.04$, $SRMR = 0.06$ for D, and $\chi^2(35) = 707$, $p < .01$, $RMSEA = 0.12$, $SRMR = 0.08$ for low Honesty-Humility, respectively. The latent correlation between D(2014) and low Honesty-Humility(2014) was $r = 0.80$, and thus smaller than unity ($\Delta \chi^2(1) = 17.02$, $p < .01$, $ER > 0.999$).

To evaluate the correlations of the unresidualized aversive traits in 2018 with D(2014) and low Honesty-Humility(2014), respectively, we estimated one model for each aversive trait in which its correlations with D and low Honesty-Humility, respectively, were allowed to vary freely, and one in which they were constrained to be equal. As can be seen in Table 2, D showed significantly stronger correlations in five out of nine cases (Egoism, Moral Disengagement, Machiavellianism, Psychopathy, and Spitefulness), whereas low Honesty-Humility showed stronger correlations to Agentic Narcissism and Self-Interest. The evidence ratios indicated moderate to strong evidence in favor of the less restrictive models (except for Sadism and Psychological Entitlement). On average, the absolute difference between the correlations corresponds to a moderate effect ($q = 0.17$).

To investigate whether either D or low Honesty-Humility predict incremental variance in the aversive traits, we regressed the unresidualized traits in 2018 on both D (again omitting the items of the to-be-predicted trait) and Honesty-Humility in 2014. As can be seen in Table 3, for each aversive trait either D or low Honesty-Humility predicted incremental variance. D explained incremental variance ($\Delta R^2 > 0.05$) in all aversive traits whereas low Honesty-Humility explained incremental variance in Agentic Narcissism, Psychological Entitlement, and Self-Interest. For Agentic Narcissism and Self-Interest, this was expected, given that both arguably due to their specific operationalizations—only show limited saturation in D, thus involving more unique variance beyond D than the other socially aversive traits considered (Moshagen et al., 2018). Psychological Entitlement, in turn, is defined as “a stable and pervasive sense that one deserves more and is entitled to more than others” (Campbell, Bonacci, Shelton, Exline, & Bushman, 2004) which closely relates to the lower pole of Honesty-Humility. In other words, unlike D, which covers a wide range of justifying beliefs (Moshagen et al., 2020), Honesty-Humility addresses entitlement specifically in describing low scorers on the modesty facet as considering themselves “as superior and as entitled to privileges that others do not have” (Lee & Ashton, 2004, p. 334). Thus, it is plausible that low Honesty-Humility captures incremental variance in Agentic Narcissism, Psychological Entitlement, and Self-Interest beyond D.

Across the aversive traits, D accounted on average for substantially more incremental variance ($\Delta R^2 = 0.14$) than low Honesty-Humility ($\Delta R^2 = 0.07$). These results illustrate why it is necessary to consider the explained variance in addition to the mere comparison of bivariate correlations. For example, judging from the correlations alone, D and low Honesty-Humility would appear to be almost functionally equivalent with respect to Sadism. Taking into account the uniquely explained variance, however, demonstrates that D comprises variance relevant for Sadism that is not comprised in low Honesty-Humility.4

For a fairer comparison of D and low Honesty-Humility, we repeatedly reran the analyses with randomly sampled subsets of only ten items loading on D to match the length of the Honesty-Humility scale. Median results were essentially equivalent to those reported above, ruling out the alternative explanation that D covers a broader range of aversive outcomes than low Honesty-Humility merely due to the larger number of items. The corresponding analysis scripts and results are provided on the OSF.

Taken together, neither the extent to which D and low Honesty-Humility are longitudinally associated with aversive traits nor the variance components they uniquely explain in aversive traits are equal. These results thus corroborate that D and low Honesty-Humility are not functionally equivalent in general and that D seems to be the better representation of the core of all aversive traits. Next to this, one can derive additional and more specific theoretically implied differences

---

4 For another clear example in the opposite direction, consider Psychological Entitlement.
If two constructs are functionally equivalent, they must—besides being strongly interrelated—show nomological consistency (Hilbig, Moshagen, & Zettler, 2016; Thielmann & Hilbig, 2019) and extrinsic convergent validity (Gonzalez et al., 2021). Investigating nomological consistency and extrinsic convergent validity is a critical test whether two constructs are associated to external criteria to similar extents and thus embedded within similar nomological nets. With regard to the current investigation, an informative criterion speaking against nomological consistency (and thus equivalence) of D and low Honesty-Humility will be differentially subsumed by or represented in—and, in turn, correlate differently with—the two constructs.

As sketched above, individuals low in Honesty-Humility pursue materialistic gains or high social status. A person high in D, on the other hand, may strive for material possessions, but not necessarily for social admiration. In fact, spiteful and sadistic behavior is often incompatible with seeking admiration. Correspondingly, money-related variables such as Materialism and Conspicuous Consumption have been found to correlate more strongly with Honesty-Humility than with a composite Dark Triad measure (Lee et al., 2013). We therefore expect the desire for social recognition and admiration to be more strongly encompassed by low Honesty-Humility than by D.

Vice versa, whereas spitefulness and sadism are a defining aspect of D, they are not covered by the theoretical definition of Honesty-Humility. Spiteful or sadistic behavior towards other people, possibly at own costs or negative consequences, is well in line with the definition of D, because utility maximization explicitly covers immaterial gains such as feelings of satisfaction one can experience from causing somebody disutility (Moshagen et al., 2018). By comparison, low Honesty-Humility is essentially limited to individualism, that is, those low in Honesty-Humility might accept causing somebody harm in order to body disutility (Moshagen et al., 2018). By comparison, low Honesty-Humility than by D.

Table 2
Unconstrained longitudinal latent correlation coefficients (Study 1).

<table>
<thead>
<tr>
<th>Dark trait 2018</th>
<th>β_D2014 (&lt;DT2014) [95% CI]</th>
<th>β_HH2014 [95% CI]</th>
<th>χ² (D)</th>
<th>χ² (HH)</th>
<th>Δχ²</th>
<th>p (Holm-Bonferroni-corrected)</th>
<th>ER</th>
<th>q</th>
<th>p (Holm-Bonferroni-corrected)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egoism</td>
<td>0.50 (0.47; 0.53)</td>
<td>0.39 (0.38; 0.50)</td>
<td>0.35</td>
<td>&lt;0.001</td>
<td>0.21</td>
<td>0.19</td>
<td>0.01</td>
<td>0.22</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Machiavellianism</td>
<td>0.35 (0.33; 0.38)</td>
<td>0.15 (0.14; 0.16)</td>
<td>0.46</td>
<td>&lt;0.001</td>
<td>0.15</td>
<td>0.14</td>
<td>0.01</td>
<td>0.24</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Moral Disengagement</td>
<td>0.67 (0.64; 0.70)</td>
<td>0.02 (0.01; 0.03)</td>
<td>0.43</td>
<td>&lt;0.001</td>
<td>0.19</td>
<td>0.18</td>
<td>0.01</td>
<td>0.26</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Agenic Narcissism</td>
<td>-0.19 (0.16; 0.21)</td>
<td>0.33 (0.30; 0.36)</td>
<td>0.33</td>
<td>&lt;0.001</td>
<td>0.05</td>
<td>0.33</td>
<td>0.01</td>
<td>0.27</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Psychological Entitlement</td>
<td>0.10 (0.07; 0.13)</td>
<td>0.53 (0.50; 0.56)</td>
<td>0.37</td>
<td>0.012</td>
<td>0.05</td>
<td>0.12</td>
<td>0.01</td>
<td>0.28</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Psychopathy</td>
<td>0.70 (0.67; 0.73)</td>
<td>0.06 (0.04; 0.08)</td>
<td>0.56</td>
<td>&lt;0.001</td>
<td>0.20</td>
<td>0.12</td>
<td>0.01</td>
<td>0.29</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sadism</td>
<td>0.32 (0.29; 0.35)</td>
<td>0.01 (0.00; 0.03)</td>
<td>0.28</td>
<td>&lt;0.001</td>
<td>0.11</td>
<td>0.01</td>
<td>0.01</td>
<td>0.30</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Self-Interest</td>
<td>-0.32 (0.30; 0.35)</td>
<td>0.81 (0.78; 0.84)</td>
<td>0.35</td>
<td>0.027</td>
<td>0.08</td>
<td>0.27</td>
<td>0.01</td>
<td>0.31</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Spitefulness</td>
<td>0.80 (0.78; 0.82)</td>
<td>-0.19 (0.17; 0.21)</td>
<td>0.43</td>
<td>&lt;0.001</td>
<td>0.25</td>
<td>0.01</td>
<td>0.01</td>
<td>0.32</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Note: N = 1,261. Standardized latent linear regression coefficients. D = Dark Factor of Personality, HH = low Honesty-Humility. χ²(D, HH): variance explained by the full model. Δχ²(HH): increase in χ² after adding Honesty-Humility to the model. ΔR²(D): increase in R² after adding D to the model.

between low Honesty-Humility and D. Confirming such differences would extend the (so far preliminary) conclusion that low Honesty-Humility and D are functionally distinct by specifying how exactly the two constructs differ.

3 Study 2
D and Honesty-Humility were assessed as in Study 1. The desire for social recognition was assessed using a German translation of the Unpretentousness Scale from the Six Factor Personality Questionnaire (6FPP; Jackson, Paunonen, & Tremblay, 2000, available via ipip.org). Each of these self-report scales were answered on a five-point-Likert scale ranging from 1 = “strongly disagree” to 5 = “strongly agree”. Spiteful behavior was assessed using a behavioral measure of sadism which is referred to as Sadistic SVO (Moshagen et al., 2020). It consists of 9 tasks that are structurally similar to the Social Value Orientation (SVO) measure which is used to assess social preferences in terms of the weight an individual attaches to their own versus someone else’s outcome (Murphy, Ackermann, & Handgraaf, 2011). Like a regular SVO, each Sadistic SVO task asks the participants to allocate points (worth 5¢ per 100 points, i.e., around $0.70 at the time of data collection) between themselves and an unknown other. In this regard, it is conceptually similar to a dictator game (Forsythe, Horowitz, Savin, & Sefton, 1994). More specifically, the items were adapted to measure the participants’ inclination for spiteful behavior and differed in how many points participants needed to forego in order to reduce the other’s outcome (see Fig. 1 for two examples; the full set of items is available on the OSF). The only motives to explain why an individual would forego own points to reduce the other’s points are competitiveness (i.e., maximizing the difference between one’s own and the other’s outcome) and/or spitefulness (i.e., minimizing the other’s outcome) and thus sadism. Any other motive, by contrast, would lead to a choice towards the opposite end of the continuum (see also Thiellmann, Böhm, Ott, & Hilbig, 2021). Thus, the Sadistic SVO is a suitable measure to assess spiteful behavior. The options were represented by a scale from 1 to 9, with higher scores on this measure representing a higher tendency towards spiteful behavior. To make the measure less extreme overall, we interweaved it with the six original SVO items (which were not included in the analyses). Additionally, for each Sadistic SVO item we created a reverse coded version with the more spiteful options on the left hand of the scale. To each participant, the Sadistic SVO items were presented in random order and direction.

3.1.2 Procedure

The study was pre-registered (https://osf.io/qsvy7) before starting a data collection. Participants were recruited and compensated through a German professional panel provider. The study consisted of two measurement occasions with about three weeks in between (M = 21.0, SD = 5.2 days). Each measurement occasion started with asking participants for informed consent and demographics. At the first measurement occasion, we randomized whether participants first completed the HEXACO-60 or the dark trait scales (which were also presented in random order). At the second measurement occasion, participants completed, again in randomized order, the Unpretentousness Scale and all 15 SVO items. Participants were informed that one of the (sadistic and original) SVO items would be drawn at random and be fully consequential for their own and the other’s bonus payment.7 They were fully debriefed about the purpose of the study after completion of the second measurement occasion.

3.1.3 Participants

In order to estimate the bifactor model, we aimed at 300 complete datasets. A total of 462 participants completed the measures at the first measurement occasion (and passed an attention check item), of which 327 also completed the measures at the second measurement occasion. We excluded 12 participants for inconsistent demographic information between the two measurement occasions and an additional two participants for speedy responses (<2 sec per item). Thus, we achieved a final sample of N = 313. Approximately 55% of the sample were female. The participants were aged between 18 and 65 (M = 41.5, SD = 12.3) years, 63% of them were employed.

3.2 Data analysis

The modeling strategy was largely identical to the one used in Study 1. We used robust standard errors and Satorra-Bentler scaled test statistics to address non-normality. Again, D was estimated using bifactor modeling. Additionally, we modeled one latent low Honesty-Humility factor, one latent Unpretentousness-factor and one latent sadistic SVO-factor based on the respective items. The basic model thus consisted of 13 latent factors, including nine specific aversive traits residualized for D.

3.3 Results and discussion

Descriptive statistics, internal consistencies of the scales, and inter-correlations are summarized in Table A3 on the OSF. Model fit statistics for the base-models were $\chi^2(4,092) = 6,571, p < .01$, RMSEA = 0.05, SRMR = 0.07 for D, and $\chi^2(35) = 184, p < .01$, RMSEA = 0.13, SRMR = 0.09 for low Honesty-Humility, respectively. The latent correlation between D and low Honesty-Humility was $r = 0.87$ and significantly smaller than unity ($\Delta \chi^2(1) = 11.23, p < .01, ER > 0.999$).

To evaluate the latent correlations of Sadistic SVO and Pretentuousness with D and low Honesty-Humility, we estimated (1) one baseline model, $\chi^2(7,013) = 10,939, p < .01$, RMSEA = 0.05, SRMR = 0.07, in which the correlations of Sadistic SVO and Pretentuousness with D and low Honesty-Humility were each allowed to vary freely (so that the criteria were allowed to exhibit different correlations to both D and low Honesty-Humility); one model each in which (2a) either the correlations of Sadistic SVO or (2b) the correlations of Pretentuousness with D and low Honesty-Humility were constrained to be equal (and the respective other allowed to vary freely), and (3) one in which both Sadistic SVO and Pretentuousness were constrained to be correlated equally strongly with D as with low Honesty-Humility.

As can be seen in Table 4, the correlations of Pretentuousness with D and low Honesty-Humility differed significantly. On the other hand, constraining the correlation of Sadistic SVO with D and low Honesty-Humility did not significantly worsen model fit. The model comparisons thus suggested to choose the model with the correlations to Pretentuousness allowed to vary freely and the correlations to Sadistic SVO constrained to be equal to $r = 0.16$ (p = .006; $\chi^2(7,122) = 13,185, p < .01$, RMSEA = 0.05, SRMR = 0.07), which was further supported by the evidence ratios and Cohen’s $q$. This is well in line with our prediction that Pretentuousness is more strongly correlated with low Honesty-Humility than with D. It is not, however, in line with our prediction that Sadistic SVO is more strongly correlated with D than with low Honesty-Humility.

To investigate incremental variance prediction, we regressed Sadistic SVO and Pretentuousness on both D and low Honesty-Humility. This was not preregistered, but given the inconclusiveness of comparable zero-order associations (Gonzalez et al., 2021; Pletzer, Benitvelzen, Oostrom, & De Vries, 2019; Sechrest, 1963), adding a test for incremental variance prediction is crucial as also demonstrated in Study 1. As can be seen in Table 5, neither D nor low Honesty-Humility explained incremental variance in Sadistic SVO. Along with the equal correlations, this suggests that both constructs explain similar portions of variance in this behavioral measure. Notably, this finding is somewhat contrary to the Study 1 findings that D substantially improved the prediction of self-reported Sadism and Spitefulness over low Honesty-Humility. Further, low Honesty-Humility explained more unique variance in Pretentuousness as compared to D. D, however, also explained incremental variance in Pretentuousness, showing that both D and low Honesty-Humility comprise different portions of variance that are relevant for

---

7 The points allocated to the other person were later randomly paid out to participants of an independent study.
4.1 Methods

4.1.1 Measures

We pretested a set of beliefs that would justify aversive behaviors. Specifically, we included constructs that reflect a negative worldview or negative expectations of others. A more detailed description and the results of the pretest are provided on the OSF. Especially distrust-related beliefs (Dangerous and Threatening World View by Sibley & Duckitt, 2009; Trust Scale by Yamagishi, 1986) showed strong relations to D. Thus, in the main study, we assessed these again with higher statistical power, and additionally replaced the Propensity to Trust Scale, which only showed weak associations to both D and low Honesty-Humility, by the IPIP Distrust Scale (Conn & Rieke, 1994), consisting of 10 items. Honesty-Humility was assessed using the corresponding 32 items of the HEXACO-200 (Lee & Ashton, 2004). D was assessed using the D35 (Moshagen et al., 2020).

4.1.2 Procedure

The study was pre-registered before the start of data collection (https://osf.io/epshf/). Through a professionally managed online panel, we recruited and compensated participants from the UK. In the beginning, participants were asked to provide informed consent and demographics. Next, all participants answered the D and Honesty-Humility beyond the scope of Honesty-Humility – in particular those related to general distrust. Hilbig, Kieslich, Henninger, Thiellmann, and Zetlter (2018) showed that the exploitative behavior of those low in Honesty-Humility is driven by temptation rather than distrust or fear of exploitation. D, by contrast, explicitly includes all beliefs that can serve as justifications for utility maximization at others’ costs which explicitly involves distrust-related beliefs. Indeed, D has shown substantial relations to several distrust-related beliefs, including Competitive Jungle and Dangerous Worldviews (Moshagen et al., 2020) and explains about twice as much variance compared to Honesty-Humility in paranoid tendencies (Hilbig et al., 2021).

Table 5

Latent regression results predicting the outcomes by low Honesty-Humility and D (Study 2).

<table>
<thead>
<tr>
<th>Outcome variable</th>
<th>( \beta_{D} ) (95% CI)</th>
<th>( \beta_{HH} ) (95% CI)</th>
<th>( R^{2}_{D} ) (HH)</th>
<th>( \Delta R^{2} ) (HH)</th>
<th>( \Delta R^{2} ) (D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sadistic SVO</td>
<td>0.27 (0.14; 0.37)</td>
<td>-0.08 (-0.14; -0.02)</td>
<td>0.04 (&lt;0.01)</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Pretentiousness</td>
<td>-0.47 (-0.79; -0.15)</td>
<td>1.20 (0.90; 1.49)</td>
<td>0.06</td>
<td>0.31</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Note: \( N = 313. \) Standardized latent regression coefficients, D = Dark Factor of Personality, HH = low Honesty-Humility. \( R^{2}(D, HH) \): variance explained in the full model. \( \Delta R^{2}(HH) \): increase in \( R^{2} \) after adding low Honesty-Humility to the model. \( \Delta R^{2}(D) \): increase in \( R^{2} \) after adding D to the model.

Pretentiousness.

Like in Study 1, we repeatedly reran the analyses with randomly sampled subsets of only ten items loading on D to match the length of the Honesty-Humility scale, which again yielded results essentially equivalent to the reported ones. The corresponding analysis scripts and results are provided on the OSF.
4.1 Participants

Based on the pretest data, we ran a power simulation (Beaujean, 2014), which showed that 500 participants would suffice to achieve desired power (1 – β ≥ 0.80) for the latent correlations between D, low Honesty-Humility, and trust, respectively. A total of 552 participants completed the survey (and passed an attention check item), 44 of which had to be excluded due to speedy responding (< 2 sec per item) or suspicious response styles (i.e., selecting the same response option for > 15 consecutive items). Thus, we included N = 508 valid datasets in our analyses. Approximately 46% of the sample was female. The participants were aged between 18 and 65 (M = 45.7, SD = 11.6) years. 55% of them were employees, 37% held a college/university degree.

4.2 Results and discussion

Descriptive results, internal consistencies and inter-correlations can be found in Table A5 on the OSF. Model fit statistics for the base models were χ²(560) = 1,297, p < .01, RMSEA = 0.06, SRMR = 0.06 for D, and χ²(464) = 2,211, p < .01, RMSEA = 0.10, SRMR = 0.09 for low Honesty-Humility, respectively. The latent correlation between D and low Honesty-Humility was r = 0.82 (p < .001) and significantly smaller than unity (Δχ²(1) = 59.91, p < .001, ER > 0.999).

For each criterion, we estimated a separate model containing the latent factors for D and Honesty-Humility, as well as one for the latent factor which was modeled from the items of the respective criterion. The factors were assigned a scale by fixing their variances to 1. We first considered the latent bivariate correlations of both D and low Honesty-Humility with the justifying beliefs. For each justifying belief, we estimated one model in which its correlations with D and low Honesty-Humility were allowed to vary freely, and one in which they were constrained to be equal. For each justifying belief, we compared the two models by examining the χ²-difference and the ER.

As can be seen in Table 6, all scales were correlated significantly more strongly with D than with low Honesty-Humility. The magnitude of differences averaged at q = 0.13, corresponding to a moderate effect size.

To investigate the incremental variance prediction, we further regressed each justifying belief on both D and low Honesty-Humility (thereby deviating from the preregistration, see above). As can be seen in Table 7, D explained incremental variance over low Honesty-Humility in each of the three justifying beliefs, though neither D nor low Honesty-Humility explained much incremental variance in Threatening World-view. Overall, distrust-related beliefs are thus more strongly subsumed in D as compared to low Honesty-Humility, that is, individuals high in D have stronger negative expectations of the world and others than do those low in Honesty-Humility.

5 Study 4

In Study 3, we corroborated both that low Honesty-Humility comprises meaningful variance beyond D (for pretentiousness), and that the same applies vice versa (for distrust). In a final study, we sought to conceptually replicate this finding by examining the role of callousness, which has been considered to be fundamental for aversive traits (Jones & Figueredo, 2013; Paulhus, 2014) and which is a defining part of D. By contrast, within the HEXACO model, empathy (the opposing pole of callousness) is part of Emotionality (more specifically, its Sentimentality facet; Ashton et al., 2014) and thus theoretically distinct from Honesty-Humility. Thus, D ought to relate more strongly to callousness than low Honesty-Humility.

5.1 Methods

Given that D was measured very broadly in the previous studies and, in some cases, with a larger item set than Honesty-Humility, D was herein assessed using the D16 (Bader, Horsten, Hilbig, Zettler, & Moshagen, 2021; Moshagen et al., 2020). On par, Honesty-Humility was assessed using the 14 items of the corresponding scale in the HEXACO-100 (Lee & Ashton, 2018). For empathy, we assessed both the positive and the negative pole: the positive pole was assessed using the Empathic Concern and Perspective Taking subscales of the Interpersonal Reactivity Index (7 items each; Davis, 1983), whereas the negative pole was assessed using the 14 items of the callousness facet from the Personality Inventory for DSM-5 (Maples et al., 2015; Zimmermann et al., 2014).

The callousness items were answered on a 4-point Likert scale (1 = “completely disagree”, 4 = “completely agree”), all other items were answered on a 5-point Likert scale (1 = “strongly disagree”, 5 = “strongly agree”). We used the respective German translations of each measure.

5.1.2 Procedure

The study was pre-registered (https://osf.io/f4bnu, blinded for peer review) before starting data collection. Participants were recruited and compensated through a German professional panel provider. Each participant first provided informed consent and demographics. The first block of the study consisted of the D and Honesty-Humility scales, the second of the empathy and callousness scales. The order of scales was randomized within each block. Additionally, we embedded one attention check item (e.g., “Please select ‘strongly disagree’ here. This serves to check your attention.”) in each block. After completion, participants were debriefed about the purpose of the study.

5.1.3 Participants

We ran a power analysis in semPower (Moshagen, 2020), aiming to achieve a power of 0.90 based on an alpha error probability of 0.05 to reject the null hypothesis that D predicts no incremental variance over low Honesty-Humility. To this end, we defined a model with regression slopes of 0.50 for D and 0.10 for low HH and a model in which we assumed a slope of zero for D. We then obtained a variance–covariance matrix from the former model and fit the latter model to it to obtain the model-implied variance–covariance matrix. Plugging these two matrices into the semPower.aPriori command revealed a required sample size of N = 204 to detect the assumed effect. However, in order to be able to estimate the structural equation model reliably, we aimed at a final sample size of 500.

A total of 542 participants completed the survey (and passed the attention check items), 57 of which had to be excluded due to speedy responding (< 2 sec per item) or suspicious response styles (showing very low variation, i.e., SD < 0.2, in responses on any of the scales that contain at least 25% reverse-keyed items, i.e., all except Callousness). Thus, we included N = 485 valid datasets in our analyses. Approximately 52% of the sample was female. The participants were aged between 18 and 66 (M = 41.6, SD = 13.1) years, 33% of them held a college/university degree.

5.2 Results and discussion

We used robust standard errors and Satorra-Bentler scaled test statistics to account for non-normality. One latent factor each was modeled...
and Callousness. Model fit statistics were

Note: N = 508. D: Dark Factor of Personality, HH: Honesty-Humility. $\chi^2_q$: scaled log-likelihood ratio test (change in model fit after restricting correlations to be equal). $ER$: normalised evidence ratio comparing the less restricted to the restricted model. $q$: difference in the (absolute) zero-order correlations with an outcome between low Honesty-Humility and D as measured by Cohen’s $q$ with associated (one-sided and Holm-Bonferroni corrected) $p$-values.

for D, low Honesty-Humility, Empathic Concern, Perspective Taking, and Callousness. Model fit statistics were $\chi^2(10.4) = 315, p < .01$, $R^2(D) = 0.611$, $SRMR = 0.06$ for D, and $\chi^2(10.4) = 728, p < .01$, $R^2(D) = 0.62$, $SRMR = 0.10$ for low Honesty-Humility, respectively. The latent correlation between D and low Honesty-Humility was $r = 0.63$ ($p < .001$) and significantly smaller than unity ($\Delta \chi^2(1) = 74.191, p < .001$, $ER > 0.999$). Descriptive results, internal consistencies, and observed inter-correlations for all scales can be found in Table A6 on the OSF.

Again, we first considered the latent bivariate correlations of both D and low Honesty-Humility with the empathy measures. To this end, we specified two separate models for empathy and callousness. More specifically, each model contained the factors for D and low Honesty-Humility, along with either two factors for empathic concern and perspective taking or one factor for callousness. We then first estimated each model allowing the correlations to vary freely. Next, we modified the empathy model such that either the correlation of empathic concern or the correlation of perspective taking with D and Honesty-Humility were constrained to be equal. Given that both restrictions lead to significant decreases in model fit, we did not estimate an additional model in which the correlations of both subscales were constrained. Additionally, we estimated the callousness model with the correlations of callousness with D and low Honesty-Humility constrained to be equal.

Again, the nested models were compared by examining the $\chi^2$-differences and the ERs. As can be seen in Table 8, all scales were correlated significantly more strongly with D than with low Honesty-Humility, corresponding to large effects for empathic concern and callousness, and a medium-sized effect for perspective taking.

To investigate incremental variance prediction, we further regressed each of the three empathy factors on both D and Honesty-Humility. As can be seen in Table 9, D explains substantial incremental variance over Honesty-Humility in all three criteria, while Honesty-Humility practically explains no unique variance. Thus, in line with theory and our predictions, empathy is more strongly subsumed in D as compared to

<table>
<thead>
<tr>
<th>Outcome variable</th>
<th>$\beta_D$ [95% CI]</th>
<th>$\beta_{HH}$ [95% CI]</th>
<th>$R^2(D, HH)$</th>
<th>$\Delta R^2_{(HH)}$</th>
<th>$\Delta R^2_{(D)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trust Scale</td>
<td>0.27 [0.17; 0.37]</td>
<td>0.11 [0.09; 0.28]</td>
<td>0.60 [0.53; 0.68]</td>
<td>0.03 [0.02; 0.05]</td>
<td>0.27 [0.22; 0.32]</td>
</tr>
<tr>
<td>Threatening world</td>
<td>0.33 [0.22; 0.46]</td>
<td>0.11 [0.09; 0.28]</td>
<td>0.60 [0.53; 0.68]</td>
<td>0.03 [0.02; 0.05]</td>
<td>0.27 [0.22; 0.32]</td>
</tr>
</tbody>
</table>

Note: N = 485. D: Dark Factor of Personality, HH: Honesty-Humility. $\chi^2_q$: variance explained in the full model. $\Delta R^2_{(HH)}$: increase in $R^2$ after adding low Honesty-Humility to the model. $\Delta R^2_{(D)}$: increase in $R^2$ after adding D to the model.

### General discussion

Recent research suggests that the shared variance of all aversive traits and thus their common core can be understood through the so-called Dark Factor of Personality, D, which is the fluid underlying disposition subsuming the aversive aspects of all aversive traits. Relatedly, the overlap of aversive traits, in particular the Dark Triad traits, has been described as the low pole of Honesty-Humility (Hodson et al., 2013; Lee et al., 2013; Muris et al., 2017), one of the basic dimensions of the HEXACO personality model (Lee & Ashton, 2008). Notwithstanding other findings casting doubt on this particular conclusion (Howard & Van Zandt, 2020; Mclarnon & Tarraf, 2021; Schreiber & Marcus, 2020), low Honesty-Humility may adequately represent the common core of all aversive traits, in turn implying that D may essentially be equivalent to Honesty-Humility.

### Table 6

Unconstrained latent correlations and model comparisons (Study 3).

<table>
<thead>
<tr>
<th>Variable</th>
<th>D [95% CI]</th>
<th>low HH [95% CI]</th>
<th>$\Delta \chi^2(1)$</th>
<th>$p$ (Holm-Bonferroni-corrected)</th>
<th>ER</th>
<th>q</th>
<th>p (Holm-Bonferroni-corrected)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trust Scale</td>
<td>0.27 [0.17; 0.37]</td>
<td>0.11 [0.09; 0.28]</td>
<td>0.60 [0.53; 0.68]</td>
<td>0.03 [0.02; 0.05]</td>
<td>0.27 [0.22; 0.32]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Threatening world</td>
<td>0.33 [0.22; 0.46]</td>
<td>0.11 [0.09; 0.28]</td>
<td>0.60 [0.53; 0.68]</td>
<td>0.03 [0.02; 0.05]</td>
<td>0.27 [0.22; 0.32]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: N = 508. D: Dark Factor of Personality, HH: Honesty-Humility. $\Delta \chi^2$: scaled log-likelihood ratio test (change in model fit after restricting correlations to be equal). $ER$: normalised evidence ratio comparing the less restricted to the restricted model. $q$: difference in the (absolute) zero-order correlations with an outcome between low Honesty-Humility and D as measured by Cohen’s $q$ with associated (one-sided and Holm-Bonferroni corrected) $p$-values.

### Table 7

Latent regression results predicting the outcomes by D and low Honesty-Humility (Study 3).
low Honesty-Humility. Thus, the present study sought to extend previous research by explicitly testing the overlap of Honesty-Humility with the common core of all aversive traits.

Meta-analytically aggregated across the present studies, D and low Honesty-Humility shared about 66% of their variance (see additional material on the OSF), which is considerably less than the overlap of 90% found between low Honesty-Humility and the common core of Dark Triad traits (Hodson et al., 2018). This difference is likely due to the broader range of aversive traits covered by D (as compared to the common core of the Dark Triad traits) which may also comprise aspects accounted for by other HEXACO dimensions. Yet, the overlap is sizable enough that D and low Honesty-Humility must be expected to show similar relations with many aversive outcomes. These similarities, however, are not sufficient to consider them functionally equivalent. Instead, a more critical test of the equivalence assumption complements the assessment of their correlation by testing their nomological consistency for a range of theoretically derived criteria. If D and low Honesty-Humility were essentially equivalent, both would have to be associated with theoretically derived criteria to approximately the same extent, and neither should account for incremental variance in said criteria over the other.

Conceptually, one crucial difference between Honesty-Humility and D is that the former was inductively derived from lexical studies and subsequently included as an orthogonal dimension in a model of basic personality structure (Ashton & Lee, 2007; Zettler et al., 2020), whereas D was derived deductively, theoretically comprises aspects related to several dimensions in the HEXACO model, and indeed empirically relates to these (Moshagen et al., 2018; Schreiber & Marcus, 2020). These theoretical differences, alone, render it unlikely that D and low Honesty-Humility represent the exact same construct. Moreover, although low Honesty-Humility and D share the aspect of utility maximization, they also differ in the extent to which utility is accompanied by or even achieved through inflicting disutility on others as well as beliefs and attitudes that are used to justify malevolent behaviors. Thus, in four studies, we investigated more closely whether D and Honesty-Humility are not only theoretically, but also empirically dissociable and how exactly they differ.

Specifically, in the first study we demonstrated that D and low Honesty-Humility do not equally determine aversive traits longitudinally. For the majority of the aversive traits considered, D outperformed Honesty-Humility—with the exception of Agentic Narcissism, Psychological Entitlement, and Self-Interest, all of which are conceptually and operationally more closely related to low Honesty-Humility than to D. Given that D and low Honesty-Humility accounted for unique variance in every aversive trait included in the study, the two cannot be equivalent representations of the common core of all aversive traits.

In Studies 2, 3, and 4 we investigated the specific differences between low Honesty-Humility and D in terms of content. To this end, we theoretically derived criteria which should be subsumed in Honesty-Humility and D to a different extent and should thus have unique variance accounted for by the two constructs. To summarize our findings, Table 10 provides an overview of the unique contributions of low Honesty-Humility and D, respectively, relative to the total explained variance in these criteria (i.e., as the relation of their respective $\Delta R^2$ to the total $R^2$).

As expected on theoretical grounds, Pretentiousness (i.e., the desire for social recognition) was more strongly linked to low Honesty-Humility than to D. This is consistent with theory, as the greed avoidance facet of Honesty-Humility includes the desire for wealth and social status at the lower pole, whereas D does not involve seeking kinship as a prominent aspect and may actually be incompatible, given that D comprises provoking disutility for others (i.e., sadistic or spiteful behavior). Note, however, that the aspect of utility maximization refers to an individual’s goals and also comprises non-materialistic utility. It is therefore not surprising that D and Pretentiousness are still related, as one can certainly derive utility from impressing others with one’s wealth and also use it to gather admirers that can then be exploited and manipulated more easily. Nevertheless, in line with our hypothesis, low Honesty-Humility was more strongly related to Pretentiousness than D.

Vice versa, we expected D to be more strongly related to spiteful behavior as operationalized by the Sadistic SVO (Moshagen et al., 2020). Unexpectedly, D and low Honesty-Humility were not dissociable via this measure and both accounted only for a comparatively (though not unusually; Thielmann, Spadaro, & Balliet, 2020) small proportion of variance. Considering that self-reported Sadism and Spitefulness were substantially predicted by both, and indeed better by D than Honesty-Humility in Study 1, this finding was unexpected. A possible explanation might be that participants had to weigh the utility of the immaterial gain of harming the other person against the utility of the possible monetary gain. Our implicit assumption was that the immaterial gain would often outweigh the material gain, which might not hold because participants could not actually experience the “suffering” of their counterpart. Thus, the subjective utility of harming the other person might have been small, at best. Future research may thus seek a more suitable behavioral measure of spiteful and/or sadistic behavior.

Then, we considered another aspect that actually defines D but is only loosely related to (some aspects of) Honesty-Humility, namely, “beliefs that serve as justifications” (Moshagen et al., 2018, p. 657) for malevolent behavior, especially distrust-related beliefs. Results consistently showed that D relates more strongly than low Honesty-Humility to those beliefs reflecting rather negative expectations of one’s surroundings and regarding other people as a potential threat. These types of beliefs serve as particularly strong justifications for malevolent behavior: If one believes others are a threat and will be exploitative, it is actually normatively necessary to behave uncooperatively to prevent being exploited (Gächter, 2004).

Finally, we demonstrated empirically the theoretically implied difference between D and low Honesty-Humility with respect to callousness or lack of empathy: as empathy is assigned to the Emotionality dimension (rather than Honesty-Humility) within the HEXACO model, it should not be comprised in Honesty-Humility. By contrast, (lack of) empathy is of central theoretical relevance to D (Moshagen et al., 2018, p. 656). Correspondingly, even though Honesty-Humility showed medium-sized bivariate correlations with callousness and two facets of empathy, it explained virtually no variance once D was accounted for. In turn, D was strongly related to all measures of (low) empathy and accounted for up to 35% of unique variance beyond Honesty-Humility.

Taken together, results from longitudinal data on a range of theoretically selected outcome criteria showed that low Honesty-Humility and D are best understood as operationally strongly related, but nonetheless functionally different and nomologically distinct constructs. Both low Honesty-Humility and D carry psychologically relevant meaning beyond each other (despite their substantial correlation). In particular, low Honesty-Humility and D longitudinally accounted for diverse aversive traits to different extents (with D predicting a larger

### Table 10

Unique contributions of D and low Honesty-Humility, respectively, relative to the total explained variances in each criterion (Studies 2–4).

<table>
<thead>
<tr>
<th>Study</th>
<th>Criterion</th>
<th>Honesty-Humility</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Sadistic SVO</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>Pretentiousness</td>
<td>45%</td>
<td>10%</td>
</tr>
<tr>
<td>3</td>
<td>Distrust</td>
<td>15%</td>
<td>69%</td>
</tr>
<tr>
<td>3</td>
<td>Trust Scale</td>
<td>9%</td>
<td>55%</td>
</tr>
<tr>
<td>3</td>
<td>Threatening World</td>
<td>25%</td>
<td>75%</td>
</tr>
<tr>
<td>4</td>
<td>Empathic Concern</td>
<td>3%</td>
<td>59%</td>
</tr>
<tr>
<td>4</td>
<td>Perspective Taking</td>
<td>3%</td>
<td>57%</td>
</tr>
<tr>
<td>4</td>
<td>Callousness</td>
<td>16%</td>
<td>43%</td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td>15%</td>
<td>57%</td>
</tr>
</tbody>
</table>

Note: The table displays the relation of $\Delta R^2$ (of each predictor) to total $R^2$ combined.
Conceptualization, Manuscript Writing.

Although low Honesty-Humility shares large parts of its respective nomological nets, they do not suggest that D and low Honesty-Humility are actually distinct with respect to most aversive traits, so that further exploration of their differences is warranted.

Notably, the findings reported herein are aligned with a recent meta-analysis showing that while low Honesty-Humility and the Dark Triad traits share large parts of their respective nomological nets, they do not overlap perfectly, indicating that they should rather be regarded as having a common theoretical basis than as being redundant constructs (Howard & Van Zandt, 2020). Correspondingly, the findings presented herein show that low Honesty-Humility is not only distinct from the common core of the Dark Triad, but more generally from the common core of all dark traits.

The findings of the present studies are further compatible with another recent meta-analysis revealing that low Honesty-Humility converged with the Dark Triad traits and other basic personality dimensions such as Agreeableness on a common higher-order factor representing D (Schreiber & Marcus, 2020). By this logic, too, D and low Honesty-Humility cannot be functionally equivalent. Importantly, this does not imply that D should either replace Honesty-Humility or be considered as a seventh basic trait in the HEXACO personality model, as D explicitly comprises aspects across basic traits and thus cannot represent an approximately orthogonal personality dimension (Moshagen et al., 2018). Rather, these meta-analytic findings are aligned with the present conclusion that neither is D perfectly indicated by low Honesty-Humility nor is low Honesty-Humility perfectly predicted by D, and thus that both carry unique variance components and theoretical meaning. Thus, although low Honesty-Humility appears to be the best proxy for the common core of aversive traits among all basic personality dimensions, and although low Honesty-Humility and D are closely related operationally and empirically, they differ conceptually—both in terms of their relations with other basic personality dimensions and in terms of defining aspects—and are functionally distinct in that both carry behaviorally relevant variance beyond each other.

Author note
This research was funded by the Deutsche Forschungsgemeinschaft (DFG) grant 2277, Research Training Group „Statistical Modeling in Psychology“ (SMIP). We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the studies. Additional materials (including data and analysis scripts) are provided online on the Open Science Framework (https://osf.io/35adh/).

CRedit authorship contribution statement
Luisa K. Horsten: Conceptualization, Data Collection, Data Analysis, Data curation, Manuscript Writing. Morten Moshagen: Conceptualization, Data Analysis, Manuscript Writing. Ingo Zettler: Conceptualization, Manuscript Writing. Benjamin E. Hilbig: Conceptualization, Manuscript Writing, Supervision.

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

Acknowledgement
None.

Appendix A. Supplementary material
Supplementary data to this article can be found online at https://doi.org/10.1016/j.jrp.2021.104154.

References
Corrigendum


Luisa K. Horsten \textsuperscript{a, *}, Morten Moshagen \textsuperscript{b}, Ingo Zettler \textsuperscript{c}, Benjamin E. Hilbig \textsuperscript{a}

\textsuperscript{a} University of Koblenz-Landau, Fortstr. 7, 76829 Landau, Germany
\textsuperscript{b} Ulm University, Albert-Einstein-Allee 47, 89081 Ulm, Germany
\textsuperscript{c} University of Copenhagen, Øster Farimagsgade 2A, 1353 Copenhagen, Denmark

The authors regret that, due to an error in the code (now corrected on the OSF), the $\Delta R^2$s reported in Tables 3, 5, 7, and 9, as well as the relative unique contributions reported in Table 10 of the original publication, are incorrect. Corrected tables are reported below. None of the conclusions drawn in the original publication are affected by these changes, thus the interpretation of the results can be maintained unaltered for each study. The authors would like to apologise for any inconvenience caused.
Table 3
Latent regression results for longitudinally predicting unresidualized dark traits by D and Honesty-Humility (Study 1).

<table>
<thead>
<tr>
<th>Dark trait 2018</th>
<th>β02014 (95% CI)</th>
<th>β02014 (95% CI)</th>
<th>R²(D, HH)</th>
<th>ΔR²(D, HH)</th>
<th>ΔR²(D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Egoism</td>
<td>0.71 (0.53; 0.93)</td>
<td>−0.17 (−0.38; 0.05)</td>
<td>0.35 (0.04; 0.20)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Machiavellianism</td>
<td>0.55 (0.39; 0.71)</td>
<td>0.15 (−0.03; 0.33)</td>
<td>0.14 (≤0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moral Disengagement</td>
<td>0.67 (0.49; 0.85)</td>
<td>−0.02 (−0.22; 0.19)</td>
<td>0.14 (≤0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agentic Narcissism</td>
<td>−0.19 (−0.41; 0.03)</td>
<td>0.71 (0.33; 0.73)</td>
<td>0.08 (≤0.00)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychological Entitlement</td>
<td>0.10 (0.09; 0.29)</td>
<td>0.53 (0.33; 0.96)</td>
<td>0.10 (≤0.00)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychopathy</td>
<td>0.70 (0.53; 0.87)</td>
<td>0.06 (−0.15; 0.27)</td>
<td>0.15 (≤0.00)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sadism</td>
<td>0.52 (0.31; 0.72)</td>
<td>0.01 (−0.21; 0.23)</td>
<td>0.08 (≤0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-Interest</td>
<td>−0.32 (−0.56; 0.09)</td>
<td>0.81 (0.56; 1.06)</td>
<td>0.17 (≤0.00)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spitefulness</td>
<td>0.80 (0.64; 0.97)</td>
<td>−0.19 (−0.38; 0.00)</td>
<td>0.20 (≤0.01)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: N = 1,261. Standardized latent linear regression coefficients. D = Dark Factor of Personality, HH = low Honesty-Humility. \( R^2(D, HH) \): variance explained by the full model. \( \Delta R^2(D, HH) \): increase in \( R^2 \) after adding Honesty-Humility to the model. \( \Delta R^2(D) \): increase in \( R^2 \) after adding D to the model.

Table 5
Latent regression results predicting the outcomes by low Honesty-Humility and D (Study 2).

<table>
<thead>
<tr>
<th>Outcome variable</th>
<th>( \beta_0 ) (95% CI)</th>
<th>( \beta_{HH} ) (95% CI)</th>
<th>( R^2_{(D, HH)} )</th>
<th>( \Delta R^2_{(D, HH)} )</th>
<th>( \Delta R^2_D )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sadistic SVO</td>
<td>0.27 (−0.14; 0.67)</td>
<td>−0.08 (−0.47; 0.32)</td>
<td>0.04 (0.02; 0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretentiousness</td>
<td>−0.47 (−0.79; −0.15)</td>
<td>1.20 (0.90; 1.49)</td>
<td>0.57 (0.33; 0.12)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: N = 313. Standardized latent regression coefficients. D = Dark Factor of Personality, HH = low Honesty-Humility. \( R^2(D, HH) \): variance explained in the full model. \( \Delta R^2(D, HH) \): increase in \( R^2 \) after adding low Honesty-Humility to the model. \( \Delta R^2(D) \): increase in \( R^2 \) after adding D to the model.

Table 7
Latent regression results predicting the outcomes by D and low Honesty-Humility (Study 3).

<table>
<thead>
<tr>
<th>Outcome variable</th>
<th>( \beta_D ) (95% CI)</th>
<th>( \beta_{HH} ) (95% CI)</th>
<th>( R^2_{(D, HH)} )</th>
<th>( \Delta R^2_{(D, HH)} )</th>
<th>( \Delta R^2_{D} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distrust</td>
<td>0.75 (0.59; 0.92)</td>
<td>−0.35 (−0.53; −0.17)</td>
<td>0.26 (0.04; 0.19)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trust Scale</td>
<td>−0.45 (−0.69; −0.24)</td>
<td>0.15 (−0.10; 0.40)</td>
<td>0.11 (≤0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Threatening world</td>
<td>0.29 (0.09; 0.49)</td>
<td>−0.13 (−0.33; 0.06)</td>
<td>0.04 (≤0.01)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: N = 508. Standardized latent linear regression coefficients. D = Dark Factor of Personality, HH = low Honesty-Humility. \( R^2(D, HH) \): variance explained in the full model. \( \Delta R^2(D, HH) \): increase in \( R^2 \) after adding low Honesty-Humility to the model. \( \Delta R^2_D \): increase in \( R^2 \) after adding D to the model.

Table 9
Latent regression results predicting the outcomes by D and low Honesty-Humility (Study 4).

<table>
<thead>
<tr>
<th>Outcome variable</th>
<th>( \beta_D ) (95% CI)</th>
<th>( \beta_{HH} ) (95% CI)</th>
<th>( R^2_{(D, HH)} )</th>
<th>( \Delta R^2_{(D, HH)} )</th>
<th>( \Delta R^2_{D} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empathic concern</td>
<td>−0.84 (−0.95; 0.25)</td>
<td>0.12 (−0.01; 0.25)</td>
<td>0.59 (0.01; 0.42)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perspective taking</td>
<td>−0.06 (−0.77; −0.54)</td>
<td>0.08 (−0.05; 0.22)</td>
<td>0.37 (≤0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Callousness</td>
<td>0.85 (−0.76; 0.94)</td>
<td>−0.03 (−0.14; 0.08)</td>
<td>0.69 (≤0.01)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: N = 485. Standardized latent linear regression coefficients. \( R^2(D, HH) \): variance explained in the full model. \( \Delta R^2(D, HH) \): increase in \( R^2 \) after adding low Honesty-Humility to the model. \( \Delta R^2_D \): increase in \( R^2 \) after adding D to the model.

Table 10
Unique contributions of D and low Honesty-Humility, respectively, to the total explained variances in each criterion (Studies 2–4).

<table>
<thead>
<tr>
<th>Study</th>
<th>Criterion</th>
<th>Honesty-Humility</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Sadistic SVO</td>
<td>58%</td>
<td>10%</td>
</tr>
<tr>
<td>2</td>
<td>Pretentiousness</td>
<td>50%</td>
<td>18%</td>
</tr>
<tr>
<td>3</td>
<td>Distrust</td>
<td>15%</td>
<td>72%</td>
</tr>
<tr>
<td>3</td>
<td>Trust Scale</td>
<td>1%</td>
<td>56%</td>
</tr>
<tr>
<td>3</td>
<td>Threatening World</td>
<td>19%</td>
<td>70%</td>
</tr>
<tr>
<td>4</td>
<td>Empathic Concern</td>
<td>2%</td>
<td>72%</td>
</tr>
<tr>
<td>4</td>
<td>Perspective Taking</td>
<td>3%</td>
<td>71%</td>
</tr>
<tr>
<td>4</td>
<td>Callousness</td>
<td>-</td>
<td>58%</td>
</tr>
</tbody>
</table>

Note: The table displays the relation of \( \Delta R^2 \) (of each predictor) to total \( R^2 \) combined.
Agreeableness is dead. Long live Agreeableness? Reply to Vize and Lynam

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

a University of Koblenz-Landau
b Ulm University
c University of Copenhagen

In our original paper (Moshagen, Zettler, Horsten, & Hilbig, 2020), we demonstrated that the Dark Factor of Personality (‘D’), i.e. the common core underlying all aversive traits, is not equivalent to Agreeableness as per the Five-Factor Model of Personality, operationally defined by the combination of Agreeableness scales from the BFAS (DeYoung, Quilty, & Peterson, 2007), the BFI-2 (Soto & John, 2017), the IPig-50 (Goldberg, 1992), and the NEO-FFI (McCrae & Costa, 2004). Specifically, we showed that D and Agreeableness have less than 50% common variance and account for substantial unique variance components in several theoretically derived criteria.

In their commentary, Vize and Lynam (2021, V&L henceforth) question our conclusions for two primary reasons: First, that the four scales we used insufficiently capture Agreeableness. Instead, they favor a broader measurement of Agreeableness (denoted AG+ in what follows) which also includes items from the Honesty-Humility and interstitial Altruism scales from the HEXACO-PI-R (Lee & Ashton, 2006). Indeed, AG+ closely aligns with D (Vize, Miller, & Lynam, in press). Secondly, V&L argue that equivalence ought to be assessed statistically by means of the ICC, and that the ICC we reported, .75, ought to be interpreted as evidence for equivalence. While we welcome V&L’s feedback and maintain that our disagreement is largely down to semantics, these two main arguments do deserve some second thought, given their implications for the Agreeableness construct within the Five-Factor Model.

The first argument, in essence, maintains that AG+ adequately represents the theoretical construct Agreeableness, whereas the Agreeableness scales from the BFAS, the BFI-2, the IPig-50, and the NEO-FFI do not. Consequently, if indeed neither of these scales – nor their combination – adequately represent the theoretical construct Agreeableness, then the vast majority of studies investigating Agreeableness as per the Five-Factor Model within the past decades must have poorly assessed Agreeableness. Although we are not inclined to take sides in a quarrel that is not ours, we do not expect that scholars who work within the Five-Factor framework agree with this conclusion.

Nonetheless, one need not dismiss V&L’s arguments so as to salvage prior work conducted within the Five-Factor Model. Instead, one may entertain the hypothesis that AG+ – as used by Vize et al. (in press) – is indeed more adequate. However, the data of Vize et al. (in press) do give some reason for pause. For one, model fit of the AG+ factor must be considered poor by all established standards (RMSEA = .16, SRMR = .10, CFI = .68). More importantly, the AG+ factor correlates approximately twice as high to three of the four remaining Five-Factor Model dimensions than Agreeableness scales typically do (for a recent meta-analysis, see Park et al., 2020) with a particularly striking correlation of .57 between AG+ and Conscientiousness. Thus, if AG+ is supposed represent Agreeableness within the Five-Factor Model, this model cannot involve approximately orthogonal dimensions. Although Vize and Lynam (2021) explicitly consider this unproblematic, others invested in the Five-Factor Model clearly do not (Goldberg, 1992; McCrae & Costa, 2003; Saucier, 2002).

With respect to the second argument – that ICC is the most adequate criterion for equivalence and ICC = .75 ought to be interpreted as corresponding positive evidence – two observations deserve attention. For one, the ICC is a problematic criterion to determine equivalence (Table 1): The ICC between two variables X1 and X2 can take a value of -1.00, even though one is uncorrelated to all outcomes considered (O1-O3) whereas the other exhibits strong correlations to these outcomes. Vice versa, the ICC can be practically zero, even though two variables (Y1, Y2) show virtually identical correlations to the outcomes in question (i.e. perfect nomological consistency, Thielmann & Hilbig, 2019). Worse yet, merely considering the opposing pole of Y2 would change the ICC from -.06 to -1, thus yielding an entirely different estimate although the to-be-compared constructs are exactly the same. Given these properties, the ICC is not a particularly useful – let alone the only or most adequate - criterion to assess equivalence.

Nonetheless, entertaining V&L’s argument that ICC=.75 signals evidence for equivalence, the data of Vize et al. (in press) show that AG+ and Conscientiousness must be considered equivalent, as they...
yield an ICC of .79. In other words, by the very logic of V&L, our data yield about as much evidence for the equivalence of D and Agreeableness as their data yield for the equivalence of AG+ and Conscientiousness. Whereas the co-occurrence of Agreeableness and Conscientiousness has previously been noted and attributed to a higher-order factor termed ‘stability’ (e.g., DeYoung, 2006), this actually rules out that they are equivalent and that this higher-order factor can be (labeled) Agreeableness.

In summary, the arguments by V&L strike us as incompatible with the traditional notion (and operationalization) of Agreeableness within the Five-Factor Model: Not only do they assert that in virtually all the published literature, Agreeableness was measured inappropriately. But their preferred operationalization (AG+) strongly reduces the independence of dimensions within the Five-Factor Model and by their preferred metric (ICC) and interpretation thereof, the constructs Agreeableness and Conscientiousness are equivalent. None of these arguments are necessarily wrong, but if true, they do strike us as the final straw for (Agreeableness as defined within) the Five-Factor Model.

Setting aside these issues – most of which arguably need to be resolved and reconciled by those devoted to the Five-Factor Model – we emphasize that there is large consensus between V&L’s and our own position. For one, we remain agreed on the notion that there is a common core to aversive traits. More importantly, Vize et al.’s (in press) finding that AG+ (which subsumes aspects of HEXACO Honesty-Humility and strongly relates to Conscientiousness) approximates D is entirely compatible with our view that “D represents a blend of basic traits”, especially Honesty-Humility, Agreeableness, and Conscientiousness (Moshagen, Hilbig, & Zettler, 2018, p. 682). We may disagree on semantics, specifically whether a super-dimension spanning Agreeableness, Honesty-Humility, Altruism, and Conscientiousness can be (or ought to be called) Agreeableness. But we are actually perfectly agreed that what is traditionally understood to be (and commonly measured under the label of) Five-Factor Model Agreeableness alone is not equivalent to D.

References
Vize, C. E., Miller, J. D., & Lynam, D. R. (in press). Examining the conceptual and empirical distinctiveness of agreeableness and “dark” personality items. Journal of Personality, 0, 0.
1 INTRODUCTION

People differ in their tendencies to behave in unethical and socially harmful ways. These differences are commonly ascribed to ethically and socially aversive (often called “dark”) personality traits. Prominent examples of such aversive traits are the Dark Triad components—Narcissism, Machiavellianism, and Psychopathy (Paulhus & Williams, 2002)—but many more aversive traits have been proposed (e.g., Sadism, O’Meara et al., 2011; Spitefulness, Marcus et al., 2014). Importantly, aversive traits are conceptually similar and highly inter-correlated (e.g., Egan et al., 2015; Jonason et al., 2017; Muris et al., 2017), and there is now strong consensus that they share a common core (Jonason...
et al., 2017; Muris et al., 2017; Schreiber & Marcus, 2020; Vize & Lynam, 2020). This common core has been defined 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” (Moshagen et al., 2018, p. 657) and termed the Dark Factor of Personality (D).

Recent research has attempted to approximate the common core of aversive traits via established personality constructs, one of which is (low) Agreeableness as traditionally conceptualized within the Big Five framework (Paulhus & Williams, 2002; Stead & Fekken, 2014; Vize, Collison, et al., 2020). By this logic, features of low Agreeableness—such as a lack of empathy and compassion, as well as a tendency for manipulation and antagonism (Graziano & Tobin, 2017)—also represent the shared features of aversive traits. Indeed, Vize, Miller, et al. (2020) explicitly suggested that, on the construct level, low Big Five Agreeableness is equivalent to the common core of aversive traits, D.

Despite the obvious conceptual overlap between D and low Big Five Agreeableness, the two have been dissociated both theoretically and empirically (Moshagen, Zettler, Horsten, et al., 2020). In particular, a broad measure of Agreeableness—modeled as the commonalities between five well-established Agreeableness scales (i.e., Big Five Aspects Scale [BFAS], Big Five Inventory 2 [BFI2], International Personality Item Pool Big-Five Factor Markers [IPIP-50], NEO Five-Factor Inventory [NEO-FFI], and HEXACO Agreeableness vs. Anger)—shared only about 40% variance with a broad measure of D (Moshagen, Zettler, Horsten, et al., 2020), thus suggesting that the two constructs overlap substantially, but are not identical. Moreover, the two constructs explained unique variance components (above each other) in a host of deductively derived criteria. Specifically, D predicted incremental variance over Agreeableness (mean $\Delta R^2 = 0.13$—a medium-sized effect according to Cohen, 1988) in behavioral dishonesty, competitive and dangerous worldviews, guilt proneness, internet trolling, and stereotypical sexualized behaviors, whereas Agreeableness predicted incremental variance beyond D in empathy ($\Delta R^2 = 0.39$ and thus above a large effect size; Moshagen, Zettler, Horsten, et al., 2020). These results imply that D and Agreeableness are nomologically and functionally distinct in the sense that they comprise unique, behaviorally relevant trait variance (Moshagen, Zettler, Horsten, et al., 2020).

Questioning the validity of this conclusion, Vize, Miller, et al. (2020) argued that Moshagen, Zettler, Horsten et al.’s (2020) “coverage of AG [Agreeableness] may have missed important aspects, particularly related to modesty (Hex-HH) [HEXACO Honesty-Humility] and straightforwardness (NEO-PI-R)” (p. 597) and may thus have underestimated the similarity of D and Agreeableness. In line with this argument, Vize, Miller, et al. (2020) noted that the overlap between Agreeableness and D is considerably higher when a particularly broad operationalization of Agreeableness, which we denote AG+ in what follows, is used. Moreover, they found that D did not predict incremental variance beyond AG+ in (self-reported) reactive/proactive aggression and crime and analogous behavior. Even though these are only two criteria, the findings appear to support the notion by Vize, Miller, et al. (2020) that D and Agreeableness are equivalent constructs.

However, to avoid jingle-fallacies, it is necessary to clarify that AG+ is not Agreeableness in the sense of a largely orthogonal dimension within the Big Five framework. Specifically, AG+ is represented by an item set compiled of 104 items from diverse scales (Crowe et al., 2018), including Big Five Agreeableness measures (BFAS, BFI, Faceted Inventory of the Five-Factor Model [FI-FFM], IPIP NEO Personality Inventory Revised [IPIP-NEO-PI-R]), items measuring HEXACO Agreeableness vs. Anger (which is not equivalent to Big Five Agreeableness; Ashton et al., 2014; Thielmann et al., 2021; Endnote 1), items measuring HEXACO Honesty-Humility, and items measuring the interstitial Altruism facet of the HEXACO P-IR (a blend of HEXACO Honesty-Humility, Emotionality, and Agreeableness vs. Anger, representing a “tendency to be sympathetic and soft-hearted toward others”; Ashton et al., 2014, p. 142). In turn, the inclusion of content clearly pertaining to basic traits other than Big Five Agreeableness is arguably the reason why AG+ and the remaining Big Five dimensions are associated considerably more strongly than what is normally found (median $r = .35$ as compared to a median of $r = .20$ in the meta-analysis by Park et al., 2020), with a particularly striking correlation of $r = .57$ with Conscientiousness. As such, AG+ cannot be included in a model of basic personality structure involving approximately orthogonal dimensions (Goldberg, 1992; Saucier, 2002) and thus does not represent Big Five Agreeableness. It may well represent some variant of a higher-order “stability” factor comprising aspects of Big Five Agreeableness and Conscientiousness (e.g., DeYoung, 2006), but this, too, logically implies that AG+ cannot be Big Five Agreeableness (Hilbig et al., 2021).

Aside from the fact that AG+ does not represent Agreeableness as conceptualized in models of basic personality structure (Big Five, FFM, HEXACO), the findings of Vize, Miller, et al. (2020) do indicate that a construct subsuming content from several basic personality dimensions, such as AG+, may closely approximate D. In fact, this is fully compatible with the previous conclusion that D, at least to a notable extent, can be understood as “a blend of basic traits” (Moshagen et al., 2018, p. 682), especially HEXACO Honesty-Humility, Agreeableness,
and Conscientiousness. The remaining question, then, is whether the particular blend of Agreeableness-related traits put forward by Crowe et al. (2018) and used in Vize, Miller, et al. (2020), AG+, is indeed equivalent to D both theoretically and empirically.

Ultimately, it is impossible to prove that two constructs are equivalent and relate similarly to all conceivable criteria. Nonetheless, it is possible to take a falsificationist approach and theoretically derive criteria that are a priori particularly likely to yield dissimilarity. Then, if the empirical evidence fails to support the hypothesis of dissimilarity, one may retain the view that the two constructs are functionally equivalent (Gonzalez et al., 2020). With regard to the case at hand, this means that if D and AG+ are non-equivalent, it must be possible to theoretically derive aspects and, ultimately, criteria on which the two can be shown to differ to a non-trivial extent.

In selecting such criteria that may differentiate between D and AG+, we relied on evidence and notions on the conceptual differences between D and Big Five Agreeableness (which constitutes the largest part of AG+). Moshagen, Zettler, Horsten, et al. (2020) pointed out that, unlike D, Agreeableness—as represented by common verbal definitions (e.g., Graziano & Tobin, 2009) and adjective lists (Goldberg, 1992; John et al., 2008)—does not capture the willingness to impose disutility on others even at some cost, nor individual differences in social cognition, as broadly as D. In line with this reasoning, D has been shown to explain notable variance beyond Agreeableness—operationally defined as the general factor across four common Big Five Agreeableness measures (BFAS, BFI2, IP1P-50, NEOFFI)—in 11 different justifying beliefs (mean $\Delta R^2 = .10$; Hilbig et al., 2022) and six antagonistic traits (including Grandiosity and Suspiciousness, as well as Deceitfulness and Manipulativeness) from the DSM-5 model (mean $\Delta R^2 = .07$; Scholz et al., 2022). In conclusion, Agreeableness does not comparably subsume the defining aspects of D that aversive tendencies are “accompanied by beliefs that serve as justifications” (Moshagen et al., 2018, p. 657) and involve inflicting disutility on others.

Moreover, D and basic personality dimensions (which are blended in AG+) conceptually differ in that D was deductively derived to represent the “aversive essence” of aversive traits. D therefore represents (only) aversive content whereas non-aversive aspects of these traits (e.g., vulnerability in Narcissism Bader et al., 2022) are, by definition, beyond the scope of D. By contrast, the basic personality dimensions blended into AG+ are inductively derived from lexical studies and therefore necessarily summarize whatever characteristics tend to co-occur—independent of whether these characteristics are aversive or not. As a consequence, Agreeableness typically also subsumes affable behaviors towards others, overlapping with Extraversion in sharing affiliation (i.e., “enjoying and valuing close interpersonal bonds and being warm and affectionate”); Depue & Morrone-Strupinsky, 2005, p. 314; see also DeYoung et al., 2013). In leaning heavily on Big Five Agreeableness and in borrowing items from the Altruism and Sympathy facets of the IP1P-NEO (e.g., “Make people feel welcome”, “Am not interested in other people’s problems”), AG+ also subsumes these affiliative tendencies. By contrast, (lack of) affiliation is not ethically or socially aversive per se and thus theoretically beyond the scope of D.

Given the above, a strict test of the (non-)equivalence of D and AG+ ought to involve those aspects subsumed more strongly in D than in AG+—that is, criteria representing the tendency to inflict disutility on others as well as criteria representing justifying beliefs—and, vice versa, aspects subsumed more strongly in AG+ than in D—that is, criteria representing affiliative tendencies. To this end, the following pre-registered study involved nine corresponding criteria, listed in Table 1: Competitive Jungle Social World View, Normlessness, and Social Dominance Orientation, representing justifying beliefs, and Pathological Selfishness and Exploitativeness, representing inflicting disutility on others, were expected to have incremental variance explained by D above AG+. Extraversion, Withdrawal, (fast) Life History Strategy, and Horizontal Collectivism, representing affiliative tendencies, were, in turn, expected to have incremental variance explained by AG+ above D.

2 | METHOD

The study and analysis plan were preregistered (https://aspredicted.org/uy9ms.pdf) before the start of data collection. Raw data, and analysis scripts are available on the Open Science Framework (OSF;https://osf.io/83sae/). The study was run based on approval by the university’s local ethics committee.

2.1 | Measures

The study was conducted in Germany, hence German translations of all measures were used (measures without an existing German translation prior to the study, i.e., FI-FFM and IP1P-NEO-PI-R, were translated under our coordination via the translation-retranslation method; Brislin, 1970; see OSF). D was assessed using the D70, a 70-item scale specifically designed to measure the latent factor underlying all aversive traits (Moshagen, Zettler, &
### TABLE 1 Overview of criterion measures and corresponding inventories.

<table>
<thead>
<tr>
<th>Category</th>
<th>Criterion</th>
<th>Pre-registered hypothesis</th>
<th>Scale (original source)</th>
<th>Number of items</th>
<th>Definition</th>
<th>Sample item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Justifying beliefs</td>
<td>Competitive Jungle Social World View</td>
<td>$\Delta R^2_{D_t} &gt; 0$</td>
<td>Two World View Scale (Sibley &amp; Duckitt, 2009)</td>
<td>6</td>
<td>“belief that the social world is a dangerous and threatening place in which good, decent people's values and way of life are threatened by bad people” (Duckitt et al., 2002, p. 78)</td>
<td>“It’s a dog-eat-dog world where you have to be ruthless at times.”</td>
</tr>
<tr>
<td>Normlessness</td>
<td></td>
<td>$\Delta R^2_{D_t} &gt; 0$</td>
<td>Normlessness Scale (Kohn &amp; Schooler, 1983)</td>
<td>4</td>
<td>“the belief that socially unapproved behaviors are required to achieve certain goals” (Ulleberg &amp; Rundmo, 2003, p. 430)</td>
<td>“If something works, it does not matter if it is right or wrong.”</td>
</tr>
<tr>
<td>Social Dominance Orientation</td>
<td></td>
<td>$\Delta R^2_{D_t} &gt; 0$</td>
<td>SDO$^7$ (Ho et al., 2015)</td>
<td>8</td>
<td>“Individual differences in the preference for group based hierarchy and inequality” (Ho et al., 2015, p. 1003)</td>
<td>“An ideal society requires some groups to be on top and others to be on the bottom.”</td>
</tr>
<tr>
<td>Disutility on others</td>
<td>Pathological Selfishness</td>
<td>$\Delta R^2_{D_t} &gt; 0$</td>
<td>Selfishness Questionnaire (Raine &amp; Uh, 2018)</td>
<td>8</td>
<td>“An inordinate focus on one's own welfare, regardless of the well-being of others.” (Raine &amp; Uh, 2018, p. 503)</td>
<td>“I know I love rewards in life, even if there is a cost to others.”</td>
</tr>
<tr>
<td>Exploitativeness</td>
<td></td>
<td>$\Delta R^2_{D_t} &gt; 0$</td>
<td>Interpersonal Exploitativeness Scale (Brunell et al., 2013)</td>
<td>6</td>
<td>“the state, condition, quality, or degree of unfairly or cynically using another person or group for profit or advantage” (Brunell et al., 2013, p. 2)</td>
<td>“I'm perfectly willing to profit at the expense of others.”</td>
</tr>
<tr>
<td>Affiliative tendencies</td>
<td>Extraversion</td>
<td>$\Delta R^2_{AG_t} &gt; 0$</td>
<td>HEXACO-60 (Ashton &amp; Lee, 2009)</td>
<td>10</td>
<td>“[Individual differences in] engagement in social endeavors (such as socializing, leading, or entertaining)” (Ashton &amp; Lee, 2007, p. 156)</td>
<td>“In social situations, I'm usually the one who makes the first move.”</td>
</tr>
<tr>
<td></td>
<td>Withdrawal</td>
<td>$\Delta R^2_{AG_t} &gt; 0$</td>
<td>PID-5 (Krueger et al., 2012)</td>
<td>10</td>
<td>“Reticence in social situations; avoidance of social contacts and activity; lack of initiation of social contact” (APA, 2013, p. 766)</td>
<td>“I keep to myself.”</td>
</tr>
<tr>
<td></td>
<td>Life History Strategy</td>
<td>$\Delta R^2_{AG_t} &gt; 0$</td>
<td>Mini-K (Figueroedo et al., 2006)</td>
<td>20</td>
<td>“A K-selected life-history strategy allocates an individual's bioenergetic and material resources preferentially to Somatic Effort over Reproductive Effort, and to Parental and Nepotistic Effort over Mating Effort, emphasizing the survival of individual organisms (whether self, offspring, or kin) over the production of new ones” (Figueroedo et al., 2006, p. 245)</td>
<td>“I often give emotional support and practical help to my friends.”</td>
</tr>
<tr>
<td></td>
<td>Horizontal Collectivism</td>
<td>$\Delta R^2_{AG_t} &gt; 0$</td>
<td>Individualism and Collectivism Scale-Revised (Singelis et al., 1995)</td>
<td>8</td>
<td>“the individual sees the self as an aspect of an in-group [...] all of whom are extremely similar to each other.” (Singelis et al., 1995, p. 244)</td>
<td>“My happiness depends very much on the happiness of those around me.”</td>
</tr>
</tbody>
</table>

*An ad-hoc translation was used, see OSF.*
Hilbig, 2020; Bader, Horsten, et al., 2021). AG+ was assessed via the 104 items identified by Crowe et al. (2018) and used by Vize, Miller, et al. (2020). An overview of the nine criterion measures (along with the corresponding hypotheses) is provided in Table 1. The Withdrawal items were rated on a 4-point Likert scale (1 = “very false” to 4 = “very true”). All other items were rated on a 5-point Likert scale (1 = “strongly disagree” to 5 = “strongly agree”).

2.2 Procedure & participants

Participants were recruited and compensated by a professional panel provider. The study comprised two measurement occasions. At T1, we assessed demographics (including sex and age), the D70, and AG+. At T2 (on average 14 days after T1), we again assessed sex and age, along with the nine criterion measures (see Table 1): Competitive Jungle Social World View, Normlessness, Social Dominance Orientation, Pathological Selfishness, Exploitativeness, HEXACO Extraversion, PID-5 Withdrawal, Mini-K, and Horizontal Collectivism. Within each measurement occasion, the order of scales was randomized and two attention checks were embedded within the scales (e.g., “Please select “Agree”. This serves to check your attention.”). Participants gave informed consent prior to each measurement occasion and they were fully debriefed about the study background at the end of T2.

In order to be able to model D and AG+ in a bifactor structure, we aimed at a final sample of approximately N = 1,000. Anticipating a certain dropout between T1 and T2, a total of 1,331 participants completed T1. In line with our preregistered exclusion criteria, 105 participants were excluded for failing at least one attention check and an additional 70 participants were excluded for suspected inattentive response behavior (based on response times of less than 2 seconds per item on average in more than 50% of the questionnaires and/or no variation in responses, i.e., SD = 0, on those scales having at least 25% reverse-keyed items). Thus, 1,156 participants were invited to T2, of whom 940 completed the survey. Of these, we had to exclude one participant for failing at least one attention check, and 12 participants for suspected inattentive response behavior. Additionally, we conservatively excluded 16 participants whose demographics did not match across the two measurement occasions (i.e., they reported to be younger or more than one year older at T2 than at T1 or indicated a different sex at T2 than at T1). Thus, a total of N = 911 participants (49% female), aged between 18 and 65 (M = 40.6, SD = 13.0) years, successfully completed both measurement occasions.

2.3 Analysis

We tested all hypotheses by estimating confirmatory factor models with the lavaan package (Rosseel et al., 2019) in R (R Core Team, 2020). Non-normality in the data was accounted for by maximum likelihood estimation with robust standard errors and scaled test statistics (as provided by the lavaan package when specifying “MLR” estimation; Yuan & Bentler, 2000; see also Savalei & Rosseel, 2022). Assuming data are missing at random, we addressed missing cases at the second measurement occasion by employing full information maximum likelihood estimation. Due to the high power of the chi-square model test (Moshagen & Erdfelder, 2016), we primarily considered the robust root mean square error of approximation (RMSEA) and the standardized root mean residual (SRMR) to assess model fit. RMSEA around .05 and SRMR around .06 are commonly considered to be indicative of adequate model fit (Browne & Cudeck, 1992). For the sake of transparency, we also report the robust comparative fit index (CFI; Brosseau-Liard & Savalei, 2014). However, given its stronger dependency on loading magnitude than on model misfit (Moshagen & Auerswald, 2018) and its limited value in evaluating absolute model fit (van Laar & Braeken, 2022), we did not rely on this index to assess model fit.

The D70 was modeled specifying a bifactor model, such that all items loaded both on a general factor and on one of five themes (Bader, Hartung, et al., 2021). The general factor in a bifactor model captures the shared variance among all items. In this case, it captures the aversive content of the indicators and thus represents the underlying aversive disposition, D. The themes capture the remaining shared variance between subsets of items, i.e., those aspects that are beyond the general disposition but characterize an individual’s specific pattern of aversive attributes in more detail (Bader, Hartung, et al., 2021). Given that specific factors capture what is beyond the common core (and thus D), they are not relevant when investigating the equivalence of D and AG+. The general factor and specific factors were constrained to mutual orthogonality and identified by setting one unstandardized loading each to 1. According to conventional guidelines (Browne & Cudeck, 1992), this model structure fit the data well, \( \chi^2(2,275) = 6,590, p < .001, \) RMSEA = .045, 90%CI [.044, .047], SRMR = .051, CFI = .799. Importantly, there was a very strong general factor (representing D) that accounted for the majority of reliable variance in the themes (.60 ≤ ECV ≤ .82, median .74). Consequently, the specific factors exhibited very low reliabilities (.07 ≤ \( \omega_{HS} \) ≤ .21, median .18), further supporting that they should not be interpreted substantively (Sellbom & Tellegen, 2019).
Analogously, AG+ was modeled specifying a bifactor structure with five themes representing the five factors extracted by Crowe et al. (2018). This model also fit the data well, $\chi^2(5,148) = 15,020$, $p < .001$, RMSEA = .044, 90%CI [.043, .045], SRMR = .061, CFI = .741. In contrast to D, the general factor representing AG+ was a less strong (.31 ≤ ECV ≤ .83, median .45). In other words, AG+ captured less variance shared by all items and, consequently, the specific factors exhibited higher reliabilities than the D specific factors (.04 ≤ $\omega_H$ ≤ .63, median .46). To ensure that results are not dependent on the choice of bifactor models for D and AG+, we replicated all analyses using single-factor specifications for D and AG+, respectively, which yielded essentially equivalent conclusions.

To test our hypotheses, we then ran separate sequential regression analyses for each criterion. That is, we first regressed each criterion on D and AG+, respectively, and then on both simultaneously to calculate the incremental portion of variance ($\Delta R^2$) explained by either AG+ or D, respectively. To gauge the strength of evidence (beyond the effect size $\Delta R^2$), we considered normalized evidence ratios (ERs) computed from weighted BICs (Wagenmakers & Farrell, 2004; Wu et al., 2020). ERs quantify the support in favor of the less parsimonious model (in this case the model in which the criterion is regressed on both predictors) and range from 0 (no evidence for the less parsimonious model, i.e., the second predictor’s contribution to the explained variance is negligible) to 1 (perfect evidence for the less parsimonious model, i.e., the second predictor contributes meaningfully to the explained incremental variance).

## RESULTS

Descriptive statistics (of observed scales), reliabilities of the factors, and zero-order inter-correlations of latent factors are summarized in Table 2. The latent correlation between D and AG+ ($r = -.59$) implied 35% of shared variance and was significantly different from unity, $\Delta \chi^2(1) = 5.56$, $p = .02$, ER > .999.

As can be seen in the upper part of Table 3, and supporting the hypotheses that justifying beliefs and inflicting disutility on others are more strongly related to D than to AG+, D explained incremental variance over AG+ for all hypothesized outcomes (i.e., for Competitive Jungle World View, Normlessness, Social Dominance Orientation, Pathological Selfishness, and Exploitativeness). Overall, the proportion of incremental variance explained in these criteria by D was in the range of medium-sized to large effects (Cohen, 1988) with a median $\Delta R^2 = .30$.

Moreover, as presented in the lower part of Table 3 and supporting the hypothesis that affiliative tendencies are
more strongly related to AG+ than to D, AG+ predicted incremental variance over D for all hypothesized outcomes in this domain (i.e., Extraversion, Withdrawal, Life History Strategy, and Horizontal Collectivism). However, the proportion of incremental variance explained in these criteria by AG+ was only in the range of small to medium-sized effects for all criteria, with a median of $\Delta R^2 = .03$. Despite varying proportions of (absolute and) incremental variances explained, evidence ratios consistently (and typically very strongly) supported all hypotheses.

In sum, the results show that D and AG+ account for different variance components in all outcome criteria and, more specifically, each construct accounts for incremental variance in exactly those criteria that it ought to based on a priori theoretical considerations.10

As an additional exploratory analysis, we followed the recommendation of an anonymous reviewer and applied Multidimensional Scaling (MDS) to the D70 and AG+ items. MDS positions the items in an n-dimensional space based on their distances, which allows evaluating which items are close to and further away from each other, respectively. The analysis and results are described in more detail in the additional materials on the OSF. In brief, we first recoded the items so that all pointed in the same direction (to be interpreted as high D). Levels of stress (a badness-of-fit measure which indicates how many axes spanned by the three axes, the centroid of the D items

was in the exact opposite octant (+, −, +). Notably, these octants were exclusively occupied by AG+ or D70 items, respectively. The average Euclidean distance was 0.26 among the D70 items, 0.18 among the AG+ items, and 0.45 across the two constructs. Thus, the items within each construct were notably closer (and thus similar) to each other than the items across constructs. In sum, the MDS results corroborate our conclusion that D and AG+ are not equivalent.

4 | DISCUSSION

Recent research has worked towards specifying and determining the most suitable representation of the common disposition underlying aversive personality traits. This quest has spurred some controversy whether Big Five Agreeableness, per se, represents this common disposition. In a nutshell, Moshagen, Zettler, Horsten, et al. (2020) dissociated the Dark Factor of Personality (D) and low Agreeableness theoretically and empirically (see also Hilbig et al., 2021; Scholz et al., 2022), whereas Vize and Lynam (2020) and Vize, Miller, et al. (2020) argued that the low pole of Agreeableness is essentially equivalent to D and that Moshagen et al.’s dissociations were problematic because their operationalization of Agreeableness was incomplete. As a remedy, Vize, Miller, et al. (2020) pointed to a particularly broad operationalization of “Agreeableness”, which we herein term AG+. AG+ comprises content from, and is substantially related to, other Big Five and HEXACO personality dimensions beyond traditional Agreeableness in the sense of a largely orthogonal dimension within the Big Five (which is why the label “Agreeableness” is conducive to the jingle-fallacy and

<p>| TABLE 3 | Latent regression results predicting the outcomes by AG+ and D. |</p>
<table>
<thead>
<tr>
<th>Category</th>
<th>Criterion</th>
<th>$R^2_{D, AG+}$</th>
<th>$\Delta R^2_D$</th>
<th>$\Delta R^2_{AG+}$</th>
<th>Unique D</th>
<th>Unique AG+</th>
<th>ER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Justifying beliefs</td>
<td>Competitive Jungle Social World View</td>
<td>.74</td>
<td>.45</td>
<td>.05</td>
<td>61%</td>
<td>6%</td>
<td>&gt;.99</td>
</tr>
<tr>
<td></td>
<td>Normlessness</td>
<td>.52</td>
<td>.33</td>
<td>.04</td>
<td>63%</td>
<td>7%</td>
<td>&gt;.99</td>
</tr>
<tr>
<td></td>
<td>Social Dominance Orientation</td>
<td>.31</td>
<td>.17</td>
<td>.03</td>
<td>56%</td>
<td>11%</td>
<td>&gt;.99</td>
</tr>
<tr>
<td>Disutility on others</td>
<td>Pathological Selfishness</td>
<td>.63</td>
<td>.24</td>
<td>.10</td>
<td>39%</td>
<td>15%</td>
<td>&gt;.99</td>
</tr>
<tr>
<td></td>
<td>Exploitativeness</td>
<td>.60</td>
<td>.30</td>
<td>.08</td>
<td>50%</td>
<td>13%</td>
<td>&gt;.99</td>
</tr>
<tr>
<td>Affiliative tendencies</td>
<td>Extraversion</td>
<td>.05</td>
<td>.02</td>
<td>.03</td>
<td>42%</td>
<td>58%</td>
<td>.99</td>
</tr>
<tr>
<td></td>
<td>Withdrawal</td>
<td>.10</td>
<td>.04</td>
<td>.02</td>
<td>34%</td>
<td>18%</td>
<td>.90</td>
</tr>
<tr>
<td></td>
<td>Life History Strategy</td>
<td>.07</td>
<td>.02</td>
<td>.03</td>
<td>25%</td>
<td>49%</td>
<td>.89</td>
</tr>
<tr>
<td></td>
<td>Horizontal Collectivism</td>
<td>.23</td>
<td>.05</td>
<td>.14</td>
<td>21%</td>
<td>61%</td>
<td>&gt;.99</td>
</tr>
</tbody>
</table>

Note: $N = 1,156$. D = Dark Factor of Personality; AG+ = Broad blend of Big Five Agreeableness as per Crowe et al. (2018). $\Delta R^2_{D, AG+}$: variance explained in the full model. $\Delta R^2_{D, AG+}$: increase in $R^2$ after adding D to the model. $\Delta R^2_{AG+}$: increase in $R^2$ after adding AG+ to the model. Unique D/Unique AG+: relation of each construct’s unique contributions relative to the total variance explained in the criteria. ER: evidence in favor of the full model over the model not including the predictor hypothesized to explain incremental variance over the other. For justifying beliefs and disutility on others, the hypothesis was $\Delta R^2_{D, AG+} > 0$, for affiliative tendencies the hypothesis was $\Delta R^2_{D, AG+} > 0$. 

10 More precisely, this was a *broad* operationalization that included not only the low pole of Agreeableness but also the high pole, which is why we termed it AG+.
thus why we prefer to label it “AG+”; Hilbig et al., 2021). Nevertheless, AG+ may be an adequate approximation of the common core of aversive traits, that is, D.

Accordingly, we herein tested whether AG+ is essentially equivalent to D, sketching theoretical differences between the two constructs and thereby deriving criteria in which the constructs ought to account for unique variance components. Specifically, we hypothesized that D should account for unique variance in attitudes and beliefs that may serve as justifications for malevolent behaviors as well as the tendency to inflict disutility on others, whereas AG+ should account for unique variance in affiliative tendencies.

In a preregistered study, we found that D and AG+ shared only about 35% of variance. At first sight, this is incompatible with the 80% reported by Vize, Miller, et al. (2020). Importantly, though, the latter was based on a modeling approach yielding poor model fit and decreased to about 40% when estimating D and AG+ (more appropriately) as bifactor structures—thus mirroring the approach reported herein. As such, given equivalent modeling approaches, the findings are largely comparable. In substantive terms, although this magnitude of shared variance indicates a sizable overlap, it nonetheless speaks against unity of the two constructs. This conclusion was further corroborated by exploratory MDS analyses showing that each construct’s centroid, around which the respective items cluster, are located in different octants in the MDS space.

More importantly, each construct captured variance beyond the other in every one of the nine criteria tested herein. Specifically, as hypothesized, D accounted for unique variance in justifying beliefs, such as that one has to be ruthless to survive (Competitive Jungle World View) or that one is superior to others (Social Dominance Orientation), and in the willingness to inflict disutility on others, as expressed in the disregard for others’ needs (Pathological Selfishness) or the active exploitation of others (Exploitativeness) representing notable effect sizes throughout (between approximately 20% and 50% of additional variance explained beyond AG+). In turn, and again as hypothesized, AG+ accounted for unique variance in affiliative tendencies, such as the interest in (being with) people (Extraversion, low Withdrawal), forming emotional bonds (Life History Strategy) and feeling connected with others (Horizontal Collectivism). These conclusions were robust across alternative modeling approaches. In other words, D subsumed a broader range of aversive content than AG+, which, in turn, accounted for a somewhat broader range of non-aversive content related to affiliation. Notably, the portion of unique variance explained by D in aversive criteria (i.e., justifying beliefs and inflicting disutility on others) was much larger than the portion of unique variance explained by AG+ in affiliative criteria (median $\Delta R^2 = .28$ and .02, respectively). In substantive terms, D is much more strongly linked to justifying beliefs and inflicting disutility on others (than AG+), but AG+ is only marginally more saturated with affiliative tendencies than D.

The present results complement previous research in which D was dissociated from basic personality dimensions that had been proposed to represent the common core of aversive traits, especially Big Five Agreeableness and HEXACO Honesty-Humility (Hilbig et al., 2020; Horsten et al., 2021; Moshagen, Zettler, Horsten, et al., 2020; Scholz et al., 2022). Hence, neither Agreeableness, nor Honesty-Humility—as defined within their respective models of personality structure—nor their (extended) combination AG+ are equivalent to D.

A limitation of our study is that data were collected during a phase of the Covid-19 pandemic when it was generally imperative to restrict in-person meetings with others. This might particularly have affected participants’ responses on the Withdrawal scale used (e.g., “I prefer not to get too close to people”, or “I avoid social events”), potentially limiting the interpretability of the respective scale score. Given that correlations are invariant to additive shifts, however, this would only have affected the scale score, but arguably not the magnitude of correlations with D and AG+. A further limitation is that we did not include a criterion representing consequential behavior. Although our results show that D and AG+ are non-equivalent with respect to self-report measures, replicating such differences on the level of actual behavior would seem prudent.

## 5 | CONCLUSION

Although D and AG+ show noticeable overlap, the two constructs are not isomorphic. Ultimately, enriching Big Five Agreeableness with content from HEXACO Honesty-Humility, Agreeableness, and Altruism to become AG+ (Crowe et al., 2018) results in a construct incompatible with the Big Five framework (Hilbig et al., 2021), which—despite its added breadth—is non-equivalent to the common core of aversive traits, D. Although there may well be other advantages of AG+ as a construct and/or operationalization, it does not represent D.

## AUTHOR CONTRIBUTIONS

Idea, conceptualization, data collection, data analysis, data curation, manuscript writing: Luisa K. Horsten.
Idea, conceptualization, manuscript writing, supervision: Benjamin E. Hilbig.
Idea, conceptualization, manuscript writing: Isabel Thielmann.
Idea, conceptualization, manuscript writing: Ingo Zettler.

In our preregistration, we divided these tendencies into “interest in (being with) people” and “feelings of connectedness with others and an emotional warm glow”, but for simplicity will subsume them as affiliative tendencies herein.

Note that HEXACO Agreeableness vs. Anger and Big Five Agreeableness are rotated variants of each other (Ashton et al., 2014). Consequently, a recent meta-analysis found a correlation between these two of $r = .69$ only (Thielmann et al., 2021), suggesting that the two constructs are not equivalent. In Moshagen, Zettler, Horsten, et al. (2020), HEXACO Agreeableness vs. Anger was nonetheless included to represent Agreeableness as broadly as possible.

Vize, Miller, et al. (2020) reported around 80% shared variance when estimating D as a single factor from the observed scores on the 12 scales from which the D70 was compiled and, analogously, AG+ from the observed scores on the respective subscales. However, in the presence of a covariate, a bifactor structure generally yields the most accurate results as it is robust to secondary loadings and covariances between parts making up the hierarchical construct and the covariate (Moshagen, 2023). When relying on the bifactor approach to model D (as is recommended, Bader, Hartung, et al., 2021) and, analogously, modeling AG+ as a bifactor, the shared variance of AG+ and D notably decreases to around 40% in Vize et al.’s dataset, thus fully conforming to prior findings.

In our preregistration, we divided these tendencies into “interest in (being with) people” and “feelings of connectedness with others and an emotional warm glow”, but for simplicity will subsume them as affiliative tendencies herein.

If any of these criteria were met at the level of individual scales, responses on the respective scale were set to NA, thus treated as missing.

According to a post-hoc power analysis for a global hypothesis test (Jobst et al., 2021) of both bi-factor models—i.e., of D ($df = 2,275$) and of AG+ ($df = 5,148$)—the power to detect even negligible model misspecification (i.e., RMSEA $= .010$) at an $\alpha$-level of .05 and given the current $N$ was extremely high with more than 99.9% in both cases.

A bifactor model is preferred to the closely related higher-order model in the current investigation both for substantive and methodological reasons. For one, the bifactor model most closely mirrors the theoretical conceptualization of the common core of aversive traits because it assumes that the general factor directly accounts for the variance shared among all indicators. By contrast, the higher-order model assumes that the higher-order factor only indirectly explains the items’ shared variance through the lower-order factors. Additionally, the bifactor model provides unbiased estimates of the correlations between the general factor and covariates (Moshagen, 2023). For a more detailed reasoning, see Bader et al. (2021) and Moshagen et al. (2018).

Additionally, we ran our analyses with modified models in which those parameters yielding the largest modification indices were freed. This alternative approach led to the same conclusions. Corresponding analysis scripts and detailed results are provided in the additional materials on the OSF.

Model fits of the single factor models, as well as the corresponding scripts and analyses, can be found in the additional materials on the OSF. In summary, as indicated both by likelihood-ratio tests and evidence ratios, the bifactor model yielded the best fit to the data, both for D and for AG+ (see additional materials on the OSF).

Modeling AG+ from the manifest scores on the 20 Agreeableness (facet) scales (see Vize, Miller, et al., 2020) yielded poor fit to the data, $\chi^2(170) = 3,617, p < .001$, RMSEA $= .143$, 90% CI [.139, .147], $SRMR = .090$. The latent correlation between D and AG+ was $r = -.78$, which was significantly smaller than unity ($\Delta \chi^2(1) = 14.52, p < .001, ER > .999$). The pattern of results was largely consistent with the one presented herein, that is, both D and AG+ predicted incremental variance beyond the other in the hypothesized criteria. As such, the analytical approach reported by Vize, Miller, et al. (2020) confirms the present conclusions. The corresponding analysis script and results are provided in the additional materials on the OSF.


---

**How to cite this article:** Horsten, L. K., Thielmann, I., Moshagen, M., Zettler, I., Scholz, D., & Hilbig, B. E. (2023). Testing the equivalence of the aversive core of personality and a blend of agreeableness(-related) items. *Journal of Personality, 00*, 1–12. [https://doi.org/10.1111/jopy.12830](https://doi.org/10.1111/jopy.12830)