Journal of Research in Personality 85 (2020) 103915

Contents lists available at ScienceDirect

Journal of Research in Personality

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

Full Length Article

The unfolding dark side: Age trends in dark personality features

Theo A. Klimstra [a,*], Bertus F. Jeronimus [b,c], Jelle J. Sijtsema [a], Jaap J.A. Denissen [a]

[a] Department of Developmental Psychology, Tilburg University, the Netherlands
[b] Department of Developmental Psychology, Faculty of Behavioural and Social Sciences, University of Groningen, the Netherlands
[c] Interdisciplinary Center Psychopathology and Emotion Regulation (ICPE), University of Groningen, University Medical Center Groningen, the Netherlands

ARTICLE INFO

Article history:
Received 18 July 2019
Revised 11 January 2020
Accepted 13 January 2020
Available online 16 January 2020

Keywords:
Dark features
Personality
Age trends
Lifespan developmental perspective
Adolescence
Adulthood

ABSTRACT

Age and gender differences across the lifespan in dark personality features could provide hints regarding these features' functions. We measured manipulation, callous affect, and egocentricity using the Dirty Dozen and their links with agreeableness in a pooled cross-sectional dataset (N = 4292) and a longitudinal dataset (N = 325). Age trends for all dark personality features were progressive through adolescence, but negative through adulthood. Men scored higher than women, but the gender gap varied with age. Trends for agreeableness partly mirrored these trends and changes in dark personality features and agreeableness were correlated. Results are discussed in light of the maturity principle of personality, gender role socialization processes, and issues regarding incremental validity of dark personality over traditional antagonism measures.

1. Introduction

Since the turn of the century, a large number of studies started examining dark personality features in the general population. These features typically reflect tendencies towards self-promotion and callous and manipulative interpersonal behavior (Paulhus & Williams, 2002). Over the last decade, studies linked such features to a wide array of outcomes variables, including workplace behavior, antisocial behavior, and mating behavior (for a review, see Furnham, Richards, & Paulhus, 2013). In this field of research, there is evidence for gender differences (with higher values for men), but these have often not been assessed while accounting for measurement issues. Furthermore, age differences have received much less attention compared to age differences in Big Five personality features. Proper measurement of gender differences in dark features as well as zooming in on age differences and interactions between gender and age differences would contribute to a better understanding of the normative expressions of narcissism (e.g., egocentrism), Machiavellianism (e.g., manipulation), and psychopathy (e.g., callous affect). In this study we employed data on 4292 individuals with an age range from 11 to 77 years to explore gender and age differences in dark personality features.

1.1. The Dirty Dozen as a measure of dark personality

Several measures have been developed to capture dark personality features. Many of these measures sought to capture the so-called Dark Triad, which consists of the interrelated features of narcissism, Machiavellianism, and psychopathy. These features have traditionally been assessed with separate measures, but after 2010 several measures were developed to assess the whole Dark Triad. Among the most frequently used of these is the Dirty Dozen (Jonason & Webster, 2010). The Dirty Dozen scales are internally consistent, its items function well, and its intended factor structure has been confirmed in several studies (e.g., Chiorri, Garofalo, & Velotti, 2017; Czarna, Jonason, Dufner, & Kossowska, 2016; Jonason & Webster, 2010; Klimstra, Sijtsema, Henrichs, & Cima, 2014; Webster & Jonason, 2013). Typically, the three scales are considered separately, but various studies also consider a general Dirty Dozen factor, modeled by means of a bifactor model (Czarna et al., 2016) or a hierarchical model (Jonason & Webster, 2010). However, bifactor models have been criticized for various reasons, including the general factor being uninterpretable and the superior fit being a symptom of overfitting (e.g., Bonifay, Lane, & Reise, 2017). Hierarchical models also come with problems. Specifically, the higher-order factor removes meaningful variance from the lower-order dimensions (e.g., the subscales), because the empirical overlap between these dimensions is modelled into the higher-order factor. Recent research using the Dirty Dozen measure suggested that working with residualized constructs can

* Corresponding author at: Department of Developmental Psychology, Tilburg University, Postbus 90153, 5000 LE Tilburg, the Netherlands.
E-mail address: t.a.klimstra@tilburguniversity.edu (T.A. Klimstra).

https://doi.org/10.1016/j.jrp.2020.103915
0092-6566/

lead to validity issues causing associations with outcome variables to be non-replicable (Vize, Collison, Miller, & Lynam, in press A). Thus, a three-factor structure likely provides the most valid representation of the Dirty Dozen. Still, a general factor onto which all items load could be of interest for examining whether constructs such as agreeableness-antagonism are indeed at the core of the Dirty Dozen (e.g., Lynam & Miller, 2019).

In the present manuscript we set out to use the Dirty Dozen as a measure of the broader Dark Triad features. However, anonymous peer reviews and literature pointing to the limitations of the Dirty Dozen (e.g., Maples, Lamkin, & Miller, 2014; Muris, Merckelbach, Otgaar, & Meijer, 2017; Vize, Lynam, Collison, & Miller, 2018) caused us to change our focus. Specifically, there is a growing number of studies pointing out that the Dirty Dozen only covers a limited part of each of the broader Dark Triad features which are multidimensional in themselves. For example, one way to subdivide narcissism is to distinguish rivalry and admiration (Back et al., 2013). Consequently, few researchers would consider narcissism to be a unidimensional construct. Similarly, psychopathy consists of multiple components (e.g., Miller et al., 2012). In addition, Machiavellianism and psychopathy are not particularly separable, especially not with brief Dark Triad measures (Maples, Lamkin, & Miller, 2014; Miller, Hyatt, Maples-Keller, Carter, & Lynam, 2017). Therefore, brief measures of the Dark Triad, such as the 12-item Dirty Dozen (Jonason & Webster, 2010) can be criticized for not measuring narcissism, Machiavellianism, and psychopathy, but related constructs (e.g., Muris et al., 2017).

Hence, the Dirty Dozen scales are likely best interpreted as proxy measures of antagonistic personality dimensions that are narrower than broad dimensions such as the Dark Triad. Thus, the scales' bandwidth may be more or less similar to scales from measures such as the Self-Report Psychopathy Scale (e.g., Neumann, Hare, & Pardini, 2014). To further illustrate this point, we listed the items in Table 1.

Table 1 shows that some reference to manipulation or a particular manipulation strategy is made in 3 of the 4 items intended to measure Machiavellianism (items 1, 2, and 3). Therefore, these items can be regarded as indicators of interpersonal tactics (Muris et al., 2017). The fourth item has been described as an indicator of disregard for conventional morality (Muris et al., 2017), but is also associated with a goal linked to malevolent manipulation: exploitation. Given that Machiavellianism scales such as the Mach-IV tend to cover a variety of content (e.g., Rauthmann, 2013), a scale with a focus limited to manipulation is therefore better described as a manipulation scale rather than a Machiavellianism scale. All four items that were intended to assess psychopathy have been described as indicators of callous affect (Muris et al., 2017). Other psychopathy-relevant content, such as disturbed interpersonal and lifestyle features, and antisocial behavior (e.g., Neumann et al., 2014), are not covered with these items. The items that were intended to assess narcissism all focus on exhibitionism, superiority/grandiosity, and entitlement (Muris et al., 2017). Narcissism features reflecting rivalry (e.g., Back et al., 2013) or leadership (e.g., Wetzel et al., 2017) are not assessed with these items and thus egocentricity seems the common denominator among the current items. Therefore, previous research and an inspection of the items strongly suggest that the Dirty Dozen scales are more accurately described as measures of egocentricity (i.e., desire for others' attention and recognition), callous affect, and manipulation (e.g., Maples et al., 2014; Muris et al., 2017; Vize et al., 2018).

Table 1
Dirty dozen items.

Item		Scale Name in Present Manuscript	Original Scale Name
1.	I tend to manipulate others to get my way.	Manipulativeness	Machiavellianism
2.	I have used deceit or lied to get my way.	Manipulativeness	Machiavellianism
3.	I have used flattery to get my way.	Manipulativeness	Machiavellianism
4.	I tend to exploit others towards my own end.	Manipulativeness	Machiavellianism
5.	I tend to lack remorse.	Callous Affect	Psychopathy
6.	I tend to be unconcerned with the morality of my actions.	Callous Affect	Psychopathy
7.	I tend to be callous or insensitive.	Callous Affect	Psychopathy
8.	I tend to be cynical.	Callous Affect	Psychopathy
9.	I tend to want others to admire me.	Egocentricity	Narcissism
10.	I tend to want others to pay attention to me.	Egocentricity	Narcissism
11.	I tend to seek prestige of status.	Egocentricity	Narcissism
12.	I tend to expect special favors from others.	Egocentricity	Narcissism

Note. We administered the Dutch-language items in all samples, but provide their English-language equivalents (Jonason & Webster, 2010) as the Dutch-language items are likely uninterpretable for a majority of readers. The Dutch-language items are provided in the Supplementary Material, Section 1.

There is a large literature examining the correlates of the Dirty Dozen's scales, which include impulsivity, antisocial behavior, mating behavior, and dimensions of major personality models (Vize et al., 2018). For example, honesty-humility, as represented in the HEXACO model, is strongly related to all three Dirty Dozen scales (Muris et al., 2017). A meta-analysis on studies using the big five suggests that the Dirty Dozen narcissism scale (i.e., egocentricity) is a combination of high extraversion and (to a lesser extent) high neuroticism with low agreeableness (Vize et al., 2018). Both the psychopathy (i.e., callous affect) and Machiavellianism (i.e., manipulation) scales appear to combine low agreeableness and conscientiousness (with small differences between the two in their associations with other personality features).

Based on these findings, one could argue that broad features covered by the HEXACO and big five models already cover the content of the Dirty Dozen measure, making the Dirty Dozen scales redundant. Especially the agreeableness-antagonism dimension has been pointed to as a potential candidate sufficiently capturing a large share of the variance of particular dark features such as psychopathy (Sherman, Lynam, & Heyde, 2014) and broader collections of dark features such as the Dark Triad (Lynam & Miller, 2019). Specifically, out of the Big Five and HEXACO, agreeableness and honesty-humility are strongly associated with the shared variance among dark features captured with the Dirty Dozen (Vize, Collison, Miller, & Lynam, in press B). However, these associations pertain to the shared variance among all Dirty Dozen scales. Therefore, one could argue that the Dirty Dozen scales offer specific facet-level operationalizations of egocentric, callous, and manipulative antagonistic tendencies.

The current paper aims to contribute to this discussion by examining gender differences, age trends, and gender by age interactions of the Dirty Dozen scales. We directly compare these to age trends of the most likely Big Five feature for explaining their variance: agreeableness. If gender and age trends of the Dirty Dozen scales differ from those of agreeableness, this would suggest that measures of dark personality dimensions may have added value compared to only considering agreeableness.

1.2. Gender differences

One frequently examined question regarding Dirty Dozen scales is the existence of gender differences in the features they represent. Next to gender differences in big five personality features

related to the Dirty Dozen scales, there are at least three additional arguments to expect gender differences. First, callous affect and manipulation are directly related to antisocial personality disorder (ASPD; American Psychiatric Association, 2013; Few, Lynam, Maples, MacKillop, & Miller, 2015), which is more prevalent in men than in women (Oltmanns & Power, 2012). Gender differences in the prevalence of narcissistic personality disorder (NPD), of which egocentricity is a key aspect, are less clear (Oltmanns & Powers, 2012). However, a meta-analysis found clear evidence for higher narcissism scores among men than women in a non-clinical sample (Grijalva et al., 2015). This suggests that men might also exhibit higher levels of egocentricity when compared to women.

Second, evolutionary theory predicts gender differences in dark personality features because they associate with measures reflecting short-term mating strategies and having more sex partners (Jonason, Li, Webster, & Schmitt, 2009; Dufner, Rauthmann, Czarna, & Denissen, 2013). Short-term mating strategies were on average more costly for women than for men, thus a more restrictive mating strategy would be relatively more adaptive for women (Schmitt, Realo, Voracek, & Allik, 2008), everything else equal. As features related to a short-term strategy may have evolved accordingly, we expect women to have lower Dirty Dozen scale scores than men.

Third, theories on gender role socialization may explain why men would score higher on dark features than women (e.g., West & Zimmerman, 1987). Gender roles are socialized from childhood onwards (Fagot, Hagan, Leinbach, & Kronsberg, 1985), for example dampening boys' initial emotional expressivity (Rydell, Berli, & Bohlin, 2003). Cultural norms also suppress the expression of assertive action and outward expression of anger more in women than in men (Chaplin, 2015; Chaplin & Aldao, 2013). Narcissism features such as egocentricity have been linked to the masculine stereotype (Grijalva et al., 2015), much like callous affect also reflects a stereotypical masculine tendency. This could potentially lead to gender differences in egocentricity and callous affect. For manipulation, stereotypes seem to imply that women do this at least as much as men do (e.g., Björkqvist, Lagerspetz, & Kaukiainen, 1992), although sex differences in physical strength may stimulate women to engage in more emotional and verbal, rather than physical manipulation strategies. This suggests no gender differences in manipulation. These gender role socialization theories have been challenged based on cross-cultural evidence suggesting that gender differences in personality tend to be larger rather than smaller in more gender-equal societies (Schmitt, Long, McPhearson, O'Brien, Remmert, & Shah, 2017). Nonetheless, our views align with those of scholars opposing reductionist views suggesting splits of nature versus nurture, genetic versus environmental effects, or in this case evolutionary versus socialization effects (e.g., Lerner & Overton, 2017). Thus, we assume that both socialization and evolutionary factors co-act to produce gender differences.

Meta-analytic empirical evidence suggests that men tend to score higher than women on all Dirty Dozen scales (Muris et al., 2017). Research also typically suggests that gender differences are larger for callous affect and other psychopathy-related scales than for Machiavellianism- and narcissism-related scales (Muris et al., 2017; Schmitt, Long, McPhearson, & O'brien, K., Remmert, B., & Shah, S. H. , 2017). However, appropriate statistical considerations, such as establishing measurement invariance before conducting gender comparisons (van de Schoot, Lugtig, & Hox, 2012) typically are not accounted for in gender comparisons on the Dirty Dozen scales (for exceptions, see Chiorri et al., 2017; Klimstra et al., 2014). Furthermore, it remains unclear whether the gender gap is equal in all age groups because age differences are rarely considered, even though such effects may partly explain differences between studies. Reviews and meta-analyses on the Dark Triad and Dirty Dozen, for example, do not even mention age as a variable of interest. In addition, most studies on the Dirty Dozen only included young adults (e.g., college students and/or young workers; Jonason & Webster, 2010; Czarna et al., 2016), who may differ in many ways from adolescents and late adults (cf. Roberts, Walton, & Viechtbauer, 2006). To provide an appropriate background for understanding what age-related differences in the magnitude of gender differences may look like, mean-level age trends need to be discussed first.

1.3. Age differences

Research specifically focusing on age trends in dark features is rare, as we found no studies examining age trends from adolescence through adulthood. However, there are strong external indications for mean-level differences in these features. For example, the Dirty Dozen scales are linked to the big five (Vize et al., 2018) and HEXACO scales (Muris et al., 2017). These models show decreases in maturity-related features, such as agreeableness, conscientiousness, and honesty-humility in early adolescence and increases in those features from middle or late adolescence into adulthood (e.g., Ashton & Lee, 2016; Denissen, Van Aken, Penke, & Wood, 2013; Roberts et al., 2006; Soto, John, Gosling, & Potter, 2011). There are negative links of agreeableness, conscientiousness, and honesty-humility with dark features (Muris et al., 2017; Vize et al., 2018). Low honesty-humility has been interpreted as the tendency to actively exploit others (Ashton & Lee, 2007), which is something it shares with the Dirty Dozen scales. This would suggest that mean-levels of these features increase during early adolescence and plateau or even a decrease after middle adolescence and through adulthood.

Hence, generalizing from age correlates of dimensions in major personality models, the prediction can be derived that mean-levels of the Dirty Dozen scales will be positively associated with age in adolescence, but negatively in adulthood. There is empirical support for the latter part of this hypothesis, as one study showed lower Dirty Dozen scale means in older employees (aged 50–59) compared to younger employees (aged 25–34) (Spurk & Hirschi, 2018), and other studies found a negative association of age with Dirty Dozen scales in adult samples (Barelds, 2016; Craker & March 2016; Fox & Rooney, 2015). However, we are unaware of studies on the association of age with the Dirty Dozen that considers a broader age range to allow for non-linear age trends or studies directly comparing age trends in the Dirty Dozen to those in relevant related dimensions, such as agreeableness. In addition, only one of the aforementioned studies (Spurk & Hirschi, 2018) examined measurement invariance between age groups.

Of the studies that examined associations between age and the Dirty Dozen scales, none seemed to examine another possibility: Moderation of gender differences by age. Age-related changes in the gender gap for the Dirty Dozen scales are likely for several reasons. First, the gender intensification hypothesis postulates increasing gender differences throughout adolescence, but it has received mixed support (Steensma, Kreukels, de Vries, & Cohen-Kettenis, 2013). Soto et al. (2011) only found gender differences in the big five personality features from late childhood and early adolescence onwards, suggesting that gender differences in personality emerge over adolescence. We expect gender differences in the Dirty Dozen scales also to be increasingly larger the older the adolescents are. During middle adulthood gender roles may stabilize or even intensify, but later in life and especially after retirement, men typically also take on more caring and stereotypically feminine roles (Arber, Davidson, & Ginn, 2003). In line with this hypothesis, the gender gap in big five features tends to be smaller in older adults (Soto et al., 2011), which also suggests a

decreasing gender gap in the Dirty Dozen scales in adulthood. For narcissism constructs, a meta-analysis suggested stable gender differences (Grijalva et al., 2015). However, their estimate was virtually restricted to student samples, and gender differences were predicted with average participant age rather than individual participant age, which is rather crude. We are unaware of research examining the gender gap in manipulation and callous affect across the life course. To fill this gap in the literature on age and gender differences in the Dirty Dozen scales, we conducted two studies.

2. Study 1: Cross-sectional age trends

Study 1 had two aims: to examine (a) gender differences in different age groups and (b) mean-level age trends in different gender groups. In all of our analyses, we also examined whether gender differences and mean-level age trends in the Dirty Dozen scales resembled those observed in agreeableness. For this purpose, we ran a large scale cross-sectional data pooling study (N = 4292, k = 12) with an age range from 11 to 77 years. All participants were drawn from Dutch-speaking populations and divided into six age cohorts: early adolescence (ages 11–13 years), middle adolescence (ages 14–16 years), late adolescence (ages 17–18 years), young adulthood (ages 19–30 years), middle adulthood (ages 31–54 years), and late adulthood (ages 55–77 years). Although cutoffs are always arbitrary and there is no full uniformity of categorizations in the literature, the age groups we used do represent commonly distinguished developmental stages in adolescence (e.g., Flanagan & Stout, 2010) and adulthood (e.g., Ebner, Freund, & Baltes, 2006). We made some alterations (i.e., we defined late adulthood as starting at age 55 rather than 60) to have sufficiently large groups to run age comparisons.

First, we examined mean-level gender differences in each age group via latent means to partly account for potential gender differences in measurement properties (cf. van de Schoot et al., 2012). In line with previous studies correcting for measurement issues, we expected higher scores for men than women on the Dirty Dozen and its scales (Chiorri et al., 2017). This gender gap was expected to be largest in young and middle adults, and largest for callous affect.

Second, we examined age trends in latent mean-levels for the Dirty Dozen scales, accounting for age-related differences in measurement properties. Given the anticipated gender differences, we examined age trends by gender group. Based on evidence from personality theories and big five trends we expected early adolescents to score lower than middle adolescents. Among the adult age groups, we expected mean levels on the Dirty Dozen and its scales to be higher in the younger than the older age groups.

2.1. Methods

2.1.1. Participants

Twelve datasets from the Netherlands and the Dutch-speaking part of Belgium (Flanders) were pooled (see Table 2). We only considered participants who completed half of the items on at least one of the three Dirty Dozen scales, which eliminated 35 of our 4330 original participants (<1%). Three multivariate outliers on the observed mean scores of the Dirty Dozen scales were also excluded based on Mahalonobis' distance values. These values reflect the distance of participants' scores from the center of the multidimensional distribution. The final sample consisted of 4292 participants (39.4% men, 60.6% women). Participants ranged in age from 11 to 77 years (M_{age} = 28.54 years, SD = 16.99). We split them into six age groups as outlined above (see Table 3). The pooled data are available at https://osf.io/pfd8b/?view_only=2fd77cf5d2c349859ccbaae448e744bf. Note that a file with the dataset numbers is available from the authors because the European privacy law (the GDPR) prohibits posting identifiable mental health information on public repositories.

Table 2
Sample characteristics.

Sample	N	% Women	Age range in years	Mean age (SD) in years	Online or Paper-and-Pencil	Ethical Protocol Number	Agreeableness/Honesty-Humility data
1	165	57.0	14–18	16.08 (0.74)	Online	EC-2013.07	PID-5
2	202	68.3	14–19	16.65 (0.81)	Paper-and-Pencil	n.a.	BFI-44
3	236	78.0	18–67	30.14 (12.74)	Online	n.a.	BFI-25
4	220	54.5	14–23	16.54 (1.22)	Paper-and-Pencil	n.a.	HEXACO-100
5	369	68.6	15–57	30.96 (15.36)	Paper-and-Pencil	n.a.	BFI-44
6	163	1.8	19–61	33.86 (11.25)	Online	n.a.	BFI-25
7	1,169	74.6	18–77	46.74 (14.00)	Online	M13.147422	NEO-FFI-3
8	351	70.7	15–76	27.54 (14.00)	Paper-and-Pencil	n.a.	BFI-44
9	150	4.0	16–62	28.53 (11.61)	Online	n.a.	BFI-25
10	92	81.5	38–56	46.01 (4.16)	Online	EC-2014.03	None
11	305	48.2	11–15	12.79 (0.77)	Paper-and-Pencil	EC-2013.09	BFI-44
12	870	53.1	11–18	13.86 (1.09)	Online	EC-2014.03	BFI-25

Note. If the ethical protocol number says n.a., this means that these datasets were collected at <BLINDED FOR REVIEW> before ethical approval was required. When collecting these datasets, we followed procedures that were highly similar to the procedures we followed when collecting datasets for which we did request and obtain ethical approval. PID-5 = Personality Inventory for DSM-5; BFI-44 = 44-item version of the Big Five Inventory; BFI-25 = 25-item version of the Big Five Inventory; HEXACO-100 = 100-item version of the HEXACO-PI-R; NEO-FFI-3 = NEO Five-Factor Inventory 3.

Table 3
Descriptive statistics by age group.

	N	Age Range in years	M_{age} (SD) in years	% Women
Early Adolescence	582	11–13	12.63 (0.50)	47.8
Middle Adolescence	1069	14–16	15.08 (0.86)	56.3
Late Adolescence	516	17–18	17.27 (0.44)	54.3
Young Adulthood	610	19–30	24.13 (3.14)	52.1
Middle Adulthood	1047	31–54	44.88 (6.09)	79.3
Late Adulthood	468	55–77	60.69 (4.74)	62.8

2.1.2. Procedure

All studies were conducted in accordance with the guidelines of the local institutional review boards (see Table 2 for ethical approval numbers of the various studies, if available). For the samples that were approached through high schools (Samples 1, 2, 4, 5, 8, 11, and 12), we first obtained permission from school principals to administer questionnaires during class. Parents were informed via a detailed letter describing the study content and goals, and were given the opportunity to object to their children's participation. After we received parental permission, students were informed about the study and asked whether they wished to participate, which they all did. They were supervised by Psychology master students while filling out the questionnaires.

Some of our high school student samples (Samples 4, 5, and 8) also included the parents of the participating adolescents. These parents reported on their own personality. Similar to data collection in the other samples including participants over 18 years old (Samples 3, 6, 7, 9, and 10), adult participants were informed about the study, and asked whether they wished to participate. They filled out the questionnaires independently, in their home environment.

A non-significant multivariate three-way interaction effect of age by gender by sample in a Multivariate Analysis of Variance suggested that age and gender differences were not confounded with sample differences ($F_{(36,\ 12434)}$ = 1.38, p = .07). In addition, effects of assessment method (i.e., online versus paper-and-pencil participants) were not significant for callous affect and manipulation and small for egocentricity (see Supplementary Section 2 for details).

2.1.3. Measures

Dark Personality. In all samples, dark personality features were self-reported by participants using the Dutch-language version (Klimstra et al., 2014) of the 12-item Dirty Dozen (Jonason & Webster, 2010). The three Dirty Dozen scales, intended to measure Machiavellianism, psychopathy, and narcissism, can more appropriately be described as measures of manipulation, callous affect, and egocentricity (e.g., Maples et al., 2014; Muris et al., 2017; Vize et al., 2018). Each scale consists of 4 items rated on a 9-point scale ranging from 1 ('strongly disagree') to 9 ('strongly agree'), but a 12-item general Dirty Dozen score can also be examined. On all items, a higher score is indicative of higher levels on the respective features. The full list of items, in English and in Dutch, is provided in the Supplementary Material Section 1. In the total sample, coefficient alphas of the manipulation, callous affect, egocentricity scales, and the general Dirty Dozen factor were 0.78, 0.72, 0.84, and 0.86, respectively. In separate samples, coefficient alphas for manipulation ranged from 0.64 to 0.87, for callous affect from 0.66 to 0.84, for egocentricity from 0.81 to 0.89, and for the general Dirty Dozen scale from 0.83 to 0.92.

Personality Constructs Related to the Dirty Dozen. We were able to examine whether the Dark Personality trajectories were unique or simply mirrored agreeableness and/or honesty-humility age trends to some extent, because we had agreeableness data available for several samples. As shown in Table 2, honesty-humility was only available in one relatively small dataset with a restricted age range (Sample 4). Data on DSM-5-related personality dimensions were only available in Sample 1 (the PID-5; Krueger, Derringer, Markon, Watson, & Skodol, 2012). Although we were not able to assess age trends for these scales, we did examine their correlations with the Dirty Dozen constructs. Manipulation, callous affect, ego-centricity, and the Dirty Dozen total score were significantly correlated with PID-5 antagonism (rs are 0.58, 0.51, 0.47, and 0.61, respectively), with HEXACO agreeableness (rs are −0.42, −0.33, −0.25, and −0.42, respectively), and with honesty-humility (rs are −0.60, −0.37, −0.52, and −0.63, respectively).

Self-reports on Agreeableness were completed by participants representing 9 of the 12 samples. In one of these samples (Sample 7), the 12-item subscale of the Dutch-language version of the NEO-FFI-3 (De Fruyt & Hoekstra, 2014) was employed. The items of the NEO-FFI-3 (e.g., 'If I don't like people, I let them know it') were scored on a 5-point scale, from 1 (strongly disagree) to 5 (strongly agree). Coefficient alpha was 0.73 for this scale. In the other 8 samples, a Dutch-language version (Denissen, Geenen, van Aken, Gosling, & Potter, 2008) of the Big Five Inventory (BFI; John, Donahue, & Kentle, 1991) was used. In for 4 of these 8 samples, participants completed the original 9-item agreeableness scale of this measure, whereas in the other 4 samples participants completed a shortened 5-item agreeableness subscale of the BFI-25 (e.g., Boele, Sijtsema, Klimstra, Denissen, & Meeus, 2017). The items of the BFI (e.g., 'Is considerate and kind to almost everyone') are scored on a 5-point scale, ranging from '1′ (completely untrue) to '5′ (completely true). To make scores comparable across these 8 samples, we created scale scores based on the 5 agreeableness items that are included in both the BFI-44 and the BFI-25. The internal consistency for the 5-item version of the scale was acceptable and ranged from 0.60 to 0.68 for 7 out of 8 samples, but it was 0.51 in the eight sample (note that there were no significant negative inter-item correlations). Note that the original 9-item and shortened 5-item version of the agreeableness scale were strongly correlated (r = 0.93) across the four samples for which we had BFI-44 data.

2.1.4. Strategy of analyses

As preliminary analyses, we first established whether the three-factor structure of the Dirty Dozen, the general Dirty Dozen factor, and the agreeableness measures were similar across gender and age groups. For this purpose, we ran Confirmatory Factor Analyses (CFAs) in Mplus 7 (Muthen & Muthen, 2012) using Maximum Likelihood Robust (MLR) estimation to examine configural, metric, and scalar measurement invariance (van de Schoot et al., 2012). MLR is the most accurate estimator when the distribution of scores deviates from a normal distribution (Satorra & Bentler, 1994), which was the case with the scores on the subscales. Configural invariance concerns the question of whether the same confirmatory factor model has an acceptable fit to the data in different groups. However, when restricted to evaluations of absolute fit indices, configural invariance tests are rather weak tests especially if there are more parsimonious plausible alternative models. In the case of the Dirty Dozen, the validity of distinctions between dark personality features has been shown to be disputable (e.g., Maples et al., 2014; Miller et al., 2017) and one-factor as well as two-factor models are therefore plausible. Hence, we compared such one- and two-factor models to the proposed three-factor model and examined whether the three-factor model outperformed those alternative models in all the groups that we distinguished. For unidimensional constructs with no plausible alternative factor structures (i.e., the agreeableness measure), such tests were not conducted.

The following two types of invariance were relevant to all measures that we used. Specifically, metric invariance (or strong invariance) refers to factor loadings being statistically equivalent across groups. Finally, scalar invariance (or strict invariance) refers to intercepts of items being statistically equivalent across groups in addition to the factor loading being equivalent. If evidence for full metric and/or scalar invariance is lacking, partial invariance can be tested. In such cases, the factor loadings (in case of metric invariance tests) or item intercept (in case of scalar invariance tests) for at least two indicators of a latent factor need to be equal across groups, but invariance constraints on the other items can be released (Steenkamp & Baumgartner, 1998). Thus, partial invariance in the case of the present study indicates that factor loadings

or intercepts of between 2 and all but one items per scale are constrained to be equal across groups.

We examined measurement invariance between gender groups in the different age groups, and across age groups for men and women, separately. Configural invariance tests showed that the intended three-factor structure of the Dirty Dozen fitted the data better than alternative one- and two-factor models (see Supplementary Section 3). Only in early adolescent girls, a two-factor model (with a combined manipulation-callous affect factor) was equivalent to the three-factor model. Because this finding only concerned 1 out of 12 age by gender groups, we decided to use three-factor models for all age and gender groups to facilitate between-group comparisons. Given the interest in the core of the Dirty Dozen and Dark Triad as a measure of antagonism, we also ran analyses on a one-factor Dirty Dozen model. It should be noted that analyses with this one-factor model yielded a poorer fit because the data were better represented by a three-factor structure.

Further tests provided evidence for full metric invariance of the three-factor Dirty Dozen model (i.e., factor loadings being equal) across gender in all six age groups. Full scalar invariance across gender was only observed in the middle and late adolescent groups. Partial (scalar) invariance across gender was observed for the early adolescence group, the middle adulthood group, and the late adulthood group. In young adults partial scalar invariance was not observed and mean-level gender differences were therefore not interpreted (see Supplementary Section 4.1 for details). Invariance tests for age showed full or partial scalar invariance between each pair of adjacent age groups (e.g., early and middle adolescence, or young and middle adulthood), for both men and women (see Supplementary Section 5.1). There often was no scalar invariance between non-adjacent age groups (e.g., early adolescence and late adulthood groups), hence mean-level differences were only interpreted between adjacent age groups.

For the general Dirty Dozen factor (i.e., the one-factor model), we found metric invariance and partial scalar invariance between gender groups in all age groups. However, the detailed description of these analyses in Supplementary Section 4.2 shows that constraints sometimes needed to be released on a large number of items to achieve partial scalar invariance (e.g., for 8 out of 12 items in the young adulthood group). Analyses presented in Supplementary Section 5.2 show that we found metric invariance and partial scalar invariance in all adjacent age groups and even between non-adjacent age groups (i.e., between early and late adolescence and between early and late adulthood) for both men and women.

For the BFI agreeableness scale, we found evidence for partial metric and scalar invariance for all gender comparisons (see Supplementary Material Section 4.3). We also found evidence for at least partial metric and partial scalar invariance between most adjacent age groups for men (e.g., between the early adolescence and middle adolescence groups), except for between the young adulthood and middle adulthood group (see Supplementary Material Section 5.3). Hence, mean-level differences were only interpreted between adjacent age groups, except for between the young adulthood and middle adulthood group.

Finally, for the NEO-FFI-3-agreeableness scale, we found evidence for metric and partial scalar invariance in women across the three adult age groups and in men across the two adult age groups for which we had sufficient data to run comparisons. We also found (partial) metric and partial scalar invariance between gender groups in the two adult age groups on which we had sufficient data to run comparisons. Details on invariance tests are presented in Supplementary Material Sections 4.4 and 5.4.

Our main research questions were addressed using latent mean comparisons of gender scores within and between all age groups using structural equation modelling. In each model, we kept invariance constraints in place to attain valid latent means. Specifically, we found partial scalar invariance in several models, which means that scale means based on simply averaging the items could have introduced bias, whereas latent variables indicated by items are less biased in such cases (Steinmetz, 2013). The young adulthood model for gender differences was not interpreted, as there was no measurement invariance across gender in that age group. In all models, men were the reference group with a mean score of zero on all scales. A score above zero indicates that women had higher mean levels than men while a score below zero would indicate women had lower mean levels than men. Because the variances of the latent factors were constrained to 1, these latent mean gender differences can be interpreted as Cohen's *d* effect sizes (Steinmetz, 2010).

Age differences within gender groups were also examined using latent mean comparisons in a structural equation modeling framework. Our model set up was based on our measurement invariance results. Hence, we ran three models (i.e., one for all adolescent groups, one for the late adolescent and young adult group, and one for all the adult groups) to examine age trends in latent means for men for both the general Dirty Dozen factor and its scales. For women, we ran five separate models in which we compared latent mean differences between each pair of adjacent age groups (i.e., early versus middle adolescence, middle versus late adolescence, late adolescence versus young adulthood, young adulthood versus middle adulthood, and middle adulthood versus late adulthood), again for both the general Dirty Dozen factor and its three scales. For agreeableness, we only compared adjacent age groups for which sufficient data was available (i.e., $n > 100$).

Note that we ran our models in the order of increasingly old age groups, which allowed for a cumulative approach of setting reference values. In each pairwise age comparison, the respective youngest group of the pair was always the reference group. For the first model comparing early adolescence with middle adolescence, we fixed the latent means for the reference group to 0. When comparing older age groups, we used the estimated value of the previous model. For example, we found in our first model that the latent mean for men in the late adolescent age group was 0.62. Therefore, the latent mean for men of the late adolescent group was constrained to 0.62 in the second model, in which late adolescent men were compared to young adult men. By using this approach, the reader gets a better idea of the age trends across all groups, despite that there is no measurement invariance across all groups.

2.2. Results

All the one-factor Dirty Dozen models we used for gender comparisons had a poor fit to the data see Supplementary Table 4.14), whereas almost all of the three-factor Dirty Dozen models and agreeableness models (see supplementary Table 4.7 for the Dirty Dozen and Table 4.20 and 4.23 for BFI agreeableness and NEO-FFI agreeableness., respectively) had an acceptable fit. The three-factor model for middle adolescents showed a CFI just below the 0.90 benchmark, which warrants a cautionary interpretation. In the young adulthood group, there was no measurement invariance across gender in the three-factor model. Hence, no gender difference is reported for that age group. The resulting latent mean comparisons by gender are presented in Table 4.

Women reported lower means on all three Dirty Dozen scales and the general Dirty Dozen factor compared to men (see Table 4), although the gender gap varied with age and by characteristic. A relatively small gender gap was reported in middle adolescence age group, with larger gender differences in the early adolescence and middle adulthood age groups. Only in the middle adolescence age group, women were equally callous as men. For egocentricity,

Table 4
Gender differences in latent means across and within age groups.

	Manipulation Gender Differences		Callous Affect Gender Differences		Egocentricity Gender Differences		General Dirty Dozen Factor Gender Differences		Agreeableness Gender Differences	
E. Ado	−0.42***	(−0.62, −0.22)	−0.37***	(−0.57, −0.17)	−0.31**	(−0.51, −0.11)	−0.43***	(−0.61, −0.24)	−0.02[a]	(−0.21, 0.18)
M. Ado	−0.20**	(−0.34, −0.06)	−0.14	(−0.28, 0.01)	−0.15*	(−0.28, −0.01)	−0.19**	(−0.32, −0.06)	0.11[a]	(−0.10, 0.32)
L. Ado	−0.34***	(−0.54, −0.15)	−0.30**	(−0.50, −0.10)	−0.10	(−0.29, 0.10)	−0.29**	(−0.49, −0.10)	0.04[a]	(−0.22, 0.30)
Y. Adu							−0.10	(−0.34, 0.13)	0.01[a]	(−0.30, 0.31)
M. Adu	−0.43***	(−0.63, −0.22)	−0.50***	(−0.73, −0.26)	−0.29**	(−0.47, −0.11)	−0.45***	−0.64, −0.27)	−0.20[a]	(−0.61, 0.21)
									0.38*[b]	(0.06, 0.69)
L. Adu	−0.19	(−0.41, 0.03)	−0.54***	(−0.82, −0.25)	−0.22*	(−0.44, −0.04)	−0.23*	(−0.44, −0.02)	0.23[b]	(−0.06, 0.52)

Note. *p < .05, **p < .01, ***p < .001. Difference values that are negative indicate that women have lower mean scores relative to the means of men. These difference scores can be interpreted in terms of effect sizes (Cohen's d). Note that mean gender comparisons in the young adulthood group for the Dirty Dozen scales are not presented in the table, because we found no evidence for (partial) scalar invariance within that age group. Gender comparisons on Agreeableness with an [a] superscript are based on BFI data, those with a [b] superscript are based on NEO-FFI-3 data.

levels were equal for men and women only in the late adolescence group. The late adulthood group was the only group in which women showed equivalent manipulation levels to men. The young adulthood group was the only group in which levels on the general Dirty Dozen factor were equal for men and women. For agreeableness, gender differences were only significant in one group and for one measure, as women had higher levels of agreeableness on the NEO-FFI in the middle adulthood group.

2.3. Age differences by gender

Age trends based on raw data are illustrated in Fig. 1. The raw data patterns were empirically smoothed using LOcally Estimated Scatterplot Smoothing (LOESS), which is based on a local regression procedure (e.g., Cleveland, 1979). Fig. 1 suggests that changes were more pronounced in the Dirty Dozen general factor and sub factors than for agreeableness, but that trends for agreeableness mirror those of the Dirty Dozen. However, Fig. 1 is based on raw data, which can be biased due to between-group differences in measurement properties. Therefore, it is more appropriate to interpret the mean-level age group comparisons presented below and in Table 5 (for men) and Table 6 (for women), which are based on latent means, for which such biases are less pronounced (Steinmetz, 2013).

Mean-Level Age Trends for Men. In men, manipulation levels were lowest in early adolescence but significantly higher in the middle- and late adolescent and young adulthood age groups. Across the latter three age groups, levels remained fairly comparable. Compared to the young adulthood age group, lower levels of manipulation were reported in the middle and late adulthood age group (see Fig. 1 and Table 5).

The early adolescence group for men showed relatively low callous affect levels. These levels were significantly higher in the middle and late adolescence groups, and even higher in the young

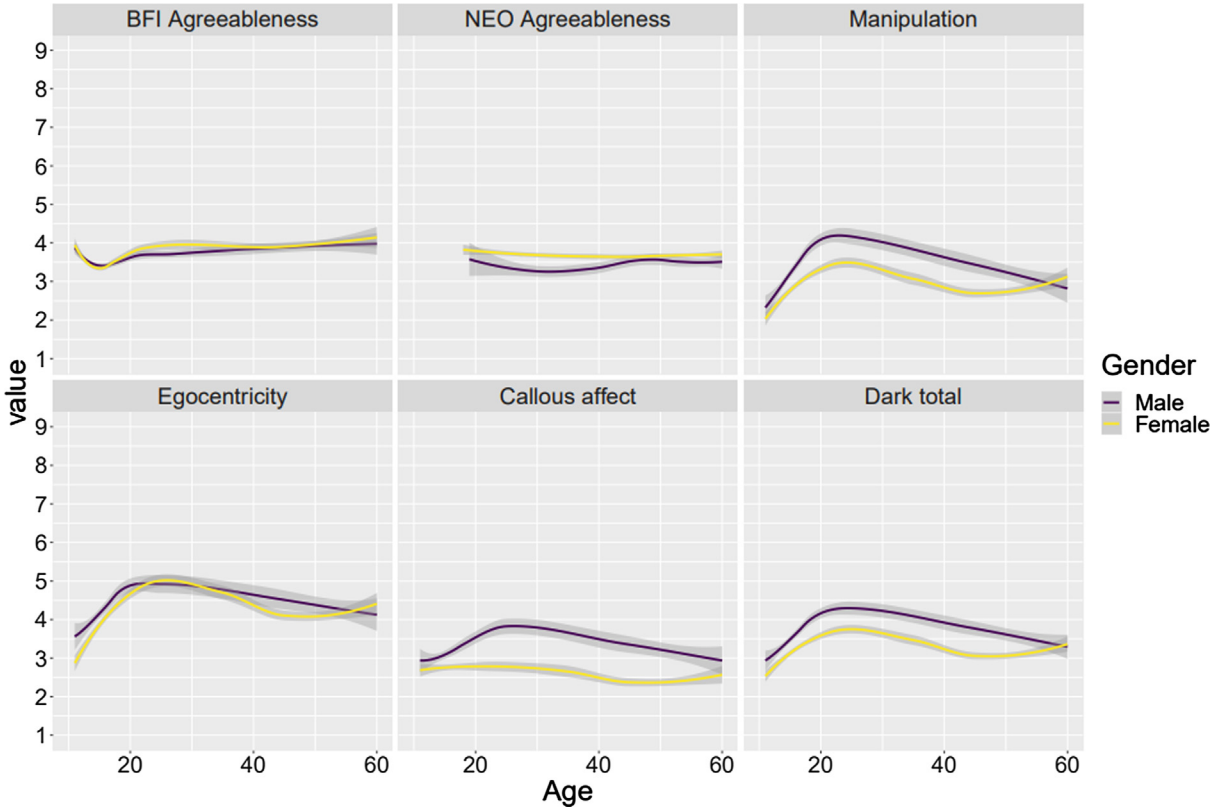

Fig. 1. Continuous age trends based on raw data for manipulation, callous affect, egocentricity, a general Dirty Dozen factor, and two measures of agreeableness in women and men. The gray shading around the lines represents 95% confidence intervals.

Table 5
Estimates and 95% Confidence Intervals of Latent Means for Dirty Dozen Scales in Men from Early Adolescence to Late Adulthood by Three Different Models.

	Model 1			Model 2		Model 3		
	EAdo	MAdo	LAdo	LAdo	YAdu	YAdu	MAdu	LAdu
Manipulation	0.00[a]	0.26[b] 0.07, 0.44	0.62[b] 0.38, 0.85	0.62[b]	0.80[b] 0.61, 1.00	0.80[b]	0.48[c] 0.25, 0.71	0.18[c] −0.06, 0.41
Callous Affect	0.00[a]	0.22[b] 0.04, 0.40	0.39[b] 0.17, 0.62	0.39[b]	0.69[c] 0.48, 0.90	0.69[c]	0.36[d] 0.12, 0.60	0.28[d] 0.03, 0.54
Egocentricity	0.00[a]	0.02[a] −0.16, 0.20	0.43[b] 0.22, 0.64	0.43[b]	0.54[b] 0.35, 0.74	0.54[b]	0.39[b,c] 0.15, 0.62	0.25[c] 0.01, 0.49
General DD	0.00[a]	0.05[a] −0.13, 0.22	0.47[b] 0.27, 0.68	0.47[b]	0.66[b] 0.47, 0.85	0.66[b]	0.24[c] −0.01, 0.48	0.07[c] −0.19, 0.33
Agr. BFI[1]	0.00[a]	−0.29[b] −0.53, −0.06	−0.24[b] −0.48, 0.00	−0.24[b]	−0.12[b] −0.35, 0.12			
Agr. NEO[2]							0.00[a]	0.08[a] −0.22, 0.38

Note. Latent means with different superscript letters are significantly different ($p < .05$). Note that comparing means obtained in different models is not warranted. For example, the latent means of men in the early adolescence group (obtained in Model 1) cannot be directly compared to latent means of men in the young adulthood group (obtained in Model 2). However, the fact that for callous affect the latent mean for the late adolescence group is larger than the latent mean for the early adolescence group and the latent means for the young adulthood group are larger than those of the late adolescence group, logically implies that the latent mean for the young adulthood group is also larger than the latent mean of the early adolescence group. Models compare early adolescents (EAdo), middle adolescents (MAdo), late adolescents (LAdo), young adults (YAdu), middle adults (MAdu) and late adults (LAdu). [1] Mean comparison for the BFI were actually run in three models with the following comparisons: EAdo versus MAdo, MAdo versus LAdo, and LAdo versus YAdu. [2] Mean comparisons on NEO-FFI-3 Agreeableness were based on one model comparing the MAdu and LAdu group.

Table 6
Estimates and 95% Confidence Intervals of Latent Means for Women on the Dirty Dozen Scales from Early Adolescence to Late Adulthood by Five Different Models.

	Model 1		Model 2		Model 3		Model 4		Model 5	
	EAdo	MAdo	MAdo	LAdo	LAdo	YAdu	YAdu	MAdu	MAdu	LAdu
Manipulation	0.00[a]	0.42[b] 0.26, 0.58	0.42[b]	0.68[c] 0.50, 0.87	0.68[c]	0.63[c] 0.44, 0.81	0.63[c]	0.20[d] 0.04, 0.36	0.20[d]	0.25[d] 0.10, 0.40
Callous Affect	0.00[a]	0.35[b] 0.17, 0.52	0.35[b]	0.32[b] 0.15, 0.49	0.32[b]	0.13[c] −0.06, 0.32	0.13[c]	−0.06[d] −0.22, 0.09	−0.06[d]	−0.01[d] −0.16, 0.14
Egocentricity	0.00[a]	0.30[b] 0.14, 0.46	0.30[b]	0.82[c] 0.66, 0.98	0.82[c]	0.89[c] 0.69, 1.08	0.89[c]	0.43[d] 0.28, 0.58	0.43[d]	0.50[d] 0.35, 0.64
General DD	0.00[a]	0.32[b] 0.16, 0.49	0.32[b]	0.50[c] 0.33, 0.68	0.50[c]	0.31[d] 0.13, 0.49	0.309[d]	−0.12[e] −0.29, 0.05	−0.12[e]	−0.04[e] −0.18, 0.10
Agr. BFI	0.00[a]	−0.27[b] −0.49, −0.04	−0.27[b]	−0.27[b] −0.54, 0.00			−0.28[b]	−0.35[b] −0.62, −0.08		
Agr. NEO[1]							0.00[a]	−0.34[b] −0.62, −0.05	−0.34[b]	−0.14[ab] −0.42, 0.15

Note. Latent means with different superscript letters are significantly different. Note that comparing means obtained in different models is not warranted. For example, the latent means of women in the early adolescence group (obtained in Model 1) cannot be directly compared to latent means of women in the late adulthood group (obtained in Model 5). However, the fact that the latent means for egocentricity of the middle adolescence group are larger than the latent means for the early adolescence group, and the latent means for the late adolescence group are larger than those of the middle adolescence group logically implies that the latent mean for egocentricity of the late adolescence group is also larger than the latent mean of the early adolescence group. Models compare early adolescents (EAdo), middle adolescents (MAdo), late adolescents (LAdo), young adults (YAdu), middle adults (MAdu) and late adults (LAdu). The estimate for the young adulthood group in Model 3 was derived from a model in which the late adolescence and young adulthood group were compared. Given lacking invariance between those groups, between-group mean comparisons from that model are not presented. [1] Mean comparisons between the three adult age groups (YAdu, MAdu, and LAdu) on NEO-FFI-3 Agreeableness are based on one three-group model. Therefore, means for the late adulthood group can be directly compared to those for the young adulthood group.

adult group. Compared to the young adult group, lower callous affect levels were observed in middle and late adulthood groups (see Table 5 and Fig. 1).

Egocentricity levels in men were comparable between the early and middle adolescence groups, significantly higher than those in the late adolescence group, and comparably high in the young and middle adulthood groups. Levels in the late adulthood group were lower than those in the younger adult groups (see Table 5 and Fig. 1).

General Dirty Dozen factor levels were comparable between men in the early and middle adolescence group, but significantly higher in the late adolescence group. There were no differences between the late adolescence and young adulthood men, but the middle adulthood group had significantly lower levels than the young adulthood group. Differences between the middle and late adulthood group were not significant (see Table 5 and Fig. 1).

For agreeableness, the early adolescence group had higher levels than the middle and late adolescence group did. These latter two groups did not significantly differ from each other. The young adulthood group also did not differ significantly from the late adolescent group. Finally, the middle and late adulthood groups also did not differ significantly from each other.

Mean-Level Age Trends for Women. Women in the early adolescence group had relatively low manipulation levels, levels for the middle adolescence group were higher, and those for the late adolescence group were even higher than that. The late adolescence and young adulthood groups reported similar levels of manipulation. The middle adulthood group had significantly lower levels of manipulation than the young adult women, but middle and late adults were comparable (see Table 6 and Fig. 1).

Women in the early adolescence group reported lower callous affect levels than those in the middle and late adolescence groups. Young adult women had lower levels than late adolescent women, and middle adult women had lower levels than young adult women. Levels of callous affect were similar between the middle and late adult women (Table 6 and Fig. 1).

The older the women, the higher their egocentricity levels were. The late adolescence and young adulthood groups did not differ significantly from each other. The middle adulthood group had significantly lower levels of egocentricity when compared to the young adulthood group. Middle and late adults did not differ significantly from each other (see Table 6 and Fig. 1).

For the general Dirty Dozen factor level, there were significant mean-level differences between all adolescent groups. The older adolescent women were, the higher their mean levels were. Young adults had lower levels than late adolescents, and the middle adults' levels on the general Dirty Dozen factor were lower than those of the young adults. There were no significant differences between the middle and late adulthood groups (see Table 6 and Fig. 1).

Women in the middle adolescence group had significantly lower levels of agreeableness than the early adolescence group.

There were no significant differences between the middle and late adolescence groups. The pattern of differences between the young and middle adulthood group was mixed: The BFI measure yielded no significant differences between these groups, whereas the NEO-FFI-3 findings suggest that the middle adulthood group had lower levels of agreeableness than the young adulthood group. The late adulthood group did not differ significantly from the middle adulthood group on agreeableness (see Table 6 and Fig. 1).

2.4. Conclusion

Results obtained in Study 1 suggest that gender differences in the Dirty Dozen sub-factors and general factor vary over the lifespan. On average, albeit not always significantly, men scored higher on the "dark" features than women. However, especially gender differences in egocentricity appeared to be smaller than what has typically been suggested in the literature (see also Grijalva et al., 2015).

Interestingly, this pattern of findings did not simply represent a mirror image of the findings for agreeableness, as gender differences for agreeableness were typically not significant. There was one exception in middle adulthood, but even that finding did not replicate across different Big Five measures. This suggests that low agreeableness may be associated with the Dirty Dozen scales as various studies have shown (e.g., Muris et al., 2017; Vize et al., 2018), but that the particular operationalization of antagonism constructs in the Dirty Dozen is more sensitive to detecting gender differences. Hence, the Dirty Dozen measure may provide at least some unique information relative to agreeableness.

Our findings generally suggest that mean-levels of dark features and a general dark factor increased with age over adolescence, stabilized in young adulthood, and then decreased with age in the later stages of adulthood. These findings will be discussed in more detail in the General Discussion below, but highlight that adolescence may be a key period for understanding the development of dark features.

This pattern of findings was more or less mirrored in our findings for agreeableness, for which we found increasingly lower levels in older adolescent age groups. The pattern for adults was harder to disentangle in our data, as we a had limited amount of agreeableness data for adult age groups, but the available data suggest that there were few mean-level differences between these groups. Previous studies (e.g., Soto et al., 2011) do suggest small increases in levels of agreeableness. Hence, the Dirty Dozen may demonstrate the same patterns of age-related increases and decreases as was already known based on research on agreeableness, although the exact shape of these patterns was more pronounced in the Dirty Dozen.

However, Study 1 had some limitations. For example, the early adolescent girls group a three-factor model did not outperform a two-factor model with a combined manipulation-callous affect factor. Given that it was only in this specific group that the three-factor model was not better than a two-factor model, we proceeded using three factors to facilitate between-group comparisons. However, this three-factor structure for early adolescent girls was suboptimal. In addition, we found no invariance between gender groups in the young adulthood age group, due to which we were unable to examine mean-level gender differences in that age group. Our study was focused was on age and gender differences in mean levels, which is why we did not further investigate the causes of the lack of invariance. However, our data are available to researchers wishing to pursue more specific research questions pertaining to measurement invariance (https://osf.io/pfd8b/?view_only=2fd77cf5d2c349859ccbaae448e744bf).

Another potential limitation was that some of our preliminary tests yielded *p*-values close to 0.05, suggesting confounds of assessment method and sample differences with gender and age differences. However, problems with over-interpreting effects for which *p*-values are close to 0.05 (e.g., Wilkinson & the Task Force on Statistical Inference, 1999) apply as much to preliminary tests as they do to tests related to substantial research questions. That being said, there was a small and significant effect suggesting a confounding effect of sample differences with age and gender differences in ego-centricity. Thus, our findings regarding ego-centricity should be interpreted more cautiously than our other findings.

In addition, there may have been (a lack of) measurement invariance between particular samples independent of age and gender differences. One way to address this concern would be to assess measurement invariance between samples within particular age groups within particular gender groups (e.g., between sample differences in early adolescent men). However, group sizes were only large enough ($n > 150$) for conducting such tests between two female samples in the middle adulthood group. Therefore, this would not have yielded a representative picture of this source of bias. Hence, we did not conduct such tests, but readers should be aware of between-dataset differences in measurement properties potentially biasing our results.

The major limitation of Study 1 was its cross-sectional nature. Birth cohort effects can confound cross-sectional age trends. The youngest participants were born in the 2000s and the oldest participants were born in the 1950s or earlier. These groups thus experienced a very different childhood (e.g., general access to mobile technology versus no general access to television) and were raised in times that were characterized by different sociocultural norms. For example, the percentage of non-religious individuals in the Netherlands increased from 18% in 1960 to 50% in 2015 (Statistics Netherlands, 2016). Cross-sectional age trends may thus be caused by differences in the way birth cohorts have been socialized rather than by lifespan developmental effects.

Note that the inability of disentangling birth-cohort effects from age effects does not mean that the age differences observed in Study 1 are not relevant. These results reflect age differences in dark features as they can currently be observed in the Dutch population. However, it is important to realize that these age trends do not necessarily reflect developmental trends, as development is an individual-level process that cannot be inferred from group mean comparisons (cf. Nesselroade & Baltes, 1974). For the study of development, longitudinal data is a necessity.

The lack of longitudinal data in Study 1 also resulted in the inability to examine correlated change of the Dirty Dozen and its scales with agreeableness to test the distinctiveness of these constructs. In addition, the use of rather broad age categories likely caused temporary fluctuations (e.g., decreases followed by increases) to be overlooked. Designs with observation points spaced less far apart (e.g., every year) are better suited for detecting change in such periods. Therefore, we wanted to extend the results of Study 1 in a second, longitudinal study.

3. Study 2: Longitudinal age trends in adolescence

To longitudinally replicate the results obtained in Study 1, a set of long-term longitudinal studies covering several adjacent age periods would have been necessary. Unfortunately, we did not have such data available at the time of writing. However, any type of longitudinal data could be useful for addressing some of the limitations of Study 1, including evidence of intra-individual change. Information on individual-level change is an absolute necessity for the study of development (e.g., Nesselroade & Baltes, 1974). Longitudinal data covering an age span of even a couple of years would indicate how much individual differences in patterns of

change there are on the Dirty Dozen and its scales. If there are large individual differences, it can be examined which factors correlate with these individual differences in change trajectories.

Longitudinal data can also provide additional information of the uniqueness of the Dirty Dozen and its scales relative to broad personality dimensions, such as agreeableness. Specifically, such data can be used to assess whether changes in agreeableness (or other broad personality dimensions) are associated with changes in the Dirty Dozen scales. In other words, longitudinal data can be used to estimate correlated change. Correlated change estimates have previously been interpreted as evidence for two dimensions sharing a common cause, or being part of the same broader construct (e.g., Allemand & Martin, 2016; Klimstra, Bleidorn, Asendorpf, van Aken, & Denissen, 2013). Thus, the absence or presence of correlated change between agreeableness and the Dirty Dozen and its scales may provide some insight into the Dirty Dozen's incremental value over agreeableness in a much more direct manner than is possible with comparisons of cross-sectional age trends.

We used a theory-inspired, but rather crude age-group categorization to estimate age trends in Study 1. This was a necessity, as alternative procedures would have resulted in a small number of observations per group and therefore unreliable estimates, but it was a limitation nonetheless. That is, development is not a linear process as our statistical models often lead us to believe, as change can be multidirectional with alternating increases, stability, and decreases. An example of this has been provided by previous research on big five features, for which change in early and middle adolescence was not always in the direction of greater maturity, but changes in late adolescence were (Denissen et al., 2013). To capture such dynamic patterns, data points should be less far apart than the ones in Study 1, in which age groups covered an age span of up to three years in adolescence.

For Study 2, we had longitudinal data covering early to mid-adolescence (ages 13 to 15) with three annual measurement occasions. This period is particularly interesting for the study of personality development, as previous studies showed that mean levels of big five features do not always change in the direction of greater maturity (Denissen et al., 2013). Results from Study 1 are in line with these changes away from maturity, as they generally suggested strong increases in dark features and decreases in agreeableness within this period. We expected to replicate this pattern longitudinally in Study 2, but acknowledge that the replication value of Study 2 is limited to adolescence. Instead, Study 2 should thus primarily be regarded as a useful extension that can provide additional information on how dark personality features change with age.

An extra incremental feature of Study 2 compared to Study 1 is that the longitudinal design allows for examining correlated change of the Dirty Dozen and its scales with agreeableness. Levels of agreeableness tend to be negatively associated with levels of the Dirty Dozen overall score and scale scores at the between-person level. In Study 1, the age-related patterns of change in agreeableness at least partly seemed to mirror those in the Dirty Dozen. Therefore, we expected changes in agreeableness to be negatively associated with changes in the Dirty Dozen overall score and scale scores.

3.1. Method

3.1.1. Participants

Respondents participated in the three-annual-wave longitudinal Study of Personality, Adjustment, Cognitions, and Emotions-II (SPACE-II), of which data of the first measurement occasion was also included in Study 1. We only included the 325 cases (M_{age} 13.31 years, SD = 1.03; 48.6% girls) who had data on at least two out of the three measurement occasions. These data are openly available: https://osf.io/ukfz2/?view_only=0e794f684c104fda897fc2afeef5ec82.

3.1.2. Measures

We used the same Dutch-language version (Klimstra et al., 2014) of the Dirty Dozen (Jonason & Webster, 2010) as in Study 1. Coefficient alphas across scales and across waves ranged from 0.70 to 0.88. To examine the extent to which change in the Dirty Dozen scales mirrors changes in related personality constructs, we included data on the Dutch-language version of the BFI-25 (Boele et al., 2017; Denissen et al., 2008; John et al., 1991; see Study 1) in our analyses. Coefficient alphas were 0.60, 0.63, and 0.52 at Waves 1, 2, and 3, respectively. Notably, despite that the alpha was low, there were no significant negative inter-item correlations.

3.2. Strategy of analyses

For the Dirty Dozen, longitudinal measurement invariance tests showed that three-factor models provided a good fit to the data on all three measurement occasions (configural invariance). This evidence for configural invariance was further supported by the fact that three-factor models had a better fit to the data than plausible alternative models (i.e., one- and two-factor models; see also Study 1). For the general Dirty Dozen factor models, we again relied on models with a general latent factor identified by all 12 items. There was evidence for full invariance between the two gender groups at all three time points, and partial metric and partial scalar longitudinal invariance for this general factor in boys, and partial metric and partial scalar longitudinal invariance in girls (see supplementary material section 6). For the agreeableness assessments, we found evidence for full metric invariance and scalar invariance across gender groups, and full metric and partial scalar longitudinal invariance for both boys and girls (see supplementary material Section 6). In these models and subsequent models, we used Maximum Likelihood Robust estimation in Mplus 7 (Muthen & Muthen, 2012), and the same fit criteria as in Study 1.

We examined potential mean-level change by running univariate latent growth models for each Dirty Dozen factor, the general Dirty Dozen factor, and agreeableness. We used items as indicators of latent means at all three measurement occasions. These latent means were subsequently used as indicators of latent growth factors (i.e., a latent intercept and a latent slope). Models based on latent means tend to result in estimates that do not optimally match the raw data's metric. To deal with this problem and to facilitate the interpretability of the estimates derived from the LGMs, we used effects coding as much as possible (Little, Slegers, & Card, 2006). The slope factor loadings were 0, 1, and 2, reflecting the fact that there was a one-year interval between each of the measurement occasions. Thus, the slope indicates the estimated amount of change per year.

To assess empirical overlap of the Dirty Dozen scales and general factor scale with agreeableness, we estimated correlated change using the simplest model possible given our limited sample size. Hence, multivariate growth models were ruled out because of their complexity and we proceeded to estimate bivariate cross-lagged panel models. Given that the focus in such models is on associations rather than means, scalar invariance tests can be omitted (Steenkamp & Baumgarther, 1998). As we found full metric invariance, we were able to use (observed) scale means rather than latent scores with items as indicators to further reduce model complexity (cf. Steinmetz, 2013). Hence, we used a series of four cross-lagged panel models in which agreeableness was linked to the Dirty Dozen total score, manipulation, callous affect, and egocentricity, respectively. These models are informative on associations between relative changes in variables (i.e., whether moving up in

the rank order on one variable is associated with moving up in the rank order on another variable; e.g., Klimstra, Noftle et al., 2018).

3.3. Results

To examine mean-level change, we first compared linear growth models (slope factor for T1 on 0, for T2 on 1, and for T3 on 2) to non-linear growth models (slope factor loading for T1 on 0 and for T2 on 1 again, but a freely estimated factor loading for T3). Both types of models fitted the data equally well ($\Delta\chi^2_{(\Delta df = 1)} \leq 3.20$, $p \geq 0.06$), indicating that an unspecified linear growth function was not distinguishable from a slightly non-linear one. In such cases, the most parsimonious model (i.e., the linear model) should be selected (Kline, 2015). The final models had an acceptable fit to the data, with CFIs $\geq$ 0.919 and RMSEAs $\leq$ 0.059. One exception to this general rule was the one-factor Dirty Dozen model, with CFI = 0.694 and RMSEA = 0.078, which was due to a three-factor model providing a better fit to the data than a one-factor model did (see Supplementary Section 7 for details).

Growth parameters obtained in the latent growth models for the entire sample are displayed in the left part of Table 7. Table 7 shows that levels of manipulation and callous affect and the mean of the generally dirty dozen factor increased significantly over the two years of this study, whereas levels of agreeableness decreased. Levels of egocentricity did not change significantly, in contrast to Study 1.

In follow-up multigroup models, we tested differences between girls and boys in levels and changes for the manipulation, callous affect, egocentricity, and agreeableness scales using a scaled Chi-square difference test (Satorra & Bentler, 2001) which is necessary when using the MLR estimator that we used. Specifically, we compared models in which levels and change rates were freely estimated for boys and girls to models in which these parameters were constrained to be equal. These tests, presented in Supplementary Section 7, showed no significant gender differences in levels or change of manipulation, callous affect, and agreeableness. Both boys and girls increased significantly in manipulation. For callous affect, boys increased significantly, whereas girls showed a non-significant increase that was nevertheless not significantly smaller. Specifically, there was no significant difference between the chi-square of a model in which change for callous affect was constrained to be equal across gender groups and a model in which change was allowed to vary between gender groups.

For the general Dirty Dozen factor, a multigroup model was not feasible because of the unfavorable ratio of the number of parameters estimated in such a model relative to the number of participants in each of the groups. Therefore, we estimated models for boys and girls, separately. Growth likely was not linear for the general Dirty Dozen factor in girls, as a non-linear model had a marginally better fit than the linear model ($\Delta\chi2_{(\Delta df = 1)} \leq 3.99$, p = .046). Boys showed a significant linear increase in mean levels on the Dirty Dozen factor, whereas girls showed an increased followed by no change.

For egocentricity, adding a quadratic slope factor improved model fit ($\Delta\chi^2_{(\Delta df = 2)} = 8.14$, $p = .03$). Follow-up analyses showed significant gender differences in the growth pattern ($\Delta\chi^2_{(\Delta df = 2)} = 5.28$, $p = .02$). Girls showed an increase followed by a decrease, whereas boys showed no significant changes.

Final growth parameters for all models are shown in Table 7 (for details on model fit, see Supplementary Material Section 7). Growth curves based on these parameters are shown in Fig. 2.

We used cross-lagged panel models to examine the empirical overlap of agreeableness with the dirty dozen scales and overall score. The fit of these models was acceptable and did not deteriorate when associations were constrained to be equal across gender and time (see Supplementary Material Section 8). Across models, relative change in agreeableness was negatively and significantly associated with relative changes in the dirty dozen subscales and total score. These associations were mostly of moderate effect size, except for the correlated change between agreeableness and egocentricity, which was of small effect size (see Table 8).

3.4. Conclusion

Using longitudinal data, we replicated some of the key findings that we obtained using cross-sectional data in Study 1, whereas other findings were not replicated. A key finding that we did not replicate longitudinally concerned the gender differences. However, we focused on the age-period in which gender differences were relatively small in Study 1, and the smaller sample in Study 2 resulted in less power. Interestingly, even with our relatively small sample size, we were able to show that girls displayed a temporary peak in egocentricity in mid-adolescence, whereas boys showed no significant changes.

The cross-sectional mean-level age-related increases in dark features that we found in Study 1 were replicated for callous affect, manipulation, the general Dirty Dozen factor, and agreeableness in Study 2. However, the relatively strong and linear cross-sectional age trends for egocentricity for both boys and girls were only partially replicated longitudinally. The longitudinal increase we found for boys was not significant, and in girls an initial increase in egocentricity was followed by a decrease. Our longitudinal findings may have been more similar to the cross-sectional findings if they would have been based on more measurement occasions, as latent

Table 7
Growth parameters for the whole adolescent sample and for adolescent boys and girls, separately.

	Overall				Boys					Girls				
	Intercept		Linear Slope		Intercept		Linear Slope		Quadr. Slope	Intercept		Linear Slope		Quadr. Slope
	M	σ^2	*M*	σ^2	*M*	σ^2	*M*	σ^2	*M*	*M*	σ^2	*M*	σ^2	*M*
Man.	2.49***	1.69***	0.34***	0.44*	2.63***	1.82***	0.31***	0.49**	n.a.	2.32***	1.46**	0.34***	0.41*	n.a.
Call.	2.84***	1.29***	0.35***	0.30	2.84***	1.48***	0.49***	0.35	n.a.	2.81***	1.08*	0.21	0.28	n.a.
EgoC	3.58***	1.79***	0.06	0.19	3.64***	2.52***	0.10	0.62	0.01	3.35***	1.64**	0.83**	0.26	−0.43**
DD	2.92***	1.84	0.38***	0.82	3.02***	1.45***	0.33***	0.35*	n.a.	2.79***	1.11*	0.64**	0.18	−0.22
Agree	3.51***	0.19***	−0.04*	0.01	3.54***	0.18***	-0.06**	0.00[a]	n.a.	3.48***	0.21***	−0.02	0.00[a]	n.a.

Note. *p < .05, **p < .01, ***p < .001. M = mean estimate, σ^2 = variance around the mean. Man. = Manipulation, Call. = Callous Affect, EgoC = Egocentricity, DD = general Dirty Dozen latent factor, Agree = Agreeableness. Quadratic slopes were only estimated for narcissism, as this was the only characteristic for which non-linear growth model provided a significantly better fit than a linear growth model (hence the n.a. for the quadratic slope factors for manipulation and callous affect). The variances of these quadratics slopes were fixed to zero to get the model identified. [a] The slope variance for agreeableness was constrained to zero, because the initial model produced negative estimates for both boys and girls. [b] Due to the complex pattern of invariance constraints as well as the large number of estimated parameters relative to sample size in the models for the general Dirty Dozen latent factor, we did not use a multigroup approach but estimated separate models for girls and boys. Further analyses on the moderating role of gender by including gender as a predictor of the intercept and slopes showed that there was no evidence for a significant effect of gender on the intercept and slope of the general Dirty Dozen factor (see Supplementary Material, Table 7.8)

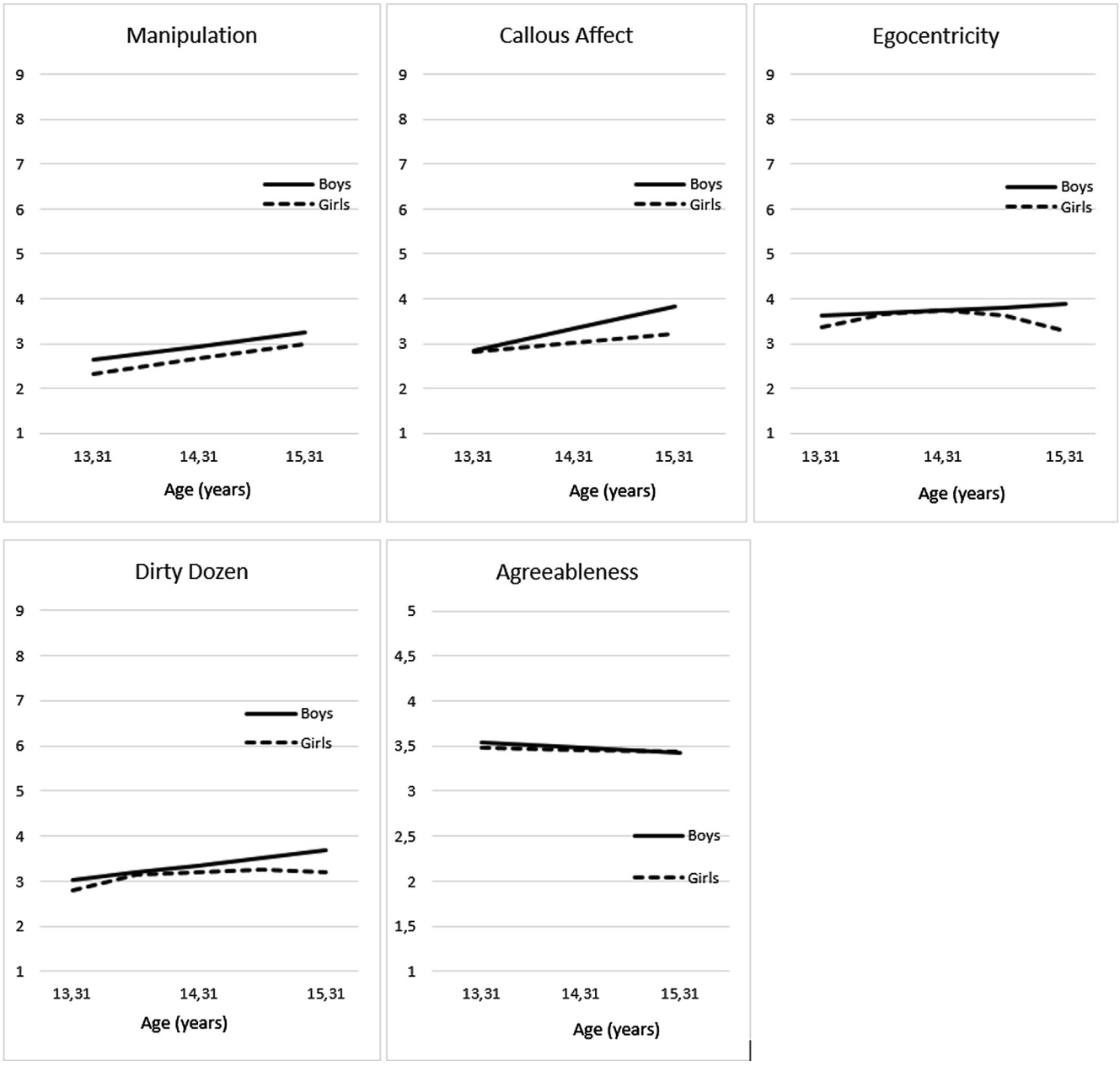

Fig. 2. Longitudinal change patterns based on estimated means from univariate latent growth models for manipulation, callous affect, and egocentricity, the general Dirty Dozen factor, and agreeableness in adolescent boys and girls.

Table 8
Time 1 correlations and correlated relative change of agreeableness with the overall dirty dozen score and subscale scores as estimated in cross-lagged panel models.

	Boys			Girls		
Agreeableness with. . .	T1	T2	T3	T1	T2	T3
Dirty Dozen total score	−0.293***	−0.359***	−0.423***	−0.431***	−0.331***	−0.422***
Manipulation	−0.314***	−0.330***	−0.365***	−0.386***	−0.292***	−0.418***
Callous Affect	−0.387***	−0.400***	−0.451***	−0.470***	−0.393***	−0.437***
Egocentricity	−0.034	−0.153**	−0.152**	−0.237**	−0.120**	−0.139**

Note. $^{**}p < .01$, $^{***}p < .001$. The T1 association is a zero-order correlation. The T2 and T3 associations are correlations between residuals, which, in a cross-lagged panel model, reflect correlated relative change. Specifically, these coefficients reflect whether changes in the rank order for one variable were associated with changes in the rank order on another variable. The unstandardized coefficients for the T2 and T3 associations were constrained to be equal across time and across gender groups, as adding such constraints did not significantly affect model fit (see Supplementary Material, Section 8). However, constraints on unstandardized estimates can still result in slight differences in the standardized coefficients, as was the case in the present study.

growth estimates increase in reliability with more measurement occasions (Willett, Singer, & Martin, 1998). In addition, the sample size of the longitudinal study was relatively small (N = 325), with larger samples typically producing more precise results.

Another way in which Study 2's findings replicated those of Study 1, is that the mean-level findings for agreeableness did not necessarily perfectly mirror the findings for the Dirty Dozen. Whereas this was most clear for gender differences in Study 1, it was more apparent for mean-level change in Study 2. Specifically, evidence for increases was stronger for the Dirty Dozen scales and general factor compared to increases in agreeableness. This suggests that the Dirty Dozen maybe more sensitive to detecting change.

However, examining similarity between mean-level patterns for agreeableness and the Dirty Dozen only provides indirect evidence for their interrelations. The longitudinal design of Study 2 allowed us to examine this overlap much more directly, using cross-lagged panel models. These models, suitable for capturing correlated relative change, showed that there was substantial empirical overlap between agreeableness and the Dirty Dozen. Previous studies also showed strong associations of the Dirty Dozen with dimensions such as agreeableness (e.g., Muris et al., 2017; Vize et al., 2018), but our findings go beyond those by showing that these associations can be generalized to relative changes in both constructs. In research on personality and psychopathology linkages (e.g., De Bolle, Beyers, De Clercq, & De Fruyt, 2012), correlated relative change has been interpreted as evidence for a spectrum/continuity model. In other words, correlated relative change may indicate that the constructs involved are on the same continuum. It was clear from our results that this overlap was much smaller for egocentricity than for the other Dirty Dozen scales and its general factor. For callous affect, manipulation, and the general Dirty Dozen factor, estimates of both within-time correlations and correlated relative change coefficients were in the medium-to-large effect size range. This suggests at least part of the general Dirty Dozen factor, manipulation, and callous affect are particular manifestations of antagonism, which is typically defined as low agreeableness.

In sum, Study 2 largely confirms that adolescence may be an important period in which levels of dark features are on the rise, but that these developments are not independent from changes in the broader construct of antagonism. The overall findings and their implications will be outlined further in the General Discussion.

4. General discussion

Lifespan development of dark features and gender differences herein are understudied. The present study addresses this gap in the literature by examining gender and age trends from 11 to 77 years using integrative data analysis on 12 datasets that included the Dutch version of the Dirty Dozen measure. We supplemented cross-sectional analyses with longitudinal analyses covering early to mid-adolescence and compared the trends for the Dirty Dozen and its scales to those for agreeableness. Both studies suggest gender and age differences in levels on all three Dirty Dozen scales and a general Dirty Dozen factor which were partly, but not entirely, mirrored in and associated with gender and age differences in agreeableness. Specifically, men typically displayed higher levels on all dark features than women, but the magnitude of this gender gap varied with age and by characteristic. There were strong age-related increases in all dark features through adolescence, whereas levels decreased again after middle adulthood. Our interpretation of these general patterns is discussed in detail below.

4.1. The Dirty Dozen scales and gender

First, our preliminary analyses showed that measurement invariance across gender groups was rare. Therefore, raw scale gender comparisons may be biased (van de Schoot et al., 2012). Using raw scores is common practice in research using the Dirty Dozen (e.g., Czarna et al., 2016; Jonason & Webster, 2010; our Fig. 1) and in psychology more generally, which could be problematic. However, the 95% confidence intervals of our latent-score-based estimates for gender differences overlapped with the 95% confidence intervals that were obtained based on raw scores of the Dirty Dozen as presented in a meta-analysis (Muris et al., 2017), which could be interpreted as evidence pointing to a lack of necessity of invariance tests. That would be a misinterpretation, as our own data suggest that using raw scores instead of latent scores can lead to both over- and underestimation of gender differences, depending on the characteristic and the age period examined (see Supplementary Material, Table 9.1). Although the latent mean comparisons we presented may still be biased in case partial invariance was obtained (i.e., in the early adolescence, middle adulthood, and late adulthood age group), they should at least be less biased than observed mean comparisons (Steinmetz, 2013).

Second, we found a lack of gender invariance in the young adulthood group, which impeded formal conclusions regarding mean-level gender differences among Dutch people in their twenties. Most studies on gender differences using the Dirty Dozen focused on young adults (Czarna et al., 2016; Jonason & Webster, 2010; Webster & Jonason, 2013), and without evidence for invariance, their estimates of gender differences may be biased. These findings could be interpreted as evidence for an unusual weak performance of the Dirty Dozen measure compared to other measures. Although we did not have similar problems with our agreeableness measures, such an interpretation would be unfair, as studies reporting measurement invariance tests before conducting group comparisons are unfortunately rare. Therefore, we would rather argue that these results illustrate that invariance tests should be conducted more often on any measure and that problems regarding invariance are unlikely to be limited to the Dirty Dozen.

Third, our main results on mean-level gender differences aligned largely with conclusions drawn in previous meta-analyses and systematic reviews on the dark triad (Furnham et al., 2013) or its specific components (Grijalva et al., 2015), despite the narrow focus of the Dirty Dozen. That is, we generally found that men scored higher than women on all three scales. However, despite that confidence intervals of gender gap estimates overlapped across age groups, there seemed to be an age-related trend in this gap. Specifically, we found a medium-to-large sized gender gap in early adolescents, which reduced to a small-to-medium gap in middle adolescents, to become slightly larger again in older groups. These cross-sectional results, combined with longitudinal analyses in Study 2, may be explained by the average adolescent girls generally being ahead of the average adolescent boy in personality development (e.g., Klimstra, Hale, Raaijmakers, Branje, & Meeus, 2009).

Interestingly, the pattern of gender differences for the Dirty Dozen was different from the pattern we found for agreeableness. As we discussed in the conclusion of Study 1, there was little evidence for gender differences in agreeableness, whereas gender differences in the Dirty Dozen were clearer. These gender differences seemed characteristic-specific, especially beyond middle adolescence. Overall, gender differences for callous affect were larger than those for manipulation, and gender differences for egocentricity were the smallest, largely in line with previous research (e.g., Muris et al., 2017; Schmitt et al., 2017).

Gender differences in manipulation were most similar to those typically observed in big five and HEXACO features, as these

became larger from adolescence onwards, and then smaller again in late adulthood (Ashton & Lee, 2016; Soto et al., 2011). Honesty-humility, and especially its facets of sincerity and fairness, are strongly associated with manipulation (Jonason & McCain, 2012). The age-related changes in the gender differences for honesty-humility (Ashton & Lee, 2016) were very similar to the ones we observed for manipulation. Ashton and Lee attributed age trends in the gender gap for honesty-humility to the intensity of intrasexual competition, which peaks at the beginning of adulthood, and may be stronger for men than for women (e.g., Karmin et al., 2015). Thus, the magnitude of gender differences in manipulation might decrease in late adulthood because of a drop in the level of perceived competition. As this drop may be more pronounced for men than for women, gender differences in manipulation may decrease as a result of this.

For callous affect, gender differences were larger in the older age groups and were of a large effect size for late adults. This pattern is not in line with developmental patterns of big five or HEXACO features, but may be explained by the corresponsive principle of personality development. According to this principle, people select environments that strengthen the personality features that led them to choose those environments in the first place (Roberts & Nickel, 2017). Accordingly, gender differences in strongly gendered features such as callous affect could lead to the selection of corresponding environments, which may intensify these gender differences with age. Relatedly, this pattern may be specific to callous affect because its associated behaviors such as physical aggression, may be perceived as typical masculine behavior (Sell, Tooby, & Cosmides, 2009). For example, physical aggression does not necessarily negatively affect social status in men, whereas it typically does in women (Crick, 1997). Behaviors that are more closely related to manipulation and egocentricity, such as gossiping and other forms of relational aggression, can enhance social status for both men and women and therefore might not have gender-specific consequences (Cillessen & Mayeux, 2004). For these reasons, gender differences may only intensify for callous affect.

Similar to Grijalva et al. (2015) results on narcissism, we found that gender differences in its related facet of egocentricity peaked at medium effect sizes. This gender gap was non-significant in late adolescence and remained small to medium in size across adulthood. These gender differences may have been particularly small due to the Dirty Dozen's egocentricity scale capturing vulnerable aspects of narcissism more so than other narcissism measures do (Maples et al., 2014). This mix of stereotypically masculine and feminine elements might result in attenuated overall narcissism differences. Speaking against this possibility, however, is that gender differences in vulnerable narcissism tend to be non-significant (Grijalva et al., 2015).

In their meta-analysis, Grijalva et al. (2015) found no evidence for moderation of gender differences by age, but they relied on the average age of their participant cohorts and also used a broader conceptualization of narcissism. We were able to extend their findings by using the individual participant age rather than average sample age resulting in more statistical power to detect continuous age differences in the gender gap, and a more specific focus on the narcissism-related facet of egocentricity.

4.2. The Dirty Dozen and age

The preliminary measurement invariance tests we conducted were already potentially informative for understanding developmental processes. First, for early adolescents the three-factor structure of the Dirty Dozen was not clearly superior to a two-factor model combining callous affect and manipulation. Similarly, the one-factor model fitted the data much better in younger age groups than in older age groups. These findings could inform a discussion on whether Machiavellianism-related and psychopathy-related features are distinguishable (Miller et al., 2017), as we show that these features' distinctiveness may be age-dependent. It should be noted that the distinction between psychopathy- and Machiavellianism-related facets is particularly blurred in the Dirty Dozen (Maples et al., 2014).

Second, we found measurement invariance between adjacent age groups (e.g., the early adolescence and middle adolescence group), but often not between non-adjacent age groups (e.g., the early adolescence and the young adulthood group). This may suggest subtle age-related shifts in the meaning of Dirty Dozen scales. Previous research showed that big five-like personality features can get more psychological depth or a more interpersonal meaning with age (Soto & John, 2014), which may be the reason why we also only found invariance between more adjacent age groups for our agreeableness measure. Similar shifts in meaning may apply to dark personality features. The measure we used has rather limited bandwidth (Miller et al., 2012), but our findings suggest that it may be worth to further explore age-effects on the distinctiveness and the meaning of dark personality features.

Our main analyses on mean-level age trends suggest that adolescence is a key period for the unfolding of dark features. However, please note that some of the latent mean comparisons we presented may still be biased, as they represent comparisons between group pairs for which only partial invariance was obtained. Still, these estimates are less biased than observed mean comparisons would have been (Steinmetz, 2013). Note that we only compared adjacent age groups causing this problem to be limited to the comparison between the middle and late adolescence groups in women as far as the three-factor Dirty Dozen model is concerned (see Supplementary Material, Section 5). For agreeableness, we found partial invariance for all group comparisons, causing these potential problems to be widely spread.

We found clearly higher mean levels in older adolescents when compared to younger adolescents for all the dark features that we included. With the exception of the changes in egocentricity for adolescent girls, these findings fully replicated in our longitudinal data. We did not anticipate these findings entirely, as the maturity principle for big five personality features would suggest age-related decreases in dark features from middle adolescence onwards (Roberts & Nickel, 2017). On the other hand, our findings for agreeableness were in line with our Dirty Dozen finding and therefore also not in line with the maturity principle, as there were no significant differences between the relatively low levels of agreeableness for the middle adolescence group and those for the late adolescence group. Our findings do align with previously reported age-related decreases in honesty-humility (Ashton & Lee, 2016). Those findings combined with ours suggest that published research on the five-factor model may not represent the full picture for personality features underlying morality.

After adolescence, age trends were less pronounced. If our cross-sectional age trends are indicative of longitudinal change in adulthood, then increases generally leveled off from late adolescence into young adulthood. Towards middle adulthood, previous increases were reversed, and decreasing levels plateaued over late adulthood. These patterns in adulthood are very much in line with our expectations based on the maturity principle (Roberts & Nickel, 2017), but only partly mirror our findings for agreeableness for which we found stability after late adolescence and even some evidence for age-related decreases in women. Decreases in the Dirty Dozen scales do also align with adult developmental tasks, such as contributing to the success and well-being of future generations (i.e., generativity; Erikson, 1950). Much like for our findings regarding gender differences, the Dirty Dozen scales appear to be more sensitive for detecting age effects than the agreeableness measures were.

Overall, levels of dark features appeared to peak in, or around, young adulthood. This is consistent with longitudinal evidence from previous studies. For example, it has been found that narcissism levels were relatively stable and Machiavellianism levels even slightly decreased in a sample of German young adults (Grosz et al., 2017). There are several potential reasons for this peak. First, broad theories, such as Erikson (1950) work on lifespan development and McAdam's theory of personality development (McAdams & Olson, 2010), postulate that adolescents and young adults increasingly focus on finding out who they are, what their morals and ethics are, and what their future plans would be. In other words, they are involved in identity formation. During this process, it is not unusual for youth to break some normative moral conventions (Erikson, 1950; Schwartz et al., 2011), as indicated by the age-crime curve (e.g., Moffitt, 1993; Sweeten, Piquero, & Steinberg, 2013). Additionally, an increased self-focus might be associated with less concerns for others, and thus higher levels of dark features. This hypothesis could be directly tested in future research.

As previously mentioned, young adulthood is also characterized by a peak in intrasexual competition. This peak generalizes to competition for jobs and not just within the sexes. Unemployment is more prevalent in young adulthood than during other periods in the lifespan (e.g., Statistics Netherlands, 2017) and therefore the social environment may be temporarily more permissive of the self-centered behavior associated with dark personality features and antagonism. Similarly, direct aggression has been linked to status goals in young adulthood, but not in later adulthood (Sijtsema, Lindenberg, Ojanen & Salmivalli, 2016). Our results, and these previous results on aggression, suggest that in young adulthood in Western societies goals can be attained through strategies that are less acceptable in other periods of the lifespan. This may also explain why age-crime-curves in delinquent and criminal behavior (e.g., Moffitt, 1993; Sweeten et al., 2013) and intimate partner violence (Johnson, Giordano, Manning, & Longmore, 2015) show overlap with the trajectories we observed for the Dirty Dozen scales, with peaks in late adolescence and young adulthood. Their potential link may be an avenue for future research.

4.3. Does the Dirty Dozen reflect low agreeableness?

As mentioned throughout, findings for the general Dirty Dozen factor and the Dirty Dozen scales did not always mirror those for agreeableness. On the one hand, the Dirty Dozen appeared to be more sensitive to detecting gender differences and age differences in the cross-sectional study. On the other hand, our longitudinal findings do suggest that changes in two out of three of the Dirty Dozen scales and the general Dirty Dozen factor are strongly associated with changes in agreeableness. Therefore, it appears as if the Dirty Dozen's egocentricity scale may tap into something rather unique and different from antagonism, but it is still likely that age and gender trends for this scale have substantial overlap with honesty-humility. This would be in line with research showing that facets of greed avoidance and modesty have their strongest correlations with the Dirty Dozen's egocentricity scale (Jonason & McCain, 2012). The other content of the Dirty Dozen does seem to be related to agreeableness.

The vast number of publications on the Dirty Dozen has highlighted the scales' potential to predict relevant outcomes in ways that are largely in line with what one would expect from a Dark Triad measure. However, as explained in the introduction, the Dirty Dozen cannot be considered a Dark Triad measure due to its limited bandwidth. Moreover, the distinctiveness of the scales, and that of dark personality features in general, has been questioned (e.g., Maples et al., 2014; Miller et al., 2017; Moshagen, Zettler, & Hilbig, 2018). Importantly, it has been suggested that such constructs represented the low end of more general personality dimensions such as agreeableness (Lynam & Miller, 2019). Our findings suggest that the scales are indeed related to, but not redundant with, low agreeableness. Furthermore, the Dirty Dozen appears to be more sensitive to detecting gender and age differences compared to the agreeableness measures we included. Therefore, the Dirty Dozen scales do also appear to represent one broad construct related to agreeableness in our data. However, given that there were also differences between the findings for the three Dirty Dozen scales, the Dirty Dozen may be best described as a measure capturing facet-level constructs representing the low end of agreeableness and, in the case of egocentricity, honesty-humility.

4.4. Strengths and limitations

The main strength of the present research was the large total sample size, which we achieved by pooling several samples. This led to increased statistical power and increased heterogeneity in the final sample. Another strength is our use of two studies of which one was longitudinal. This allowed us to replicate part of our cross-sectional between-person age trends as within-person developmental changes (see Fig. 1). Finally, the inclusion of agreeableness data alongside the Dirty Dozen data was a strength, as this allowed us to examine whether the Dirty Dozen constructs were not simply reflective of low agreeableness.

However, there also are several limitations. First, our longitudinal replication study was restricted to adolescents. In addition, we only had three longitudinal measurement occasions, whereas the reliability of growth estimates increases substantially when more occasions are added (Willett et al., 1998). Cohort-sequential studies with a larger sample, larger age range, and more longitudinal follow-ups are needed to draw more generalizable conclusions.

Second, we focused on mean-level age trends and devoted little attention to individual differences in growth. Significant variance estimates for several slope factors showed that such individual differences existed. Growth mixture modeling could be used to capture such individual differences. The longitudinal sample in the current study was rather small, making it less appropriate for such techniques. However, future studies could employ such analyses to examine whether there are different developmental trajectories of dark personality that align with the ones that have been proposed for antisocial behavior (Moffitt, 1993). Future research should also try to explain individual differences in developmental change. Recently, Grosz et al. (2017) showed that life events predicted individual difference in developmental change in Machiavellianism and narcissism for young adults. Such work could be extended to include other predictors and age groups, and broad-bandwidth measures of multiple dark features.

A third, related point is that more focus on individual differences in change could also shed light on the processes of change. Given the evolutionary explanations for the existence of dark personality features (e.g., McDonald, Donnellan, & Navarette, 2012), it would make sense to test evolutionary principles such as frequency-dependency. For example, particular dark feature levels may only predict success in attaining status if there are few individuals with high levels of such features in the social environment, whereas high levels may become counterproductive when these become more frequent (Penke, Denissen, & Miller, 2007).

Fourth, the present data are representative for Dutch-speaking individuals. There are similarities between age trends in Big Five dimensions as observed in a meta-analysis on predominantly Western European countries (Denissen et al., 2013) and those observed in a sample representing English speaking countries (78% North Americans; Soto et al., 2011). As such, the age trends we found may also generalize to other Western countries, but are unlikely

to be representative for all countries within and especially outside of the Western world. However, gender differences in personality features are known to vary by culture, typically with smaller gender differences in more egalitarian countries (Schmitt et al., 2017) like Belgium and The Netherlands. Future studies could thus examine the cross-national generalizability of the gender differences and age trends we observed.

Fifth, we examined gender and age differences in the Dirty Dozen scales without accounting for their empirical overlap. Accounting for overlap by using bi-factor models or by simply entering multiple related constructs simultaneously into regression-based models is common practice in research on the Dark Triad and the Dirty Dozen (Czarna et al., 2016; Muris et al., 2017). However, validity issues may results from such procedures, as the resulting residual-based variables may or may not provide a good representation of the original construct. Hence, recent research showed that the replicability of results based on associations with residual-based analyses for popular Dark Triad measures including the Dirty Dozen can be problematic (Vize et al., in press A). For that reason, we refrained from using bi-factor models and working with residualized variables.

Sixth, we had sufficient data on agreeableness to compare age and gender trends and examine longitudinal associations with the Dirty Dozen scales, but were unable to include honesty-humility in these analyses. Future studies could conduct analyses similar to the analyses we presented for agreeableness to better examine what the Dirty Dozen scales capture. However, we were able to include both the BFI and the NEO-FFI-3 agreeableness scales. This is important as, compared to the NEO-FFI, the BFI correlates somewhat weaker with dark personality features (Miller, Gaughan, Maples, & Price, 2011; Vize et al., in press B). Unfortunately, we only had both measures available in the young and middle adulthood groups for women in Study 1 (and not in Study 2), which is insufficient to draw robust conclusions on the similarity and differences of age and gender trends for BFI and NEO-FFI-3 agreeableness compared to those trends for the Dirty Dozen. Post-hoc analyses presented in Section 10 of the Supplementary Material do suggest that especially the Dirty Dozen total score and its manipulation scale are more strongly associated with the NEO-FFI-3 agreeableness score than with the BFI agreeableness score. Those findings do suggest that we might have found more evidence for similarities in age and gender trends in Study 1 and for correlated change in Study 2 with the NEO-FFI-3.

5. Conclusion

The present study provides a step forward in our understanding of gender and age trends in dark personality features. Compared to women, men tended to have higher levels on egocentricity, callous affect, and manipulation, but the magnitude of these differences varied with age. For both men and women, levels of all three features increased through adolescence to peak in young adulthood, after which they gradually decreased towards late adulthood. There were similarities in the gender and age trends between these dark features and agreeableness, but the Dirty Dozen measure that we used to capture dark features seemed slightly more sensitive than agreeableness measures in detecting these trends. Therefore, our findings, combined with previous findings, suggest that the dark features captured with the Dirty Dozen may represent facets capturing the low end of agreeableness and, in the case of egocentricity, honesty-humility. Our findings provide some hints, but also raise further questions about for whom, under what conditions, and in which age periods certain levels of dark features may be more and less functional. The rather strong age-related changes in mean levels across the lifespan suggest that a developmental-contextual perspective to the study of the dark side of personality is crucial.

Author contributions

Theo A. Klimstra was responsible for conceiving of and designing the studies, collecting part of the data, analysing and interpreting the data, and writing the paper. Bertus F. Jeronimus and Jelle J. Sijtsema designed and collected data of several of the studies from which we used to create the pooled cross-sectional dataset. Bertus F. Jeronimus, Jelle J. Sijtsema, and Jaap J. A. Denissen actively assisted in designing the studies, analysing and interpreting the data, and writing and revising the paper.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Preparation of this manuscript was supported by the Netherlands Organization for Scientific Research via a Veni grant (016.195.405) to Bertus F. Jeronimus and a Vidi grant (452–14-013) to Theo A. Klimstra.

Acknowledgement

The two studies that are presented in this manuscript were not preregistered. We wish to thank all participants and collaborators of all the studies included in this manuscript for their participation in and invaluable contribution to this research project.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jrp.2020.103915.

References

Allemand, M., & Martin, M. (2016). On correlated change in personality. *European Psychologist, 21*, 237–253.
American Psychiatric Association (2013). *Diagnostic and statistical manual of mental Disorders* (5th ed.). Washington, DC: American Psychiatric Association.
Arber, S., Davidson, K., & Ginn, J. (2003). Changing approaches to gender and later life. In S. Arber, K. Davidson, & J. Ginn (Eds.), *Gender and ageing: Changing roles and relationships* (pp. 1–14). Maidenhead, PA: Open University Press.
Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 11*, 150–166.
Ashton, M. C., & Lee, K. (2016). Age trends in HEXACO-PI-R self-reports. *Journal of Research in Personality, 64*, 102–111.
Back, M. D., Küfner, A. C. P., Dufner, M., Gerlach, T. M., Rauthmann, J. F., & Denissen, J. J. A. (2013). Narcissistic admiration and rivalry: Disentangling the bright and dark sides of narcissism. *Journal of Personality and Social Psychology, 105*, 1013–1037.
Barelds, D. P. H. (2016). Psychometric properties of the Dark Triad Dirty Dozen (DTDD) among working adults. *Gedrag & Organisatie, 29*, 347–364.
Björkqvist, K., Lagerspetz, K. M. J., & Kaukiainen, A. (1992). Do girls manipulate and boys fight? Developmental trends in regard to direct and indirect aggression. *Aggressive Behavior, 18*, 117–127.
Boele, S., Sijtsema, J. J., Klimstra, T. A., Denissen, J. J. A., & Meeus, W. H. J. (2017). Person-group dissimilarity in personality and peer victimization. *European Journal of Personality, 31*, 220–233.
Bonifay, W., Lane, S. P., & Reise, S. P. (2016). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science, 5*, 184–186.
Chaplin, T. M. (2015). Gender and emotion expression: A developmental contextual perspective. *Emotion Review, 7*, 14–21.
Chaplin, T. M., & Aldao, A. (2013). Gender differences in emotion expression in children: A meta-analytic review. *Psychological Bulletin, 139*, 735–765.
Chiorri, C., Garofalo, C., & Velotti, P. (2017). Does the dark triad manifest similarly in men and women? Measurement invariance of the dirty dozen across sex. *Current Psychology, 1–7*. https://doi.org/10.1007/s1214.
Cillessen, A. H. N., & Mayeux, L. (2004). From censure to reinforcement: Developmental changes in the association between aggression and social status. *Child Development, 75*, 147–163.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*, 829–836.

Craker, N., & March, E. (2016). The dark side of Facebook®: The Dark Tetrad, negative social potency, and trolling behaviours. *Personality and Individual Differences, 102*, 79–84.

Crick, N. R. (1997). Engagement in gender normative versus nonnormative forms of aggression: Links to social-psychological adjustment. *Developmental Psychology, 33*, 610–617.

Czarna, A. Z., Jonason, P. K., Dufner, M., & Kossowska, M. (2016). The dirty dozen scale: Validation of a Polish version and extension of the nomological net. *Frontiers in Psychology, 7*, 445.

De Bolle, M., Beyers, W., De Clercq, B., & De Fruyt, F. (2012). General personality and psychopathology in referred and nonreferred children and adolescents: An investigation of continuity, pathoplasty and complication models. *Journal of Abnormal Psychology, 121*, 958–970.

De Fruyt, F., & Hoekstra, H. (2014). *NEO-PI-3 persoonlijkheidsvragenlijst*. Amsterdam: Hogrefe.

Denissen, J. J. A., Geenen, R., van Aken, M. A. G., Gosling, S. D., & Potter, J. (2008). Development and validation of a Dutch translation of the Big Five Inventory (BFI). *Journal of Personality Assessment, 90*, 152–157.

Denissen, J. J. A., Van Aken, M. A. G., Penke, L., & Wood, D. (2013). Self-regulation underlies temperament and personality: An integrative developmental framework. *Child Development Perspectives, 7*, 255–260.

Dufner, M., Rauthmann, J. F., Czarna, A. Z., & Denissen, J. J. (2013). Are narcissists sexy? Zeroing in on the effect of narcissism on short-term mate appeal. *Personality and Social Psychology Bulletin, 39*, 870–882.

Ebner, N. C., Freund, A. M., & Baltes, P. B. (2006). Developmental changes in personal goal orientation from young to late adulthood: From striving for gains to maintenance and prevention of losses. *Psychology and Aging, 21*, 664–678.

Erikson, E. H. (1950). *Childhood and society*. New York, NY: Norton.

Fagot, B. I., Hagan, R., Leinbach, M. D., & Kronsberg, S. (1985). Differential reactions to assertive and communicative acts of toddler boys and girls. *Child Development, 56*, 1499–1505.

Few, L. R., Lynam, D. R., Maples, J. L., MacKillop, J., & Miller, J. D. (2015). Comparing the utility of DSM-5 Section II and III antisocial personality disorder diagnostic approaches for capturing psychopathic traits. *Journal of Personality Disorders, 6*, 64–74.

Flanagan, C. A., & Stout, M. (2010). Developmental patterns of social trust between early and late adolescence: Age and school climate effects. *Journal of Research on Adolescence, 20*, 748–773.

Fox, J., & Rooney, M. C. (2015). The Dark Triad and trait self-objectification as predictors of men's use and self-presentation behaviors on social networking sites. *Personality and Individual Differences, 76*, 161–165.

Furnham, A., Richards, S. C., & Paulhus, D. L. (2013). The dark triad of personality: A 10 year review. *Social and Personality Psychology Compass, 7*, 199–216.

Grijalva, E., Newman, D. A., Tay, L., Donnellan, M. B., Harms, P. D., Robins, R. W., & Yan, T. (2015). Gender differences in narcissism: A meta-analytic review. *Psychological Bulletin, 141*, 261–310.

Grosz, M. P., Göllner, R., Rose, N., Spengler, M., Trautwein, U., Rauthmann, J. F., ... Roberts, B. W. (2017). The development of narcissistic admiration and machiavellianism in early adulthood. *Journal of Personality and Social Psychology, 116*, 467–482.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The big five inventory – Versions 4a and 54*. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

Johnson, W. L., Giordano, P. C., Manning, W. D., & Longmore, M. A. (2015). The age–IPV curve: Changes in the perpetration of intimate partner violence during adolescence and young adulthood. *Journal of Youth and Adolescence, 44*, 708–726.

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

Jonason, P. K., & McCain, J. (2012). Using the HEXACO model to test the validity of the Dirty Dozen measure of the dark triad. *Personality and Individual Differences, 53*, 935–938.

Jonason, P. K., & Webster, G. D. (2010). The dirty dozen: A concise measure of the Dark Triad. *Psychological Assessment, 22*, 420–432.

Karmin, M., Saag, L., Vicente, M., Sayres, M. A. W., Järve, M., Talas, U. G., ... Pagani, L. (2015). A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Research, 25*, 459–466.

Klimstra, T. A., Bleidorn, W., Asendorpf, J. B., van Aken, M. A. G., & Denissen, J. J. A. (2013). Correlated change of big five personality traits across the lifespan: A search for determinants. *Journal of Research in Personality, 47*, 768–777.

Klimstra, T. A., Hale, W. W., Raaijmakers, Q. A. W., Branje, S. J. T., & Meeus, W. H. J. (2009). Maturation of personality in adolescence. *Journal of Personality and Social Psychology, 96*, 898–912.

Klimstra, T. A., Noftle, E. E., Luyckx, K., Goossens, L., & Robins, R. W. (2018). Personality development and adjustment in college: A multifaceted, cross-national view. *Journal of Personality and Social Psychology, 115*, 338–361.

Klimstra, T. A., Sijtsema, J. J., Henrichs, J., & Cima, M. J. (2014). The dark triad of personality in adolescence: Psychometric properties of a concise measure and associations with adolescent adjustment from a multi-informant perspective. *Journal of Research in Personality, 53*, 84–92.

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford.

Krueger, R. F., Derringer, J., Markon, K. E., Watson, D., & Skodol, A. E. (2012). Initial construction of a maladaptive personality trait model and inventory for. DSM–5. *Psychological Medicine, 42*, 1879–1890.

Lerner, R. M., & Overton, W. F. (2017). Reduction to absurdity: Why epigenetics invalidates all models involving genetic reduction. *Human Development, 60*, 107–123.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variable in SEM and MACS models. *Structural Equation Modeling, 13*, 59–72.

Lynam, D. R., & Miller, J. D. (2019). The basic trait of antagonism: An unfortunately underappreciated construct. *Journal of Research in Personality, 81*, 118–126.

Maples, J. L., Lamkin, J., & Miller, J. D. (2014). A test of two brief measures of the dark triad: the dirty dozen and the short dark triad. *Psychological Assessment, 26*, 326–331.

McAdams, D. P., & Olson, B. D. (2010). Personality development: Continuity and change over the life course. *Annual Review of Psychology, 61*, 517–542.

McDonald, M. M., Donnellan, M. B., & Navarette, C. D. (2012). A life history approach to understanding the Dark Triad. *Personality and Individual Differences, 52*, 601–605.

Miller, J. D., Few, L. R., Seibert, A., Watts, A., Zeichner, A., & Lynam, D. R. (2012). An examination of the dirty dozen measure of psychopathy: A cautionary tale about the costs of brief measures. *Psychological Assessment, 24*, 1048–1053.

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

Miller, J. D., Hyatt, C. S., Maples-Keller, J. L., Carter, N. T., & Lynam, D. R. (2017). Psychopathy and Machiavellianism: A distinction without a difference? *Journal of Personality, 85*, 439–453.

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review, 100*, 674–701.

Moshagen, M., Zettler, I., & Hilbig, B. E. (2020). Measuring the dark core of personality. Psychological Assessment (in press), http://dx.doi.org/10.1037/pas0000778.

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

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nesselroade, J. R., & Baltes, P. B. (1974). Adolescent personality development and historical change, 1970–1972. *Monographs of the Society for Research in Child Development, 39*, 1–80.

Neumann, C. S., Hare, R. D., & Pardini, D. A. (2014). Antisociality and the construct of psychopathy: Data from across the globe. *Journal of Personality, 83*, 678–692.

Oltmanns, T. F., & Powers, A. D. (2012). Gender and personality disorders. In T. A. Widiger (Ed.), *Oxford handbook of personality disorders* (pp. 206–218). New York: Oxford University Press.

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

Penke, L., Denissen, J. J. A., & Miller, G. F. (2007). The evolutionary genetics of personality. *European Journal of Personality, 21*, 549–587.

Rauthmann, J. F. (2013). Investigating the MACH–IV with item response theory and proposing the Trimmed MACH. *Journal of Personality Assessment, 95*, 388–397.

Roberts, B. W., & Nickel, L. (2017). A critical evaluation of the Neo-Socioanalytic Model of personality. In J. Specht (Ed.), *Personality development across the lifespan*. San Diego, CA: Elsevier.

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life-course: A meta-analysis of longitudinal studies. *Psychological Bulletin, 132*, 1–25.

Rydell, A. M., Berlin, L., & Bohlin, G. (2003). Emotionality, emotion regulation, and adaptation among 5-to 8-year-old children. *Emotion, 3*, 30–47.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. V. Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference Chi square test statistic for moment structure analysis. *Psychometrika, 66*, 507–514.

Schmitt, D. P., Long, A. E., McPhearson, A., O'brien, K., Remmert, B., & Shah, S. H. (2017). Personality and gender differences in global perspective. *International Journal of Psychology, 52*, 45–56.

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology, 94*, 168–182.

Schwartz, S. J., Beyers, W., Luyckx, K., Soenens, B., Zamboanga, B. L., Forthun, L. F., ... Waterman, A. S. (2011). Examining the light and dark sides of emerging adults' identity: A study of identity status differences in positive and negative psychosocial functioning. *Journal of Youth and Adolescence, 40*, 839–859.

Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences, 106*, 15073–15078.

Sherman, E. D., Lynam, D. R., & Heyde, B. (2014). Agreeableness accounts for the factor structure of the Youth Psychopathic Traits Inventory. *Journal of Personality Disorders, 28*, 262–280.

Sijtsema, J. J., Lindenberg, S. M., Ojanen, T. J., & Salmivalli, C. (April, 2016). Aggression and the balance between status and affection goals throughout the lifespan. Invited presentation at the symposium on Aggression and Bullying, Groningen, the Netherlands.

Soto, C. J., & John, O. P. (2014). Traits in transition: The structure of parent-reported personality traits from early childhood to early adulthood. *Journal of Personality, 82*, 182–199.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big-Five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology, 100*, 330–348.

Spurk, D., & Hirschi, A. (2018). The Dark Triad and competitive psychological climate at work: A model of reciprocal relationships in dependence of age and organization change. *European Journal of Work and Organizational Psychology, 27*, 736–751.

Statistics Netherlands (2016). Statline: Kerkelijke Gezindte en Kerkbezoek; Vanaf 1849; 18 jaar of Ouder. [Statline: Religious Denomination and Church Attendance; From 1849 Onwards; 18 Years or Older]. Available from https://opendata.cbs.nl/#/CBS/nl/dataset/37944/table?ts=1514905629072.

Statistics Netherlands (2017). Statline: Arbeidsdeelname en Werkloosheid per Maand [Statline: Labor Participation and Unemployment by Month]. Available from http://statline.cbs.nl/Statweb/publication/?DM=SLNL&PA=80590NED&D1=13&D2=0&D3=1-3&D4=12,25,38,51,64,77,90,103,116,129,142,155,168,181&HDR=T&STB=G1, G2,G3&CHARTTYPE=1&VW=G.

Steenkamp, J. M., & Baumgartner, H. (1998). Assessing measurement invariance in crossnational consumer research. *Journal of Consumer Research, 25*, 78–90.

Steensma, T. D., Kreukels, B. P., de Vries, A. L., & Cohen-Kettenis, P. T. (2013). Gender identity development in adolescence. *Hormones and Behavior, 64*, 288–297.

Steinmetz, H. (2010). Estimation and comparison of latent means across cultures. In P. Schmidt, J. Billiet, & E. Davidov (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 85–116). Oxford: Routledge.

Steinmetz, H. (2013). Analyzing observed composite differences across groups: Is partial measurement invariance enough?. *Methodology, 9*, 1–12.

Sweeten, G., Piquero, A. R., & Steinberg, L. (2013). Age and the explanation of crime, revisited. *Journal of Youth and Adolescence, 42*, 921–938.

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*, 486–492.

Vize, C. E., Collison, K. L., Miller, J. D., & Lynam, D. R. (in press A). Using item-level analyses to better understand the consequences of partialing procedures: An example using the Dark Triad. Journal of Personality, http://dx.doi.org/10.1111/jopy.12521.

Vize, C. E., Collison, K. L., Miller, J. D., & Lynam, D. R. (in press B). The "core" of the dark triad: A test of competing hypotheses. Personality Disorders: Theory, Research, and Treatment, https://doi.org/10.1037/per0000386.

Vize, C. E., Lynam, D. R., Collison, K. L., & Miller, J. D. (2018). Differences among Dark Triad components: A meta-analytic investigation. *Personality Disorders: Theory, Research, and Treatment, 9*, 101–111.

Webster, G. D., & Jonason, P. K. (2013). Putting the "IRT" in "dirty": Item response theory analyses of the dark triad dirty dozen – An efficient measure of narcissism, psychopathy, and machiavellianism. *Personality and Individual Differences, 54*, 302–306.

West, C., & Zimmerman, D. H. (1987). Doing gender. *Gender & Society, 1*, 125–151.

Wetzel, E., Brown, A., Hill, P., Chung, J. M., Robins, R., & Roberts, B. W. (2017). The narcissism epidemic is dead; long live the narcissism epidemic. *Psychological Science, 28*, 1833–1847.

Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604.

Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology, 10*, 395–426.