Test-retest reliability and construct validity of the brief dark triad measurements

Citation for published version:
Dragostinov, Y & Mõttus, R 2022, 'Test-retest reliability and construct validity of the brief dark triad measurements', *Journal of personality assessment*. https://doi.org/10.1080/00223891.2022.2052303

Digital Object Identifier (DOI):
10.1080/00223891.2022.2052303

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Journal of personality assessment

JOURNAL OF PERSONALITY ASSESSMENT
https://doi.org/10.1080/00223891.2022.2052303

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Test-Retest Reliability and Construct Validity of the Brief Dark Triad Measurements

Yavor Dragostinov[1] and René Mõttus[1,2]

[1]Department of Psychology, University of Edinburgh; [2]Institute of Psychology, University of Tartu

ABSTRACT

Despite the widespread use of the Dirty Dozen (DD) and Short Dark Triad (SD3) as inventories for antagonist personality constructs, appropriately powered studies on their test-retest reliability (r^{tt}) are lacking. We report the 12-day r^{tt}-s of the DD and SD3 scales. Leveraging the test-retest data, we also calculated their convergent and discriminant correlations while controlling for measurement error. Median r^{tt}-s were .87 and .90 ($N = 500$) for the DD and SD3 scales, respectively, substantially higher than their internal consistencies. Convergent correlations were .77, .63 and .64 for Machiavellianism, Narcissism and Psychopathy, respectively. Discriminant correlations between the Machiavellianism and Psychopathy scales had a median of .65, pointing to their being effectively indistinguishable traits in the SD3 and DD. The DD and SD3 items had median r^{tt}-s of .69 and .71, respectively. We emphasize the importance of the r^{tt} for scale development and validation.

ARTICLE HISTORY
Received 24 September 2021
Accepted 7 March 2022

Since Paulhus and Williams (2002) proposed the *Dark Triad* (DT) model of personality constructs – narcissism, Machiavellianism and psychopathy – thousands of DT studies have generated numerous peer-reviewed articles, books and coverage by the lay media (Miller et al., 2019). Although several instruments have been created to measure the DT constructs, there is little work yet to evaluate their test-retest reliability (r^{tt}). We address this gap, while also highlighting the importance of the retest method for estimating one of the key properties of psychometric scales – their reliability – over the more commonly used method of internal consistency. Finally, we leverage the test-retest data to assess the construct validity of the DT constructs, controlling for random and situation-specific measurement errors.

Origins and core of the DT constructs

Rather than defining narcissism and psychopathy through their clinical origin (see Furnham & Crump, 2005) and thereby contrasting normal with abnormal, the DT model conceptualizes the two constructs from a subclinical perspective, expecting them to vary continuously through population (Furnham et al., 2013; Paulhus & Williams, 2002). A person who is comparatively high in psychopathy tends to be impulsive and have little empathy, whereas someone comparatively high in narcissism tends to feel entitled and superior to others. Completing the DT, Machiavellianism, reflects the belief that interpersonal manipulation, lack of principles and cynicism are key to life success, and some follow this belief more, some less.

At least in part, variance in these overlapping DT constructs could be driven by an underlying antagonism trait which varies throughout the population (see Vize et al., 2020). If so, labeling the constructs with the term “dark” may not be helpful since all people have some levels of them. For example, Vize et al. (2020) argue that the term inadvertently stigmatizes “antagonistic individuals who may seek out treatment” (p. 98).

Measurement of the DT constructs

Due to the growing interest in the DT constructs, brief and efficient scales for measuring them are widely sought. Currently, two instruments dominate the DT literature: the Dirty Dozen (DD; Jonason & Webster, 2010), a 12-item questionnaire that has currently been cited over 1300[1] times, and the Short Dark Triad (SD3; Jones & Paulhus, 2014), a 27-item questionnaire that has been cited over 1400[1] times.

However, their brevity may have come at a cost (see Maples et al., 2014). For example, previous studies have shown that the DD has low convergent validity in relation to longer DT scales (Jonason & Webster, 2010; Maples et al., 2014; Miller et al., 2012), possibly due to poor coverage of some features of psychopathy such as antagonism,

CONTACT Yavor Dragostinov y.dragostinov@sms.ed.ac.uk Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK.

Supplemental data for this article is available online at https://doi.org/10.1080/00223891.2022.2052303.

[1]According to Google Scholar.

grandiosity, manipulativeness and disinhibition (Miller et al., 2012). On the other hand, the SD3 appears to neglect the vulnerable features of narcissism (Maples et al., 2014). Furthermore, neither the DD nor the SD3 Machiavellianism scales align with expert intuitions about the trait since their scores are negatively associated with conscientiousness, whereas experts consider machiavellian individuals to be planful, deliberate, ambitious, and strategic (Miller et al., 2017). Finally, psychopathy and Machiavellianism are often so highly correlated in both DD and SD3 that these constructs may be considered redundant (Vize et al., 2018), raising concerns about their construct validity in the SD3 and DD.

Reliability in personality measurements

Reliability is a fundamental property of psychological assessment scales, showing how much useful information the scale scores contain and how strongly they can correlate with anything else. Among the common methods of estimating reliability is internal consistency, often measured using the coefficient alpha (α; Cronbach, 1951). It is based on correlations between a scale's items being administered simultaneously, assuming they all evaluate exactly the same trait. Therefore, any inconsistencies among their scores should reflect measurement unreliability (APA Dictionary of Psychology, 2020). The r^{tt}, on the other hand, represents the degree to which scores of the same items remain stable when administered multiple times.

Important psychometric textbooks such as Nunnally and Bernstein (1994) have suggested, for example, that "coefficient α usually provides a good estimate of reliability because sampling of content is usually the major source of measurement error for static constructs" (p. 252). Accordingly, α is routinely reported in test manuals and journal articles. However, many authors have highlighted its limitations, including Cronbach himself (Cronbach et al., 1963; Cronbach & Shavelson, 2004; Murphy & Davidshofer, 2001; Sijtsma, 2009), primarily because the assumption that all scale items measure exactly the same trait is untenable. On top of their shared variance that contributes toward internal consistency, most personality scale items contain unique variance that has the key trait properties of stability over many years, cross-method agreement (self- versus informant-ratings), unique etiology (e.g., sibling similarity or developmental trajectories) and predictive validity (Mõttus et al, 2017, 2019, 2020; Seeboth & Mõttus, 2018). Internal consistency is therefore prone to underestimate reliability as it misclassifies veridical unique trait information in items – or, personality "nuances" (McCrae, 2015; Mõttus et al., 2017) – as measurement error. But it may also inflate reliability because occasion-specific (state-like) methodological artifacts such as mood can influence responses to all items alike.

It is not surprising then that scales' internal consistencies do not track their validities, even though they should because reliability is a key assumption of validity (Henry et al., 2022; McCrae, 2015; McCrae et al., 2011): items' unique variance is a valid part of scale scores, but it is misclassified as measurement error by internal consistency. In fact, high internal consistency can even have undesirable consequences. First, achieving it by writing highly similar items leads to construct content not being covered sufficiently broadly, thereby limiting the scale's validity. Second, achieving it by including a high number of items constrains the number of constructs that can be measured, because typically only a limited number of items can be administered.

The r^{tt} assumes that individual differences in constructs that interest researchers are stable over at least relatively short time periods and that observed fluctuations represent measurement error (Murphy & Davidshofer, 2001). Therefore, reliable variance, for the purpose of measuring a trait that *is defined* as stable, is the variance that *is* in fact stable. The r^{tt} does not rely on the assumption that all items measure nothing but a single unidimensional trait, and it is less distorted by state-like artifacts. Indeed, unlike internal consistency, scales' r^{tt}-s track their validities (Henry et al., 2022; McCrae, 2011), making it the preferred method of estimating reliability (Lowman et al., 2018; McCrae, 2015; Revelle & Condon, 2019;). Besides, it can be calculated for individual test items, allowing researchers to select the most reliable items into their scales. Also, corrections of correlations between scale scores for measurement error that use internal consistencies often result in correlations above 1.00, whereas using r^{tt} rarely results in such off-limit correlations (Lowman et al., 2018).

Retest reliability of the DT scales

For the three DD scales, r^{tt}-s between .71 and .88 have been reported (Jonason & Webster, 2010), but these were calculated in a sample of only 94 participants, only 60 of whom provided complete data; it takes much higher sample sizes to calculate reliable correlations (Schönbrodt & Perugini, 2013). In a sample of 112 Spanish participants assessed with the DD, r^{tt} -s of .60, .70, and .59 were reported for the Machiavellianism, Narcissism, and Psychopathy scales, respectively (Maneiro et al., 2019); however, the retesting interval was 6 months, meaning that measurement unreliability may have been conflated with actual trait change. Shorter retesting intervals – around two weeks – are more appropriate, because true change is less likely to occur (Chmielewski & Watson, 2009; Henry et al., 2022; Mõttus et al., 2019). Macedo et al. (2017) assessed 30 Portuguese participants twice over 6 weeks and reported an r^{tt} of .70 for the DD overall score (sum of all items). In a German sample of 221 participants tested twice over 4 weeks with the SD3, r^{tt}-s of .81, .74, and .83 were reported for the Machiavellianism, Narcissism and Psychopathy scales, respectively (Malesza et al., 2019); to our knowledge, this is only the second published study to assess the r^{tt}-s of the SD3.[2]

[2] At the Annual Convention of the Society for Personality and Social Psychology in 2011, Paulhus and Jones reported r^{tt} for the SD3 (with coefficients ranging from .77 to .84) at a conference. However, further details are unknown as the details of these findings have not been published – (Paulhus, & Jones, 2011).

Table 1. Intercorrelations of r^{tt}-s, α-s, ω-s and MICs for the SD3 and DD subscales.

Trait	Measurement	Test-Retest	Alpha	Omega	MIC
Machiavellianism	Dirty Dozen	.83 [.78; .88]	.85	.87	.58
	Short Dark Triad	.85 [.80; .90]	.78	.82	.26
Narcissism	Dirty Dozen	.81 [.76; .86]	.80	.82	.48
	Short Dark Triad	.88 [.84; .92]	.75	.79	.25
Psychopathy	Dirty Dozen	.80 [.75; .85]	.71	.80	.37
	Short Dark Triad	.85 [.80; .90]	.76	.81	.24

Note: MIC – mean inter-item correlations. Numbers in square brackets represent 95% confidence intervals.

The current study

More appropriately powered research is needed to establish the key psychometric property, r^{tt}, of the most widely used DT scales. Among other things, this would allow researchers to better interpret the associations of the DT scales with one another, as well as with other variables, and appropriately correct these associations for unreliability. We estimated the DD and SD3 scales r^{tt}-s and compared these to their internal consistencies, expecting r^{tt}-s to exceed internal consistencies. The test-retest data also allowed us to estimate the convergent and discriminant validities – collectively, the construct validity – of the DT scales, while considering random and situation-specific measurement errors. For this, we calculated cross-lagged correlations between all six subscales of these two instruments, correcting these for their r^{tt}-s as the theoretical upper limit of the cross-lagged correlations.

Methods

Participants and procedures

Participants were recruited via Prolific Academic, using a cohort of returning participants ($N = 639$) that took part of an ongoing research project. A total of 509 people completed the SD3 and the DD twice over a 12-day interval. Following recommendations from other researchers (Henry et al., 2022; Wood et al., 2017), we excluded participants whose profile consistency (q, calculated as the correlation between the profiles of corresponding item responses at each measurement occasion) was lower than three standard deviations below the sample median of .79. The vast majority of participants had a high or very high profile consistency (25th quantile = .68, 3 SDs below Mdn = .32), which is impossible to achieve with random responding; those who did not respond consistently had remarkably lower profile correlations (9 had profile correlations below .32, some even negative). This left us with a final sample of $N = 500$ (M age = 30.17, SD = 10.12; 235 males, 262 females and 3 individuals who did not identify as either male or female). Each of these participants received £1.20 for their participation.

Measures

The survey had a total of 39 items: the SD3, consisting of 27 items, nine per subscale, rated using a 5-point Likert scale; and the DD, consisting of 12 items, had four items per subscale with a 9-point Likert scale. These measures were described in detail earlier.

Data analyses

The r^{tt}-s of the overall scores (sums of all items, regardless of the DT trait for which they were designed), the Machiavellianism, Narcissism and Psychopathy scale scores and individual items were estimated as the correlations between their respective scores at time 1 and time 2.[3] We also calculated Cronbach's α-s for the total scores (all items) and DT scales for comparison, alongside mean inter-item correlations (MIC). Previous work (Dinić et al., 2018) has found MICs of the DT scales to be above the .50 upper bond suggested by Clark and Watson (1995), indicating potential item redundancy. Omegas (ω) (Revelle & Condon, 2019) were also calculated using the default argument values of the omega function in the "psych" R package (Revelle, 2021; see Table 1).

To estimate the convergent and discriminant validity of the six DD and SD3 scales, we calculated their cross-lagged correlations (i.e., cross-scale, cross-time-point correlations) and divided each by the geometric mean of the r^{tt}-s of the scales involved. In doing so, we assumed a model whereby trait measurements at each time-point contain a) reliable variance shared by two traits (including general rater-specific method effects), b) reliable variance unique to each trait (including trait-specific rater-specific method effects), c) random error and d) state-specific effects. Cross-lagged correlations reflect a), whereas r^{tt}-s reflect a) and b): hence, dividing the former by the latter gives a purer estimate of the degree of shared variance among traits than cross-sectional correlations. We expect the corrected (convergent) correlations among the corresponding DD and SD3 scales (e.g., DD Narcissism and SD3 Narcissism) to be higher than the corrected (discriminant) correlations among non-corresponding scales (e.g., DD Narcissism and SD3 Psychopathy or DD Narcissism and DD Psychopathy). However, we expected the convergent and discriminant correlations for Machiavellianism and psychopathy to be more similar than other discriminant correlations (Vize et al., 2018).

Data availability statement

Code, data, and materials that may be used to reproduce all analyses can be found at https://osf.io/mwygk/.

[3] We also estimated the degree to which DT items reflect unique variance, or personality "nuances" (McCrae, 2015; Mõttus et al., 2017) which are available at https://osf.io/2gfyh/.

Table 2. Correlations of the DT scales.

	Machiavellianism		Narcissism		Psychopathy	
Measure	DD	SD3	DD	SD3	DD	SD3
DD Machiavellianism	—					
SD3 Machiavellianism	.77 [.71; .83]	—				
DD Narcissism	.55 [.48; .62]	.49 [.41; .57]	—			
SD3 Narcissism	.34 [.26; .42]	.35 [.27; .43]	.63 [.56; .70]	—		
DD Psychopathy	.60 [.53; .67]	.59 [.52; .66]	.31 [.23; .39]	.14 [.05; .22]	—	
SD3 Psychopathy	.71 [.65; .77]	.64 [.57; .71]	.41 [.33; .49]	.39 [.31; .47]	.64 [.57; .71]	—

Note: DD – Dirty Dozen; SD3 – Short Dark Triad. Numbers in square brackets represent 95% confidence intervals.

Results

The DD had an overall score r^{tt} of .87, versus an α of .85. Its Machiavellianism, Narcissism and Psychopathy scales had r^{tt}-s of .83, .81, and .80 respectively, whereas the respective α-s were .84, .78, and .69 (Table 1). The SD3 overall scores had a r^{tt} of .90, versus an α of .85. Its Machiavellianism, Narcissism and Psychopathy scales had r^{tt}-s of .85, .88, and .85, respectively, compared to respective α-s of .77, .74, and .74. Thus, r^{tt}-s were generally higher than internal consistencies; across the six subscales of the two instruments, the median r^{tt} was .84, whereas the median α was .76. On average, omegas were closer to r^{tt}-s than α-s, but still lower in most cases. Notably, the r^{tt}-s were only slightly higher for the longer SD3 than for the shorter DD.

Convergent and discriminant validity of the DT scales

In the cross-lagged correlations corrected for r^{tt}-s, convergent correlations (Table 2) for the Psychopathy and Machiavellianism scales were .64 and .77, respectively, whereas their discriminant correlations ranged from .59 to .71 with a median of .65. Pointing to poor construct validity, DD Psychopathy was even significantly more strongly correlated with DD Machiavellianism than with SD3 Psychopathy ($t = 3.32$, $p < .001$). The convergent correlation of the two Narcissism scales was .63, whereas their discriminant correlations ranged from .14 to .55, with a median of .35.

Test-Retest reliability of items

The DD items (Table S2 in the supplemental materials) had a median r^{tt} of .71 (.71, .68, and .70 for the Machiavellianism, Narcissism and Psychopathy scales, respectively). The SD3 items had a median r^{tt} of .68 (.61, .70, and .68 for the Machiavellianism, Narcissism and Psychopathy scales, respectively). Despite these fairly consistent mean r^{tt}-s of the scales, there were notable differences between the items within both instruments in their r^{tt}-s, ranging from .52 to .84 for SD3 and from .53 to .81 for the DD. After removing the DT variance from items, their unique variance retained much of the reliability, with mean r^{tt}-s of .49 and .57 for SD3 and DD, respectively (see the supplemental materials).

Discussion

Despite the DD and SD3 being some of the most widely used instruments to measure the DT constructs of Machiavellianism, narcissism and psychopathy, there is a drought of studies to evaluate their test-retest reliability (r^{tt}), an essential property of psychological tests. To our knowledge, this is the first study to provide a sufficiently powered examination of the r^{tt} for both the DD and the SD3. Both instruments had reasonably and comparably high reliabilities, with r^{tt}-s over a 12-day interval ranging from .81 to .88 for the three DT trait scales. For most scales, the r^{tt}-s were higher than the internal consistencies (α-s), suggesting that the reliability of these DT scales has been underestimated so far, since most attempts have relied on internal consistency. We also found that the scales had moderate convergent validity, with the corresponding scales having unreliability-corrected correlations between .63 (Narcissism) and .77 (Machiavellianism), but the Psychopathy and Machiavellianism scales had poor discriminant validity.

The advantages of Test-Retest reliability

Presumably, the r^{tt}-s were generally higher than internal consistencies (respective medians were .84 and .76 for the DT trait scales), because individual items measure systematic variances beyond those for which they have been designed (McCrae & Mõttus, 2019), as also evidenced by the substantial r^{tt}-s of items' unique variance (see the supplemental materials). This finding is also in line with those of Henry and colleagues (2022) who evaluated the r^{tt}-s of the HEXACO-PI-R (Lee & Ashton, 2004). This suggests that internal consistency provides misleading estimates of scales reliabilities and should generally be avoided (McCrae, 2015). However, the difference between r^{tt} and α is most pronounced in case of very low internal consistencies (e.g., Henry et al., 2022); in the current work, there were no very low internal consistencies, so the differences between r^{tt}-s and α-s were generally small.

Additionally, by relying solely on internal consistency measures, researchers miss the possibility of detecting unreliable and therefore potentially flawed items. The r^{tt}-s of the DT items varied from .52 to .84 (with standard errors hence below .04); even when allowing for some sampling error in these estimates, this reflects considerable variation in their reliabilities. Items with low r^{tt}-s should eventually be replaced in these scales and avoided in the future. Therefore, we highly recommend that researchers collect r^{tt} as a routine practice when assessing scales.

Construct validity of the brief DT measures

Having the test-retest data for both instruments, SD3 and DD, we could assess the convergent and discriminant validities of their scales (see Table 2) while considering random and situation-specific measurement errors. As for convergent correlations between different scales designed to measure the same constructs, we found them to be in .60s for narcissism and psychopathy and in .70s for Machiavellianism. These results are similar with the convergent correlations findings from Maples and colleagues (2014) – .61, .54 and .65 for Machiavellianism, narcissism and psychopathy, respectively – although these correlations were not corrected for measurement error. We think that correlations in this range do not provide strong evidence for the convergent validity of the DT measures, as the measures would very often misclassify people. For example, a correlation of .60 means that if people were divided into equal groups of low, medium, and high scores in both measures, only about 53% of people would get a similar result in both tests (Mõttus, 2021).

In line with Maples et al. (2014), our findings also highlight weak discriminant validity between the scales between SD3 Psychopathy and DD Machiavellianism (.71), as well as DD Psychopathy and SD3 Machiavellianism (.59). This means that the Psychopathy and Machiavellianism scales overlapped about as strongly as each of them overlapped with their ostensible counterpart; in fact, one of the discriminant correlations was even significantly higher than one of the convergent correlations. Therefore, instead of a *triad* of antagonistic personality constructs, the two instruments measure a *dyad*, with psychopathy and Machiavellianism effectively being psychometrically indistinguishable.

Some experts (e.g., Furnham et al., 2013) have recommended partialing DT scores from one another to understand their unique relations with various outcomes. However, considering the high overlap between these scores, the residualized constructs tend to become completely different compared to the original constructs, complicating their interpretation (Sleep et al., 2019). Hence, if DT researchers prefer to keep Machiavellianism and psychopathy separate, they should consider refining the definition and/or measurement of these constructs.

Limitations

Our sample was recruited from a paid online service, indicating a possibility of selection bias. Furthermore, most of the participants either were living or were from Western countries, suggesting a potentially W.E.I.R.D. sample (Henrich et al., 2010). Also, future studies should estimate the r^{tt}-s of other DT instruments, as well as the cross-rater agreement of the DT scales since these allow separating rater-specific method effects from trait variance (McCrae & Mõttus, 2019), for estimating both reliability and construct validity.

Conclusions

According to our findings, the SD3 and DD scales appear to provide reliable measurements of the DT traits but, do not have strong convergent validity and, especially psychopathy and Machiavellianism, show poor discriminant validity. We strongly advise researchers to treat r^{tt} as a crucial step in scale development.

Conflict of interest

We have no conflicting interests regarding this article.

Ethics approval statement

This study received ethical approval on the November 10, 2020 (reference number 45-2021/9) from The University of Edinburgh: PPLS Research Ethics.

Open Scholarship

This article has earned the Center for Open Science badges for Open Data and Open Materials through Open Practices Disclosure. The data and materials are openly accessible at https://osf.io/mwygk/. To obtain the author's disclosure form, please contact the Editor.

Funding

This study was supported by the University of Edinburgh: School of Philosophy, Psychology, and Language Sciences (PPLS).

References

APA Dictionary of Psychology. (2020). Retrieved 22 December 2020, from https://dictionary.apa.org/internal-consistency

Chmielewski, M., & Watson, D. (2009). What is being assessed and why it matters: The impact of transient error on trait research. *Journal of Personality and Social Psychology*, *97*(1), 186–202. https://doi.org/10.1037/a0015618

Clark, L.A. & Watson, D. (1995). Construct validity: Basic issues in objective scale development. *Psychological Assessment*, *7*, 309–319.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*(3), 391–418. https://doi.org/10.1177/0013164404266386

Furnham, A., & Crump, J. (2005). Personality traits, types, and disorders: an examination of the relationship between three self-report measures. *European Journal of Personality*, *19*(3), 167–184.

Furnham, A., Richards, S. C., & Paulhus, D. L. (2013). The Dark Triad of personality: A 10 year review. *Social and Personality Psychology Compass*, *7*(3), 199–216. https://doi.org/10.1111/spc3.12018

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2–3), 61–83.

Henry, S., Thielmann, I., Booth, T., & Mõttus, R. (2022). Test-retest reliability of the HEXACO-100-And the value of multiple measurements for assessing reliability. *PloS One*, *17*(1), e0262465.

Jonason, P. K., & Webster, G. D. (2010). The dirty dozen: A concise measure of the dark triad. *Psychological Assessment*, *22*(2), 420–432.

Jones, D. N., & Paulhus, D. L. (2014). Introducing the short dark triad (SD3): A brief measure of dark personality traits. *Assessment*, *21*(1), 28–41.

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, *39*(2), 329–358. https://doi.org/10.1207/s15327906mbr3902_8

Lowman, G. H., Wood, D., Armstrong III, B. F., Harms, P. D., & Watson, D. (2018). Estimating the reliability of emotion measures over very short intervals: *The utility of within-session retest correlations. Emotion*, *18*(6), 896.

Macedo, A., Araújo, A. I., Cabaços, C., Brito, M. J., Mendonça, L., & Pereira, A. T. (2017). Personality dark triad: Portuguese validation of the dirty dozen. *European Psychiatry*, *41*(S1), S711–S711. https://doi.org/10.1016/j.eurpsy.2017.01.1268

Malesza, M., Ostaszewski, P., Büchner, S., & Kaczmarek, M. C. (2019). The adaptation of the Short Dark Triad personality measure–psychometric properties of a German sample. *Current Psychology*, *38*(3), 855–864. https://doi.org/10.1007/s12144-017-9662-0

Maneiro, L., López-Romero, L., Gómez-Fraguela, J. A., Cutrín, O., & Romero, E. (2019). Pursuing the Dark Triad: Psychometric properties of the Spanish version of the Dirty Dozen. *Journal of Individual Differences*, *40*(1), 36–44. https://doi.org/10.1027/1614-0001/a000274

Maples, J. L., Lamkin, J., & Miller, J. D. (2014). A test of two brief measures of the dark triad: The dirty dozen and short dark triad. *Psychological Assessment*, *26*(1), 326–331.

McCrae, R. R. (2015). A more nuanced view of reliability: Specificity in the trait hierarchy. *Personality and Social Psychology Review : An Official Journal of the Society for Personality and Social Psychology, Inc*, *19*(2), 97–112. https://doi.org/10.1177/1088868314541857

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review : An Official Journal of the Society for Personality and Social Psychology, Inc*, *15*(1), 28–50.

McCrae, R. R., & Mõttus, R. (2019). What personality scales measure: A new psychometrics and its implications for theory and assessment. *Current Directions in Psychological Science*, *28*(4), 415–420. https://doi.org/10.1177/0963721419849559

Miller, J. D., Few, L. R., Seibert, L. 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*(4), 1048–1053.

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*(4), 439–453.

Miller, J. D., Vize, C., Crowe, M. L., & Lynam, D. R. (2019). A critical appraisal of the dark-triad literature and suggestions for moving forward. *Current Directions in Psychological Science*, *28*(4), 353–360. https://doi.org/10.1177/0963721419838233

Mõttus, R. (2021, August). What does a correlation say about me? A tutorial on translating correlational research findings to their implications for individual people. https://doi.org/10.31234/osf.io/bpm9y

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*, *112*(3), 474–490. https://doi.org/10.1037/pspp0000100

Mõttus, R., Sinick, J., Terracciano, A., Hřebíčková, M., Kandler, C., Ando, J., Mortensen, E. L., Colodro-Conde, L., & Jang, K. L. (2019). Personality characteristics below facets: A replication and meta-analysis of cross-rater agreement, rank-order stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*, *117*(4), e35–e50.

Mõttus, R., Wood, D., Condon, D. M., Back, M., Baumert, A., Costantini, G., Epskamp, S., Greiff, S., Johnson, W., Lukaszewski, A., Murray, A., Revelle, W., Wright, A., Yarkoni, T., Ziegler, M., & Zimmermann, J. (2020, September). Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the Big Few traits. *European Journal of Personality*, *34*(6), 1175–1201.

Murphy, K. R., & Davidshofer, C. O. (2001). *Psychological testing* (5th ed.). McGraw-Hill.

Nunnally, J. C., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*(6), 556–563. https://doi.org/10.1016/S0092-6566(02)00505-6

Paulhus, D. L., & Jones, D. N. (2011). A short measure of the Dark Triad. *Presented at meeting of the Society for Personality and Social Psychology*, Texas.

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

Revelle, W. (2021). How to use the psych package for mediation/moderation/regression analysis. *The Personality Project*. http://personality-project.org/r/psych/HowTo/mediation.pdf.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Seeboth, A., & Mõttus, R. (2018). Successful explanations start with accurate descriptions: Questionnaire items as personality markers for more accurate predictions. *European Journal of Personality*, *32*(3), 186–201. https://doi.org/10.1002/per.2147

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's Alpha. *Psychometrika*, *74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Sleep, C. E., Lynam, D. R., Widiger, T. A., Crowe, M. L., & Miller, J. D. (2019). Difficulties with the conceptualization and assessment of Criterion A in the DSM–5 alternative model of personality disorder: A reply to Morey (2019).

Vize, C. E., Collison, K. L., Miller, J. D., & Lynam, D. R. (2020). The "core" of the dark triad: A test of competing hypotheses. *Personality Disorders: Theory, Research, and Treatment*, *11*(2), 91.

Vize, C. E., Lynam, D. R., Collison, K. L., & Miller, J. D. (2018). Differences among dark triad components: A meta-analytic investigation. *Personality Disorders*, *9*(2), 101–111. https://doi.org/10.1037/per0000222

Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science*, *8*(4), 454–464. https://doi.org/10.1177/1948550617703168