Personality and Individual Differences 83 (2015) 185–191

Contents lists available at ScienceDirect

Personality and Individual Differences

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

A Mokken analysis of the Dark Triad 'Dirty Dozen': Sex and age differences in scale structures, and issues with individual items

Gregory Louis Carter [a,*], Anne C. Campbell [a], Steven Muncer [b], Katherine A. Carter [c]

[a] Psychology Department, University of Durham, Durham DH1 3LE, England, United Kingdom
[b] Programme in Clinical Psychology, University of Teesside, Middlesbrough TS1 3BA, England, United Kingdom
[c] Physics Department, University of Durham, Durham DH1 3LE, England, United Kingdom

ARTICLE INFO

Article history:
Received 13 October 2014
Received in revised form 6 April 2015
Accepted 7 April 2015
Available online 23 April 2015

Keywords:
Dark Triad
Narcissism
Machiavellianism
Psychopathy
Mokken analysis
Sex differences
Age differences

ABSTRACT

The Dark Triad (DT: Machiavellianism, narcissism, and subclinical psychopathy) have often been measured using a 12-item scale: The 'Dirty Dozen'. Many articles report participants' scale scores as well as their total score because structural models, based on classical test theory analysis, have indicated DT can be represented both as three correlated scales and a single scale. As DT are proposed to underlie a 'male' reproductive strategy of short-term, low-investment mating, sex differences have been of particular theoretical interest. Using two samples – one of student-aged participants; another comprised of a broader national sample – we applied Mokken analysis to investigate whether the same hierarchical structure existed across sex and age. For student women, the exclusion of one psychopathy item produced a single hierarchical DT scale. For student men, items formed a three-item narcissism scale and a six-item Machiavellianism–psychopathy scale. For non-student women and men, all twelve items constituted a unidimensional DT scale. Across all groups, item 'difficulty' was similar: Narcissism items were most easily endorsed and psychopathy items had the lowest rate of endorsement. Results are discussed in relation to the problematic empirical status of the Dirty Dozen psychopathy subscale, and in relation to sex and age differences.

1. Introduction

The Dark Triad of personality (DT: Machiavellianism; narcissism; subclinical psychopathy) has received considerable empirical attention since the concept appeared, just over a decade ago (Paulhus & Williams, 2002). The traits that comprise the Triad can each be measured with a separate inventory: For narcissism, this is the Narcissistic Personality Inventory (NPI, Raskin & Terry, 1988), which consists of 40 dyadic statements; for Machiavellianism, it is the 20-item Likert-scale Mach-IV (Christie & Geis, 1970), and for psychopathy, it is the Self-Report Psychopathy questionnaire, the most frequently-used version being the 31-item Likert-scale SRP-III (Paulhus, Neumann, & Hare, 2009). However, a total of 91 items across three measures (often in conjunction with other inventories) is burdensome to participants.

To address this, Jonason and Webster (2010) developed a 12-item inventory called the "Dirty Dozen" (DD). Correlations between DD subscales and original measures used to evaluate the three constructs ranged between $r = .34$ and $r = .47$. Internal consistency ($\alpha = .83$) and test–retest reliability ($r = .89$) were both high. There was also evidence of construct validity: Correlations between the DD and other inventories (e.g., measures of the Big 5) showed the predicted pattern of results. Subsequent research has supported and extended these findings (Jonason & McCain, 2012; Lee & Ashton, 2005). Since its development, the DD has been cited or used in peer-reviewed, DT-related papers more than 60 times; it is also the focus of the present study.

Whatever instruments are used, a key issue with the DT construct has been the extent to which the traits should be considered as three correlated scales, or as constituting a single scale (Furnham, Richards, Rangel, & Jones, 2014). Exploratory and confirmatory factor analyses have been used to examine this issue (Jonason, Li, & Buss, 2010; Jonason, Li, Webster, & Schmitt, 2009). Although confirmatory factor analyses used in the development of the DD concluded that a model specifying three correlated constructs fitted the data better than a single-factor model (Jonason & Webster, 2010), later analyses (Jonason, Kaufman, Webster, &

* Corresponding author.
E-mail addresses: g.l.k.carter@durham.ac.uk (G.L. Carter), a.c.campbell@durham.ac.uk (A.C. Campbell), s.muncer@tees.ac.uk (S. Muncer), k.a.carter@durham.ac.uk (K.A. Carter).

http://dx.doi.org/10.1016/j.paid.2015.04.012
0191-8869/

Geher, 2013; Jonason & Luévano, 2013) concluded that a bi-factor model (with items loading on both a general factor and three separate factors) showed the best fit to DD data. Structural equation modelling (SEM), conducted in relation to mate retention strategies and sociosexuality, indicated the former was best explained by a three-measure model, and the latter by a single-measure model (Jonason, Kavanagh, Webster, & Fitzgerald, 2011). Because of disagreement about the use and interpretation of multivariate models (Furnham et al., 2014), there is on-going debate as to whether high correlations between the traits constitute grounds for believing they represent a single latent construct. Consequently, many authors report both subscale and composite DD scores.

Increasingly, psychologists are moving beyond classical test theory (CTT) in evaluating psychometric measures. CTT is predicated on item correlations that test whether people respond similarly to items intended to measure the same trait. Most traits are normally distributed, and individuals endorse some items and not others (Watson, Deary, & Austin, 2007). Two individuals could therefore receive the same trait score despite having endorsed non-overlapping items. For example, in a test of arithmetic ability, someone who correctly answered '2 + 2 = ?' would receive the same score as someone who correctly answered '(234 − 56)/4 = ?'. Item response theory (IRT), however, examines items' structure by ordering them according to difficulty. It is based on the premise that an individual who achieves a high overall score would be more likely to get the latter question correct than someone who gets a lower overall score. This can also be applied to personality traits, to reveal hierarchical item structure.

Webster and Jonason (2013) used multidimensional IRT to evaluate the DD's item structure. Item discrimination (the degree to which an item can discriminate between people with the same level of the latent trait) was adequate, whilst analysis of item difficulty (the amount of the latent trait necessary to have a 50% chance of endorsing the item) was quite low, suggesting that the social undesirability of items created a high endorsement threshold. This was particularly true for psychopathy and Machiavellianism items. The possibility that men and women respond differently to DD items is pertinent because evolutionary psychologists have argued that DT underlies a male-typical strategy of short-term, low-investment mating (e.g., Jonason et al., 2009). The possibility that student-aged individuals and older, non-student adults may respond differently to DD items is important because most work on DT is, largely for convenience, conducted with student samples. It is therefore important to demonstrate that DD items function invariantly over sex and age because the validity of assertions regarding DD as a universal measure of DT depends upon this. Webster and Jonason (2013) examined differential scale functioning in relation to sex, and found that men had lower endorsement thresholds, especially for psychopathy. However, because item-level data were not examined, conclusions cannot be reached about whether specific items functioned differently in men and women. This is a key aim of the present study. (Note that item differential functioning is distinct from a sex difference: the former indicates an item has a different 'difficulty' in relation to total overall score in the two sexes.) Age differences have seldom been explicitly studied in this field, but significant sex differences in DT traits found in most student samples (e.g., Jonason et al., 2009) are not consistently replicated in samples of older participants (Carter, Campbell, & Muncer, 2014; James, Kavanagh, Jonason, Chonody, & Scrutton, 2014).

To explore these issues, we use Mokken analysis, a non-parametric form of IRT (Mokken, 1971; Molenaar, 1982). Although based on Guttman scaling, Mokken does not assume error-free data. Nor does it include assumptions about the sigmoid shape of item characteristic curves that can cause rejection of many items and so decrease the resultant measure's reliability. Two Mokken models have been outlined: The Monotone homogeneity model (MHM) and Double monotonicity model (DMM). These differ slightly in their requirements. Both require data to have unidimensionality (items assess the same latent trait), monotonicity (the probability of any given response is a non-decreasing function of that trait), and item independence (participants' response to any given item is not influenced by their response to other items). DMM additionally requires the non-intersection of items (such that item characteristic curves do not touch or overlap). Invariant item ordering (IIO) means that items can be ranked by difficulty (or endorsement frequency), allowing for hierarchical ordering. This requires the calculation of three coefficients. Coefficient H for each item provides a measure of scalability (and unidimensionality). From these values, an H coefficient for the full scale can be calculated, which indexes the extent to which items accurately order respondents. H^T reverses the roles of persons and items, and thus indexes the extent to which individuals agree on item ordering (Sijtsma, Meijer, & van der Ark, 2011). Together, H and H^T are indicative of scale strength and structure. Ultimately, if a DMM fits the data, and IIO can be established, it can be concluded that item ordering is robust across populations and subgroups (Sijtsma et al., 2011). Mokken analysis works by building a scale in a 'bottom-up' fashion from item-level data. When an item relationship is found that cannot be incorporated into the first extracted scale, the process iterates to determine the second (and further) scales present in the data. If the best solution to the data matrix is a three-scale structure, the Mokken program will identify these scales and constituent items.

As noted, the DD measure has chiefly been used on undergraduate samples. The present study considers two samples separately. The first ($N = 279$) consists of a student sample typical of existing work on DT ($M_{age} = 20.02$). The second ($N = 465$) is comprised of a national sample ($M_{age} = 35.37$), recruited via CrowdFlower, an internet platform that functions in a similar way to Amazon's "Mechanical Turk" system (Buhrmester, Kwang, & Gosling, 2011). Questionnaire responses were collected in standard Likert format. Mokken analysis was originally developed to deal with dichotomous (i.e., binary response) data of this kind, however, a model for polytomous data was subsequently introduced (Molenaar, 1982). We use these data to examine whether DD items constitute a single scalable dimension, whether scale structure varies between women and men, between student-aged and older samples, and whether item difficulty varies as a function of sex and/or age. We also examined narcissism, Machiavellianism, and psychopathy subscales for each sex.

2. Study 1: student sample

2.1. Participants

Two hundred and seventy-nine individuals (48.39% men), aged between 18 and 34 ($M = 20.04$, $SD = 2.17$), completed the Dirty Dozen. They were recruited as a convenience sample via a departmental participant pool (course credit was awarded for participation) at a UK university.

2.2. Measure

The Dirty Dozen (Jonason & Webster, 2010) is a 12-item measure of DT, consisting of three four-item subscales for narcissism, Machiavellianism, and psychopathy (see Table 1 for items). Respondents indicated the extent to which they agreed or disagreed with how well each statement reflected their own personalities on a five-point Likert scale (1 = Not at all like me;