Journal of Research in Personality 61 (2016) 87–98

Contents lists available at ScienceDirect

Journal of Research in Personality

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

Equivalence of Narcissistic Personality Inventory constructs and correlates across scoring approaches and response formats

Eunike Wetzel [a,b,*], Brent W. Roberts [a], R. Chris Fraley [a], Anna Brown [c]

[a] University of Illinois at Urbana-Champaign, United States
[b] University of Konstanz, Germany
[c] University of Kent, UK

ARTICLE INFO

Article history:
Received 13 May 2015
Revised 20 November 2015
Accepted 21 December 2015
Available online 29 January 2016

Keywords:
Narcissism
Narcissistic Personality Inventory
Forced-choice
Response format
Thurstonian item response model

ABSTRACT

The prevalent scoring practice for the Narcissistic Personality Inventory (NPI) ignores the forced-choice nature of the items. The aim of this study was to investigate whether findings based on NPI scores reported in previous research can be confirmed when the forced-choice nature of the NPI's original response format is appropriately modeled, and when NPI items are presented in different response formats (true/false or rating scale). The relationships between NPI facets and various criteria were robust across scoring approaches (mean score vs. model-based), but were only partly robust across response formats. In addition, the scoring approaches and response formats achieved equivalent measurements of the vanity facet and in part of the leadership facet, but differed with respect to the entitlement facet.

1. Introduction

Narcissism is characterized by inflated and grandiose self-views, feelings of superiority, a sense of entitlement, fantasies of unlimited power, success, or beauty, exhibitionism, and a lack of empathy (e.g., Cain, Pincus, & Ansell, 2008). A considerable amount of research effort over the past decades has been invested into understanding narcissism – both from clinical and personality psychology perspectives. As a result of this research effort, many findings have been reported regarding the relationships of narcissism with a variety of traits, such as the Big Five and self-esteem, and sociodemographic variables, such as gender and age. The validity of these findings depends on the psychometric soundness of the instruments used to measure narcissism. The most widely used instrument is the Narcissistic Personality Inventory (NPI; Raskin & Hall, 1979; Raskin & Terry, 1988). According to Cain et al. (2008), 77% of the research on narcissism in social and personality psychology relies on the NPI. The NPI consists of 40 item pairs that are presented in a forced-choice format. Participants are presented with item pairs and are instructed to endorse the response option that is closest to their feelings and beliefs. Each item pair in the NPI consists of one narcissistic response option (Option A in the example) and one non-narcissistic response option (Option B in the example).

	Most like me
Example for the forced-choice format	
Option A I have a natural talent for influencing people	☐
Option B I am not good at influencing people	☐

Despite the popularity of the NPI, its psychometric integrity as a measure of narcissism has been questioned, especially with respect to its factor structure (e.g., Ackerman et al., 2011). Furthermore, the predominant procedure for scoring the responses to NPI items is to count the number of narcissistic response options a respondent endorsed, thereby disregarding the forced-choice nature of the items. It has been shown that the forced-choice format violates the assumption of independence (i.e., the options in the forced-choice pair are not independent; Brown & Maydeu-Olivares, 2011; Meade, 2004). This raises the question of whether the findings reported on narcissism, which are based mainly on total scores from the NPI, are robust to the problematic scoring procedure. The present research investigates whether this scoring procedure results in biased estimates of correlations with external variables. The present investigation uses data from several studies, including an online experiment in which response formats for the NPI are systematically varied and then related to external criteria that have been linked to narcissism in past research.

* Corresponding author at: University of Konstanz, Department of Psychology, Box 31, 78457 Konstanz, Germany.
E-mail address: eunike.wetzel@uni-konstanz.de (E. Wetzel).

http://dx.doi.org/10.1016/j.jrp.2015.12.002
0092-6566/

We first summarize some of the important findings from research on narcissism based on NPI scores. Second, we describe psychometric issues related to the NPI in detail. Third, we describe the study we conducted to investigate whether previous findings on narcissism are confirmed when the forced-choice format is modeled appropriately and when the response format is varied. Finally, we report the results of these analyses and discuss their implications for the use of the NPI in psychological research.

2. Findings on narcissism as a personality trait

Narcissism has fascinated researchers, in part, because of its complex nature. Narcissism can be salubrious or deleterious. For example, trait narcissism as assessed by the NPI is positively related to extraversion and emotional stability, but negatively related to agreeableness (Ackerman et al., 2011; Emmons, 1984; Rhodewalt & Morf, 1995; Trzesniewski, Donnellan, & Robins, 2008). Furthermore, people higher on narcissism also tend to report higher self-esteem (Rhodewalt & Morf, 1995; Trzesniewski et al., 2008). However, studies that distinguish between the adaptive (e.g., grandiosity, leadership, vanity) and maladaptive (entitlement, exploitativeness) components of narcissism find that these two components show differential relationships to other traits. For example, neuroticism is positively related to the entitlement/exploitativeness facet of the NPI, but negatively or not related to adaptive NPI facets (Ackerman et al., 2011; Emmons, 1984). Extraversion shows the opposite pattern: positive correlations with adaptive narcissism and no correlations with maladaptive narcissism (Ackerman et al., 2011; Emmons, 1984). Both components of narcissism are negatively associated with agreeableness (Ackerman et al., 2011; Rhodewalt & Morf, 1995).

Narcissism has also been studied in relation to sociodemographic variables. Several studies have reported that NPI scores decline with age (Foster, Campbell, & Twenge, 2003; Hill & Roberts, 2012; Roberts, Edmonds, & Grijalva, 2010). A recent meta-analysis on gender differences in narcissism found that men tend to report higher levels of narcissism than women (overall $d = .25$; Grijalva et al., 2015). Other research has linked narcissism to higher socioeconomic status (Piff, 2014).

Thus, research on narcissism as assessed by the NPI has revealed relationships between narcissism and other personality traits and external variables. These findings have been important for a number of reasons. Understanding the association between NPI scores and personality traits, for example, has been crucial for illuminating both the adaptive and maladaptive sides of narcissism. Considering that these findings use, as their foundation, an instrument that has potentially questionable scoring practices, it seems important to verify the validity of these findings.

3. Psychometric issues with the NPI

3.1. Dimensionality and factorial structure

The validity of the external correlations of the NPI rests on the assumption that the scale is both reliable and valid. Unfortunately, the NPI has a somewhat inconsistent record regarding its factor structure. The most persistent inconsistency of the NPI is the varying number of factors reported in exploratory factor analyses of the measure. The original study by Raskin and Terry (1988), from which the 40-item version in use today originated, identified seven subscales (authority, exhibitionism, superiority, vanity, exploitativeness, entitlement, self-sufficiency). Other studies found fewer factors. The studies by Corry, Merritt, Mrug, and Pamp (2008) and Kubarych, Deary, and Austin (2004) reported two factors (leadership/authority, exhibitionism/entitlement in Corry et al. and power, exhibitionism in Kubarych et al.), although Kubarych et al. (2004) suggested that a third factor (being a special person) might exist. Ackerman et al. (2011) also identified three factors (leadership/authority, grandiose exhibitionism, entitlement/exploitativeness) while Emmons (1984) identified four factors (leadership/authority, superiority/arrogance, self-absorption/self-admiration, exploitativeness/entitlement). Leadership/authority factors tend to be measured by a larger number of items and are related to more adaptive traits and outcomes such as high self-esteem and extraversion. In contrast, exhibitionism/entitlement/exploitativeness factors tend to be measured by fewer items and are related to rather maladaptive traits and outcomes such as high neuroticism and low relationship quality (Ackerman et al., 2011).

Of particular importance to the present study, Ackerman, Donnellan, Roberts, and Fraley (2015) investigated the impact of changing the response format from the original forced-choice format to a dichotomous true/false or polytomous rating scale format on the resulting factor structure. They found that the factor solutions differed across response formats with three factors (leadership, vanity, exhibitionism) being sufficient in the forced-choice format whereas two additional factors (manipulativeness, superiority) were found in the true/false and rating scale format. Furthermore, Ackerman et al. (2015) found that several item pairs assessing manipulativeness and superiority consisted of statements that did not reflect the same trait (i.e., unidimensional forced-choice) but rather different traits (i.e., multidimensional forced-choice). To summarize, despite finding differing factor structures, previous research is consistent with two conclusions: (1) the NPI is not a unidimensional scale and (2) items describing adaptive content (e.g., leadership, authority, vanity) are more prevalent than items describing maladaptive content (e.g., exploitativeness, entitlement).

3.2. Scoring of the NPI

Another important psychometric issue that appears to have been neglected in previous research is related to the scoring of the NPI items. In most applications the number of narcissistic responses endorsed by a participant are counted to form the NPI total score. This scoring practice essentially treats responses to the NPI's forced-choice items as responses to single-stimulus items where each item is rated separately. The forced-choice nature of the NPI items is ignored. For unidimensional item pairs, where the two response options reflect different levels of the same trait, this might not distort the validity of the scores. This is because the latent response tendency for such an item pair is simply the difference of the item utilities, which represent the similarity between the behavior described in the item and the respondent's own behavior (Maydeu-Olivares & Brown, 2010). With the utilities of items i and k described by the linear factor analysis model, the latent response tendency has a simple form:

$$y^*_{ik} = (\text{mean}_i - \text{mean}_k) + (\text{loading}_i - \text{loading}_k) \cdot \text{trait} + (\text{error}_i - \text{error}_k). \quad (1)$$

Assuming that the factor loading for the positively keyed narcissism item i is positive, and the factor loading for the negatively keyed narcissism item k is negative, the difference of the two factor loadings is positive. Therefore, selecting the positively keyed narcissism item will contribute positively to the measurement of the trait (narcissism). Assigning the score 1 in this case reflects the judgment for the whole pair, not for an individual item. Summing up such binary scores as in the classical scoring approach to derive the total score is an acceptable simplification that in most cases does not distort correlations with external variables (McDonald, 1999).

However, a recent study suggests that some item pairs in the NPI may be multidimensional (Ackerman et al., 2015), in which case partially ipsative scores are obtained (Brown & Maydeu-Olivares, 2013). As the name suggests, partially ipsative scores impose a partial constraint on the total test score which may result in distorted correlations with other variables. Because the relationships between NPI scores and other variables have only been investigated with the scores assuming unidimensionality in item pairs, it is unclear whether distortions may have taken place.

One obvious solution to the partially ipsative data of the NPI is to model the responses using appropriate forced-choice models, which can incorporate both the unidimensional and multidimensional item pairs. This will be referred to as the *model-based* scoring approach in the following sections. Another solution to the partially ipsative data of the NPI is to use alternative response formats. For example, other narcissism measures such as the Pathological Narcissism Inventory (Pincus et al., 2009) or the Narcissistic Admiration and Rivalry Questionnaire (Back et al., 2013) use Likert rating scales rather than paired comparisons. It is possible that the psychometric properties of the NPI would be clarified if the 40 couplets were rated as individual items. In fact, multiple studies have applied the NPI in a single-stimulus format (Bogart, Benotsch, & Pavlovic, 2004; Boldero, Bell, & Davies, 2015; Brown & Zeigler-Hill, 2004; Egan & Lewis, 2011; Gerbasi & Prentice, 2013; Jordan, Spencer, Zanna, Hoshino-Browne, & Correll, 2003; Lee, Gregg, & Park, 2013; Morf & Rhodewalt, 1993), although these investigations did not check whether equivalent constructs were measured across formats. Also, in most cases, investigators only presented the narcissistic response options to participants. The present study will investigate the effects of both of these approaches on the relationships between NPI scores and other variables.

4. Aim of this study

The aim of this study is to investigate (1) whether the same constructs are measured across (a) scoring approaches and (b) response formats and (2) whether external correlates are affected by different (a) scoring approaches and (b) response formats. Research question 1a addresses whether the same constructs are measured in the classical (mean or sum score) scoring approach, which ignores the forced-choice nature of the items, and the model-based scoring approach, which takes the forced-choice nature of the items into account. Research question 1b addresses whether the same constructs are measured when the items are presented using different response formats. Three NPI versions based on different response formats are compared: (1) the original forced-choice response format, (2) a dichotomous true/false response format, and (3) a polytomous rating scale response format.

Research question 2a is concerned with whether external correlates of NPI constructs with criteria and other personality traits differ between the classical and model-based scoring approach. This will allow us to test whether findings based on NPI scores that have been reported in previous research can be confirmed when the forced-choice nature of the NPI's forced-choice response format is taken into account. Research question 2b investigates whether external correlates of NPI constructs differ across response formats (forced-choice, true/false, and rating scale). Both research questions will be addressed at the overall scale level and at the facet level.

5. Method

5.1. Samples

We analyzed data from two samples. For the first sample (*between sample*) the data collection was based on a between-subjects design and participants were randomly assigned to one of the three response formats. In the second sample (*within sample*) participants took both the forced-choice and rating scale versions of the NPI.

5.1.1. Between sample

The between sample consisted of $N = 17{,}434$ participants who took part in an online survey. After responding to several demographic questions, participants were randomly assigned to one of the three response format conditions and filled out that version of the NPI. The first response format condition was the original forced-choice format with pairwise comparisons (see example given in the introduction). For the other two response format conditions the 80 statements in the NPI were presented separately. In the second condition respondents rated whether each statement was true or false for them.

		True	False
Example for the true/false response format			
Item A	I have a natural talent for influencing people	☐	☐
Item B	I am not good at influencing people	☐	☐

In the third response format condition each statement was rated on a five-point rating scale ranging from *strongly disagree* to *strongly agree*.

		Strongly disagree	Dis-agree	Neither disagree nor agree	Agree	Strongly agree
Example for the rating scale response format						
Item A	I have a natural talent for influencing people	☐	☐	☐	☐	☐
Item B	I am not good at influencing people	☐	☐	☐	☐	☐

The sample sizes for each response format condition, descriptive statistics, and model-based reliability estimates (coefficient ω; McDonald, 1999) are presented in Table 1. After completing the questionnaire, participants received feedback on their NPI scores. Note that this sample subsumes the sample of $N = 7185$ participants analyzed in Ackerman et al. (2015) since the data came from the same online survey that continued collecting data after the data for the Ackerman et al. study had been extracted.

5.1.2. Within sample

Participants in the within sample were $N = 1246$ persons who took part in the Project Talent Personality Inventory (PTPI) validation study (Pozzebon et al., 2013). The sampling procedure for the validation study was designed to oversample young adults (age 20s) and old adults (age 60s) to resemble the ages of the original Project Talent sample at wave 1 (1960) and today. Table 1 shows descriptive statistics for the within sample. The within sample had a slightly lower percentage of females (66%) compared to the between sample (70%). Furthermore, the mean age of participants in the within sample was higher than the mean age of participants in the between sample (55.07 vs. 30.01). Besides the NPI,

Table 1
Descriptive statistics for the between sample and the within sample.

	N	Age	%	Leadership			Vanity			Entitlement		
		M (SD)	W	*M* (SD)	POMP	ω	*M* (SD)	POMP	ω	*M* (SD)	POMP	ω
Between sample												
Forced-choice	6690	29.30 (11.36)	70	0.44 (0.22)	44	.89	0.35 (0.24)	35	.87	0.28 (0.26)	28	.52
True/false	5510	30.24 (12.05)	71	0.43 (0.17)	43	.93	0.37 (0.20)	37	.91	0.33 (0.18)	33	.46
Rating scale	5234	30.69 (12.18)	70	1.94 (0.41)	48.5	.91	1.76 (0.55)	44	.89	1.65 (0.41)	41.25	.44
Within sample												
Forced-choice	1246	55.07 (17.33)	66	0.31 (0.23)	31	.93	0.19 (0.20)	19	.92	0.13 (0.20)	13	.76
Rating scale				1.76 (0.44)	44	.92	1.32 (0.54)	33	.89	1.43 (0.43)	35.75	.64

Note. W = women, POMP = percent of maximum possible, ω = model-based reliability coefficients. POMP scores are linear transformations from the original metric into percentages with range 0–100. For obtaining ω negative factor loadings in the true/false and rating scale data were reversed. Negative factor loadings occurred because the factor structure of the forced-choice format was imposed on the other response formats. For both samples only participants with ages between 18 and 75 were included in the analyses. The theoretical midpoint for the forced-choice and the true/false formats is 0.5. The theoretical midpoint for the rating scale format is 2.0 (range 0–4).

participants also filled out demographic information, the PTPI, the Satisfaction with Life Scale (Diener, Emmons, Larsen, & Griffin, 1985), as well as a few short scales assessing traits such as gratitude. Participants received remuneration in the form of token points (Zoompoints) that they could spend in the online system of Zoomerang, the online survey company that recruited the participants for this study.

5.2. Instruments

5.2.1. Demographic information

Participants in both samples provided information on their age, gender, and education level. In the between sample participants additionally rated their socioeconomic status (SES) on a ten-point rating scale depicted as a ladder. The instruction was "Please click the step corresponding to the position on the ladder where you think you stand at this time in your life, compared to people in your country." In the within sample participants additionally responded to a number of other demographic questions including monthly household income and ethnicity (see Pozzebon et al., 2013).

5.2.2. Narcissistic Personality Inventory

The 40-item NPI (Raskin & Terry, 1988; description see above) was administered in three versions that differed with respect to the response format: forced-choice, true/false, and rating scale with five response options (see examples above).

5.2.3. Project Talent Personality Inventory

The PTPI assesses ten personality traits that are relevant to normal high-school student populations: vigor, calmness, mature personality, impulsiveness, self-confidence, culture, sociability, leadership, social sensitivity, and tidiness. Participants rated how well the 108 statements applied to them on a scale ranging from *not very well* (1) to *extremely well* (5). Descriptions of the constructs can be found in Pozzebon et al. (2013). The PTPI was only filled out by the within sample.

5.2.4. Satisfaction with Life Scale

Participants in the within sample further filled out the Satisfaction with Life Scale (SWLS; Diener et al., 1985), a five-item measure assessing general life satisfaction. An example for an item is "I am satisfied with my life." Responses were given on a seven-point rating scale ranging from *strongly disagree* (1) to *strongly agree* (7).

6. Analyses

First, pre-analyses were conducted using exploratory structural equation modeling (Asparouhov & Muthen, 2009) in order to find the adequate factor structure for the NPI in our samples and in order to derive scores at the facet level. Following a classical scoring approach, we computed a mean score for narcissism on the overall scale level and mean scores for each facet confirmed in the factor analyses. Furthermore, following a model-based approach, we modeled overall narcissism and the NPI facets as latent traits. Research question 1, asking whether the same constructs are measured across scoring approaches and response formats, was addressed by comparing correlations of the same constructs across scoring methods and response formats. Research question 2, asking whether external correlates are affected by different scoring approaches and response formats, was investigated by comparing correlations with criteria and traits across scoring approaches and across response formats. The components of our analyses are described in more detail in the following section.

6.1. Pre-analyses on NPI factor structure

It is important to note that due to the classical scoring scheme of the NPI, previous research only factor analyzed the items representing narcissistic responses. This is only appropriate for unidimensional item pairs where the score of 1 or 0 reflects the judgment for the whole item pair. In our study both response options in the item pair will be taken into account in the factor analysis of the forced-choice format. In addition, since we also presented items separately in single-stimulus response formats, the factor structure of *all* individual response options was investigated. The factor structure of the NPI was investigated using exploratory structural equation modeling (ESEM; Asparouhov & Muthen, 2009) in Mplus (Version 7.11; Muthén & Muthén, 1998–2014). ESEM models with one to six factors were estimated for each of the response formats using the unweighted least squares (ULS) method with mean and variance-corrected Satorra–Bentler (Satorra & Bentler, 1994) goodness-of-fit tests (denoted ULSMV estimator in Mplus). For the forced-choice format an exploratory version of the Thurstonian item response model was applied (Brown & Maydeu-Olivares, 2011, 2013). An annotated Mplus syntax for this model can be found in supplemental material S1. The Geomin oblique rotation method was applied in all models. Fig. 1 illustrates the ESEM model with three factors. As shown in Fig. 1, the factor loadings of all items on all factors are estimated (e.g., λ_{11}, λ_{12}, λ_{13} for the factor loadings of item 1). Pre-analyses of the factor structure were conducted on the between sample because it was larger and more heterogeneous than the within sample and data for all three response formats were available.

Several criteria were applied to evaluate the factor models: (1) the conceptual clarity and interpretability of the factors, (2) the similarity of the factors across response formats, and (3) goodness of fit. Conceptual clarity and interpretability refers to whether the

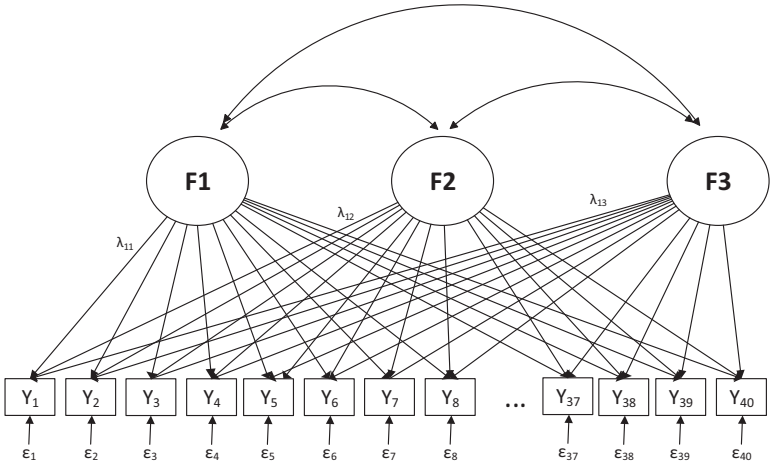

Fig. 1. Exploratory structural equation model with three factors. For clarity of presentation, only the paths for the first item are labeled. For the forced-choice response format there are 40 items (as depicted), for the true/false and rating scale response formats there are 80 items.

factors could unambiguously be identified as personality traits based on the content of the items that loaded on them. Model fit was evaluated using the root mean square error of approximation (RMSEA; Steiger, 1990), comparative fit index (CFI; Bentler, 1990), and Tucker–Lewis index (TLI; Tucker & Lewis, 1973). For the RMSEA, values below .08 indicate reasonable and values below .05 indicate close fit (Browne & Cudeck, 1993). For the CFI and TLI, values above .90 (.95) indicate acceptable (good) model fit (Hu & Bentler, 1999). Items were allocated to a factor if at least 12% of their variance was explained by the respective factor. The cut-off of 12% was chosen because it can be considered a substantial and meaningful amount of variance explanation, indicating that the item measures the factor. This cut-off corresponds to a threshold of .25 for the factor loadings in the forced-choice model and a threshold of .35 for the factor loadings in the true/false and rating scale models.[1] The threshold of .35 was also used in previous analyses of the NPI's factor structure (Emmons, 1984; Kubarych et al., 2004).

6.2. Mean scores and latent trait correlations

After the factor structure was established, mean scores for the overall narcissism scale and for the facets resulting from the pre-analyses were computed. Mean scores in the forced-choice format are based on the current classical scoring practice of summing the narcissistic response options that a participant endorsed. Average mean scores and percent of maximum possible (POMP) scores (Cohen, Cohen, Aiken, & West, 1999) for the two samples on overall narcissism and the facets are reported in Table 1. Correlations based on mean scores will be referred to as observed correlations.

We also applied a model-based approach to model overall narcissism and the facets as latent traits. This approach allows us to obtain estimates of the correlations between latent variables that are not attenuated by unreliability (latent correlations; Adams, Wilson, & Wang, 1997). The latent correlations were estimated based on the respective response format model. For the true/false data the dichotomous two-parameter logistic model was applied. For the rating scale data the graded response model (Samejima, 1969) was applied. These two models linking categorical item responses to continuous latent traits are explained in detail in standard textbooks on item response theory (e.g., De Ayala, 2009; Embretson & Reise, 2000). For the forced-choice data, the Thurstonian item response model (Brown & Maydeu-Olivares, 2011, 2013) was applied. In contrast to the classical scoring approach, this model takes the forced-choice nature of the items into account (i.e., the dependencies between items within pairs are modeled correctly). Supplemental material S2 shows a sample Mplus syntax for the application of the Thurstonian item response model to the NPI data (see also Brown & Maydeu-Olivares, 2012, for a tutorial on fitting the Thurstonian item response model). The three item response models all allow items to differ regarding their difficulty (i.e., the probability of endorsing an item) and their discrimination (i.e., their ability of differentiating between different trait levels). Thus, the models share a common modeling framework and only differ with respect to the specific type of response format data they accommodate.

6.3. Analyses to compare scoring approaches

Analyses to compare the classical and the model-based scoring approach were based on correlations. To investigate whether the same constructs were measured by both scoring approaches (research question 1a), we obtained latent correlations between latent trait overall narcissism and mean score overall narcissism. Thus, mean score overall narcissism was added to the Thurstonian item response model as the observed covariate and its correlation with the latent trait narcissism was estimated. In this case unreliability in measurement is accounted for on the side of the latent trait, but not for the observed mean score. The same analyses were also conducted at the facet level.

To investigate whether external correlates differed between the classical and model-based scoring approaches to the forced-choice format (research question 2a), we correlated overall narcissism and the NPI facets with the criteria age, sex, SES, and education level for the between sample and with the PTPI personality traits and satisfaction with life for the within sample. Correlations with the criteria and other traits were compared between observed correlations based on mean scores (classical scoring approach) and latent correlations (model-based scoring approach).

6.4. Analyses to compare response formats

To investigate whether different response formats in collecting responses to the NPI items measured the same underlying

[1] The cut-off loading in the forced-choice model is different from the one in the single-stimulus models because the loadings are standardized with respect to item pairs, not individual items. The cut-off of .25 in the forced-choice model and the cut-off of .35 in the single-stimulus models correspond to the same unstandardized loadings.

constructs (research question 1b), we obtained latent correlations between NPI constructs in the forced-choice format and NPI constructs in the rating scale format for the within sample. The true/false format could not be included in these analyses because the within sample only filled out the forced-choice and rating scale versions of the NPI. Lastly, to investigate whether correlations of NPI constructs with criteria and other traits differed across the three response formats (research question 2b), we compared latent correlations with the criteria age, sex, SES, and education level across the three response formats for the between sample. Furthermore, for the within sample, latent correlations with the PTPI personality traits and satisfaction with life were compared between the forced-choice and rating scale formats. The strength of correlations will be evaluated according to the classification by Cohen (1988): .10 is considered a weak correlation, .30 a moderate correlation, and .50 a strong correlation.

7. Results

In the following, the results of the pre-analyses on the NPI's factor structure will be described. Then, the results of the analyses investigating whether the same constructs are measured across scoring approaches and response formats (research question 1) and whether external correlates are affected by different scoring approaches or response formats (research question 2) will be reported for the overall scale level and for the facet level.

7.1. NPI factor structure

The evaluation of the competing ESEM models according to the criteria of conceptual clarity, comparability across response formats, and ESEM model fit showed that the three-factor exploratory model was the most adequate to describe the factor structure of the NPI across the three response formats. According to the RMSEA, model fit was good in the forced-choice format (RMSEA = 0.038, 90% CI [0.037, 0.038], the true/false format (RMSEA = 0.033, 90% CI [0.033, 0.034]), and the rating scale format (RMSEA = 0.056, 90% CI [0.055, 0.056]). According to CFI and TLI, goodness of fit was just below acceptable for the forced-choice format (CFI = 0.90; TLI = 0.88), and less than acceptable for both the true/false format (CFI = 0.80, TLI = 0.79) and the rating scale format (CFI = 0.74, TLI = 0.72).[2] The fit indices for ESEM models with one to six factors are available in supplemental Table S3.

The three factors were defined as *leadership*, *vanity*, and *entitlement*. According to the factor loadings in the forced-choice format, 19 item pairs indicated leadership, 12 indicated vanity, and 4 indicated entitlement. The five remaining item pairs did not show any loading greater than .25 in the forced-choice format. Leadership refers to respondents' tendency to assume leadership positions, see themselves as good leaders, and to gain power for power's sake alone. An example pair measuring leadership is "A: I am not sure if I would make a good leader. – B: I see myself as a good leader." Vanity refers to respondents' tendency to take excessive pride in their own appearance or achievements as for example assessed by the pair "A: I like to look at myself in the mirror. – B: I am not particularly interested in looking at myself in the mirror." Entitlement refers to participants' tendency to feel that they have the right to something (e.g., praise, recognition, favorable treatment). An example for an item pair assessing entitlement is "A: I will never be satisfied until I get all that I deserve. – B: I take my satisfactions as they come." The first two factors correspond well with those reported in Ackerman et al. (2015; also called leadership and vanity) and Ackerman et al. (2011; labeled leadership/authority and grandiose exhibitionism). In contrast, our third factor mainly comprised items with content related to feelings of entitlement while the third factor in Ackerman et al. (2015) comprised exhibitionism items and the third factor in Ackerman et al. (2011) comprised items addressing entitlement and exploitativeness.

Supplemental Table S4 shows a comparison of the factor structures between Ackerman et al. (2015) and this study. It is important to note that Ackerman et al. and our study pursued different goals. Ackerman et al. explicitly investigated what the best-fitting factor structure was for each response format. In contrast, the goal of our factor analyses was to find the factor structure that was most comparable across response formats in order to base the following analyses on a common framework of facets that had the same meaning across response formats. Thus, due to the differing goals, it is not surprising that the factor structures differed slightly between the two studies. Nevertheless, for the forced-choice format, the allocation of items to the leadership and vanity facets corresponded very well. Some of the items on Ackerman et al.'s exhibitionism facet were part of leadership or vanity in our study, leaving only four items that we found are best characterized by the term entitlement (see Table S4).

Observed correlations between mean scores based on our factor structure and mean scores based on Ackerman et al.'s factor structure substantiate the similarity of the response formats (see supplemental Table S5). The leadership and vanity facets showed a large amount of overlap for all response formats (correlations between .77 and .95). Our entitlement facet correlated strongly with Ackerman et al.'s exhibitionism facet in the forced-choice format (r = .73) and the true/false format (r = .50). For the rating scale format, our entitlement facet correlated moderately to strongly with Ackerman et al.'s exhibitionism (r = .36) and manipulativeness (r = .49) facets.

Table 2 contrasts the unstandardized factor loadings for 12 item pairs on vanity for the three response formats. Unstandardized factor loadings are comparable across response formats because residual variances in the ESEM for forced-choice data were adjusted to account for the presentation of items as pairs (see syntax in supplemental material S1). The complete set of factor loadings for all NPI items on the three facets is available in supplemental Table S6. First, it can be seen in Table 2 that the factor loadings for the two items in the true/false and rating scale formats are opposite in sign, as expected. It can also be seen that the factor loadings are largely similar for the two single-stimulus formats. Since the difference in factor loadings of the two items is estimated in the unidimensional forced-choice format (see Eq. (1)), only one factor loading is printed; whereas for the single-stimulus formats (true/false and rating scale), both items in the item pair have separate factor loadings. For most item pairs, the factor loadings for single-stimulus items are very similar to the forced-choice loading (e.g., −.80 for forced-choice, −.83 for true-false, and −.63 for rating scale for item 15A "I don't particularly like to show off my body.").

However, there are also a few item pairs for which the forced-choice and the single-stimulus formats differ with respect to which of the factors the items load highest on. For example, for the two item pairs 9 and 20 in Table 2, the highest loading for the forced-choice format is on vanity whereas for true/false the highest loading for the first item in the pair is on entitlement. Out of the 40 NPI item pairs, there are 7 item pairs that would be allocated to different facets according to their factor loadings in the forced-choice, true/false, and rating scale formats.

Furthermore, there are a few item pairs that are multidimensional (i.e., the two items in the pair measure different traits). Multidimensionality is indicated by the two items in a pair showing their highest loading on different facets of narcissism. For instance,

[2] Note that the RMSEA is more informative than CFI and TLI for our models because it is sample size independent (Meade, Johnson, & Braddy, 2008).

Table 2
Unstandardized factor loadings for vanity across three response formats.

Item	Item content	Factor loading		
		Forced-choice	True/false	Rating scale
4A	When people compliment me I sometimes get embarrassed.	−0.60	−0.41	−0.41
4B	I know that I am good because everybody keeps telling me so.		0.26	0.09
7A	I prefer to blend in with the crowd.	−0.77	−0.36	−0.42
7B	I like to be the center of attention.		0.81	0.86
9A	I am no better or worse than most people.	−0.42	−0.07	−0.28
9B	I think I am a special person.		0.52	0.18
15A	I don't particularly like to show off my body.	−0.80	−0.83	−0.63
15B	I like to display my body.		0.99	0.73
19A	My body is nothing special.	−1.10	−0.69	−0.38
19B	I like to look at my body.		1.00	0.58
20A	I try not to be a show off.	−0.53	−0.44	−0.73
20B	I am apt to show off if I get the chance.		0.65	0.70
26A	Compliments embarrass me.	−0.75	−0.50	−0.41
26B	I like to be complimented.		0.67	0.40
28A	I don't care about new fads and fashions.	−0.61	−0.54	−0.38
28B	I like to start new fads and fashions.		0.54	0.45
29A	I like to look at myself in the mirror.	1.09	1.08	0.68
29B	I am not particularly interested in looking at myself in the mirror.		−0.95	−0.60
30A	I really like to be the center of attention.	0.71	0.79	0.88
30B	It makes me uncomfortable to be the center of attention.		−0.56	−0.62
37A	I wish somebody would someday write my biography.	0.39	0.33	0.36
37B	I don't like people to pry into my life for any reason.		−0.22	−0.16
38A	I get upset when people don't notice how I look when I go out in public.	0.63	0.57	0.65
38B	I don't mind blending into the crowd when I go out in public.		−0.31	−0.50

Note. Only items with standardized factor loadings >.25 in the forced-choice format are depicted. The unstandardized factor loadings are shown in this table because only they are directly comparable across the three response formats. For the forced-choice and rating scale format, the sign was switched in order to make items reflect vanity rather than lack of vanity. $N = 6690$ for forced-choice, $N = 5510$ for true/false, and $N = 5234$ for rating scale.

item39A ("I am more capable than other people.") shows moderate positive factor loadings on leadership for all response formats (.43 forced-choice, .57 true/false, and .51 rating scale) whereas item 39B ("There is a lot that I can learn from other people.") has low loadings on leadership (.04 for true/false and .13 for rating scale). Instead, item 39B loads about −.71 on entitlement for true/false and -.46 for rating scale, indicating that it measures entitlement rather than leadership. Over all NPI item pairs, there are 12 item pairs where the first item in the pair shows the highest loading on a different facet than the second item in the pair for the true/false format. For the rating scale format the number of multidimensional item pairs is 13. With the Thurstonian item response model, this multidimensionality can be modeled in the analysis of forced-choice NPI data (see below).

Multidimensionality can also be determined based on the correlations of the items in one item pair. Ackerman et al. (2015) defined multidimensional item pairs as those in which the two response options in one pair correlated equal to or below |.30| in the true/false and rating scale data. This method led to 19 multidimensional item pairs. Applying Ackerman et al.'s criterion to our between sample which subsumes the sample Ackerman et al. analyzed yielded 20 item pairs with multidimensionality (item 16 in addition to the ones listed in Ackerman et al., 2015). Thus, both criteria for multidimensionality indicate that a substantial number (between 30% and 50%) of the item pairs in the forced-choice NPI consist of items measuring different traits.

7.2. Results regarding the equivalence of constructs and external correlates across scoring approaches and response formats

In the following, we will report our results for the two research questions. For the facet-level analyses, the results reported in this section are all based on the factor structure for the forced-choice format, i.e. the items allocated to the three factors were identical across response formats. The discussion will address how the results would change if the "correct" factor structure were used for each response format (i.e., allocating items to factors according to factor loadings in the rating scale format for rating scale, and according to factor loadings in the true/false format for the true/false format).

7.2.1. Research question 1a: Are the same constructs measured across scoring approaches?

To evaluate whether the same constructs were being measured by the classical scoring approach based on summing item responses and by the model-based approach taking dependencies between items in each pair into account, we estimated the correlation between the two within the Thurstonian item response model. At the level of overall narcissism, the mean overall narcissism score and latent trait narcissism correlated at .99 in the between sample (.96 in the within sample) indicating that the two corresponded almost perfectly.

At the facet level, leadership modeled as a latent trait correlated almost perfectly with the observed mean score on leadership ($r = .98$). The same was the case for vanity with a correlation of .96. Thus, for these two facets both types of scoring procedures provided equivalent estimates of participants' latent trait levels. In contrast, latent trait entitlement and mean score entitlement correlated at only .79. This indicates that for entitlement, it made a difference whether the dependencies between items were taken into account or not in the scoring of the items and estimates of participants' trait levels may therefore differ between scoring procedures.

7.2.2. Research question 1b: Are the same constructs measured across response formats?

In order to investigate whether the same construct of overall narcissism was measured in both the forced-choice and the rating scale format, we obtained the latent correlation between the two overall narcissism latent traits in a combined Thurstonian and graded response model. It was .90, indicating that there was substantial overlap between overall narcissism as measured in the forced-choice format and overall narcissism as measured in the rating scale format, but that the match was not perfect.

Next, we tested whether the same facet-level constructs were measured across response formats by obtaining the latent correlations between all NPI facets from the forced-choice data and all NPI facets from the rating scale data for the within sample. The resulting correlation matrix is depicted in Table 3. The latent correlation for leadership in the forced-choice format with leadership in the rating scale format was .91. The latent correlation for vanity assessed with different response formats was even higher at .94. These correlations were higher than the correlations between traits within one response format of which the highest one was .85 between leadership and entitlement in the rating scale format. This indicates that the forced-choice and rating scale formats appear to provide near equivalent measurements of leadership and vanity.

For entitlement, the latent correlation between forced-choice and rating scale was .69, which was lower than several correlations within one response format. This shows that the forced-choice and

rating scale versions of the NPI did not provide equivalent measurements of the entitlement facet.

7.2.3. Research question 2a: Are external correlates affected by different scoring approaches?

To test whether external correlates of overall narcissism differed between the classical and model-based scoring approach, we compared the observed mean score correlations with the latent model-based correlations. Correlations of overall narcissism with the criteria age, sex, SES, and education level are depicted in Table 4. The NPI mean score was negatively related to age and positively related to SES. Men reported higher overall narcissism levels than women. All correlations were small (absolute values between .01 and .14). The NPI mean score was not related to education level. Latent correlations from a Thurstonian item response model with the overall narcissism dimension were very similar to the observed correlations except that the correlation with SES appeared higher (.22), although the difference between the two correlations was not significant. In general, latent correlations can be expected to be slightly higher than correlations based on mean scores since they are not attenuated by measurement error.

At the observed level, the NPI mean score was strongly related to PTPI-leadership (r = .59) and vigor (r = .36). Latent correlations between overall narcissism and the PTPI traits were overall similar, but higher in some cases such as leadership with a correlation of .71 (see Table 5). The mean absolute difference between observed and latent correlations across the 11 traits was 0.06, indicating similar relationships across scoring techniques.

The same analyses were conducted at the facet level. Latent correlations were obtained from two versions of the Thurstonian item response model. The first assumes all item pairs to be unidimensional as in the classical scoring approach. The second takes multidimensionality in item pairs into account by allowing items within the item pair to load on different facets if the loadings from the single-stimulus formats indicated multidimensionality.

The facet level correlations reveal a differentiated picture of the relationships with the criteria (see Table 4). For example, consistently across the two scoring approaches leadership did not show a correlation with age. Instead, the negative correlation found for overall narcissism was mainly driven by the negative correlations between age and vanity as well as age and entitlement. Latent correlations were slightly higher than observed correlations in particular for entitlement: r = −.27 for the unidimensional model and r = −.18 for the multidimensional model vs. r = −.15 for the classical scoring approach, although these differences were not significant. Men tended to score higher than women on leadership and entitlement (e.g., r = .14 for leadership and r = .21 for entitlement in the multidimensional model). SES was mainly related to leadership and vanity. Education level was not related to any of the NPI facets except for a small negative correlation with entitlement in the multidimensional model (r = −.09). Overall, the correlations were similar across the scoring approaches, indicating that differences between scoring procedures appear to be negligible when correlations with criteria are of interest.

Furthermore, for the within sample, the correlations between the NPI facets and the ten PTPI personality traits and satisfaction with life were computed. As Table 5 shows, leadership correlated moderately with several traits such as vigor, mature personality, and self-confidence. Leadership as assessed by the NPI and the PTPI showed a strong overlap with r = .66 (.81) for forced-choice based on observed scores (latent correlations). The comparison of observed correlations with latent correlations within the forced-choice format indicated minor differences between relationships with other traits for the two scoring approaches: For leadership and vanity the average absolute difference was .06 and for entitlement it was slightly higher at .08. Thus, as for the criteria, relationships between NPI facets and other traits did not appear to be strongly distorted by using the classical scoring approach as opposed to the model-based approach.

Table 3
Correlation matrix for Narcissistic Personality Inventory facets measured with the forced-choice and rating scale format.

		Forced-choice			Rating scale	
		Lead	Vanity	Ent	Lead	Vanity
Observed scores						
Forced-choice	Vanity	.52				
	Ent	.35	.31			
Rating scale	Lead	**.82**	.47	.28		
	Vanity	.51	**.78**	.27	.58	
	Ent	.31	.31	**.51**	.38	.43
Latent traits						
Forced-choice	Vanity	.59				
	Ent	.70	.71			
Rating scale	Lead	**.91**	.55	.59		
	Vanity	.57	**.94**	.60	.66	
	Ent	.67	.56	**.69**	.85	.62

Note. Lead = leadership, Ent = entitlement. Correlations reflecting the relationship between the same trait measured with different response formats are in bold. N = 1246.

7.2.4. Research question 2b: Do external correlates differ across response formats?

To test whether external correlates of overall narcissism are affected by varying the response format, we compared latent correlations with criteria and PTPI traits across response formats. For the criteria age, sex, SES, and education level, latent correlations with overall narcissism were very similar across response formats (see Table 4). For example, the correlation between overall narcissism and age was −.11 for forced-choice, −.13 for true/false, and −.10 for rating scale. Across the four criteria, the mean absolute difference in latent correlations was 0.02 between forced-choice and true/false and 0.03 between forced-choice and rating scale.

Regarding the correlations with PTPI traits, latent correlations for the rating scale format were generally higher than the ones found for the forced-choice format (see Table 5). For instance, the latent correlation between overall narcissism and vigor was .40 for the forced-choice format and .51 for the rating scale format. The correlations differed significantly between forced-choice and rating scale for 9 out of 11 traits, indicating that varying the response format had an effect on the relationships between overall narcissism and other traits.

To investigate whether external correlates at the facet level were affected by varying the response format, we compared the latent correlations between NPI facets and criteria as well as the latent correlations between NPI facets and other traits obtained from different response formats. Latent correlations of leadership, vanity, and entitlement with the four criteria were in general very similar for the forced-choice, true/false and rating scale formats. For example, age consistently showed a negative correlation with vanity and entitlement: r = −.15 and r = −.27 for unidimensional forced-choice, r = −.17 and r = −.22 for true/false and r = −.12 and r = −.18 for rating scale, respectively.

Latent correlations between vanity and the four criteria did not differ notably across response formats (e.g., mean absolute difference across criteria 0.02 between unidimensional forced-choice and true/false). Differences for leadership were also small (e.g., 0.04 between unidimensional forced-choice and rating scale for latent correlations). Across all facets, none of the correlations with criteria differed significantly between the forced-choice and true/false or rating scale format, respectively. Thus, correlations

Table 4
Correlations between Narcissistic Personality Inventory facets and criteria.

	Response format														
	Unidimensional forced-choice				Multidimensional forced-choice			True/false				Rating scale			
	NPI	Lead	Van	Ent	Lead	Van	Ent	NPI	Lead	Van	Ent	NPI	Lead	Van	Ent
Observed															
Age	−.11	−.02	−.13	−.15				−.13	−.03	−.15	−.17	−.10	−.03	−.12	−.14
Sex	.12	.13	.05	.08				.11	.14	.02	.07	.13	.17	.05	.08
SES	.14	.14	.12	.02				.16	.16	.14	.02	.17	.17	.15	−.01
Edu	−.01	.01	−.01	−.01				−.01	−.00	−.01	−.01	−.01	.00	−.01	−.01
Latent															
Age	−.11	−.02	−.15	−.27	.01	−.18	−.18	−.11	−.02	−.17	−.22	−.10	−.05	−.12	−.18
Sex	.13	.14	.06	.14	.14	.06	.21	.11	.14	.03	.13	.15	.17	.06	.16
SES	.22	.22	.18	.05	.22	.15	.22	.26	.26	.20	.23	.27	.26	.23	.24
Edu	−.03	−.01	−.04	−.03	−.00	−.04	−.09	−.04	−.03	−.04	.00	−.07	−.07	−.05	−.07

Note. Lead = leadership, Van = vanity, Ent = entitlement, NPI = total score on Narcissistic Personality Inventory, Observed = mean score correlations, Latent = latent correlations, SES = socioeconomic status, Edu = education. The coding for sex was 0 = women, 1 = men. Education level was assessed with a 9-point rating scale from *some high school* to *completed graduate/professional degree*. SES was assessed with a 10-point rating scale depicted as a ladder. $N = 6690$ for forced-choice, $N = 5510$ for true/false, and $N = 5234$ for rating scale.

Table 5
Correlations between Narcissistic Personality Inventory facets and personality traits.

Observed scores	NPI	NPI	Leadership		Vanity		Entitlement	
Trait	FC	RS	FC	RS	FC	RS	FC	RS
Vigor	.36	.39	.37	.37	.25	.29	.07	.04
Calmness	.09	.05	.16	.13	.00	−.06	−.17*	−.33*
Mature personality	.23	.19	.32	.30	.07	−.01	−.07	−.17
Impulsiveness	.24	.27	.21	.24	.23	.27	.11	.18
Self-confidence	.29	.29	.38	.43	.15	.19	−.08	−.18
Culture	.24	.30	.24	.25	.21	.22	.03	−.03
Sociability	.30	.34	.28	.30	.29	.30	−.01	−.07
Leadership	.59	.64	.66	.70	.37	.39	.16	.16
Social Sensitivity	.09	.05	.12	.06	.07	−.04	−.13*	−.26*
Tidiness	.15	.18	.18	.20	.08	.07	−.01	−.03
Satisfaction with life	.13	.14	.16	.18	.08	.12	−.09	−.18
Average absolute difference	0.03		0.03		0.04		0.08	
Correlation	.99		.99		.96		.99	

Latent traits	NPI	NPI	Leadership		Vanity		Entitlement	
Trait	FC	RS	FC	RS	FC	RS	FC	RS
Vigor	.40*	.51*	.42*	.52*	.32	.37	.14*	.48*
Calmness	.06*	.26*	.20*	.40*	−.04	−.05	−.25*	.44*
Mature personality	.20*	.48*	.33*	.56*	.05	−.01	−.11*	.68*
Impulsiveness	.32	.39	.21	.28	.34	.38	.34	.42
Self-confidence	.40*	.58*	.51*	.68*	.22	.27	−.12*	.50*
Culture	.35*	.54*	.36*	.55*	.32	.34	.15*	.65*
Sociability	.37*	.52*	.36*	.52*	.36	.36	.05*	.60*
Leadership	.71*	.82*	.81*	.90*	.49	.51	.30*	.72*
Social Sensitivity	.11*	.33*	.17*	.42*	.08	.02	−.14*	.57*
Tidiness	.14*	.28*	.19*	.33*	.07	.10	.01*	.43*
Satisfaction with life	.12	.22	.18	.26	.08	.12	−.19*	.25*
Average absolute difference	0.16		0.15		0.03		0.51	
Correlation	.95		.95		.98		.34	

Note. FC = forced-choice, RS = rating scale. $N = 1246$.
* $p \leqslant .05$.

between criteria and the NPI facets did not appear to be affected by varying the response format.[3]

In contrast, latent correlations between the NPI facets and PTPI personality traits and satisfaction with life differed more strongly between the forced-choice and rating scale format (see Table 5). For leadership correlations with nine out of 11 traits differed significantly between forced-choice and rating scale, leading to an average absolute difference of 0.15. For vanity, none of the correlations differed significantly (average absolute difference between forced-choice and rating scale correlations 0.03). The correlational pattern for entitlement differed significantly across response formats for most traits (10 out of 11; see Table 5). For example, the latent correlation between vigor and entitlement was .14 for the forced-choice format while it was .48 for the rating scale format (difference significant at $\alpha = 0.05$). Consequently, the average absolute difference in the correlations across all traits was higher compared with the other NPI facets: 0.51 for latent correlations. In sum, latent correlations between vanity and other

[3] We also computed observed correlations with criteria for mean scores based on Ackerman et al.'s (2015) facets. These are included in supplemental Table S7. For facets that largely corresponded between our study and Ackerman et al. (i.e., leadership and vanity), the correlations were very similar. This indicates that correlations with criteria were also robust across slightly different facet compositions.

traits were not affected by varying the response format, whereas for leadership and in particular entitlement, correlations were affected by varying the response format.

8. Discussion

The correlates of narcissism reported in previous research have largely been based on NPI total scores and it was unclear whether they may have been artifacts of how the NPI's forced-choice responses were scored. This study indicates that the correlations between NPI scores and external variables are highly similar for the classical scoring approach and the model-based approach, even when multidimensionality in item pairs was modeled. However, correlations differed substantially across the three NPI facets, indicating that some relationships are not adequately represented when only the NPI total score is used. Furthermore, while correlations with external variables were robust to classical scoring, scores on the entitlement facet were not. Thus, estimates of participants' trait levels may differ depending on whether the forced-choice nature of the items is modeled or not.

In addition, we systematically varied the response format in order to investigate whether correlations differed across response formats. For overall narcissism, results were stable across response formats with respect to the correlations with criteria, but not with respect to correlations with other traits. Results were the most stable across response formats for the vanity facet, indicating that the forced-choice, true/false, and rating scale formats all capture the same underlying construct and can be considered alternate forms for the assessment of vanity. For the leadership facet, correlations with criteria were robust across response formats, whereas correlations with traits were not, indicating that the equivalence of the measured construct is questionable. Relationships between entitlement and other traits showed the largest differences across response formats. Thus, the three response formats are equally poor measures of entitlement and differ with respect to the nature of the underlying construct. In the following, we will compare previously reported results and our results on the relationships between NPI scores and criteria and discuss the implications of our results for the use of the NPI and for the measurement of narcissism.

8.1. Relationships between NPI facets and other variables

The correlation between the NPI total score and age in our study (−.11) was smaller than the ones reported previously which were between −.22 (Foster et al., 2003) and −.32 (Hill & Roberts, 2012; Roberts et al., 2010). Facet-level correlations revealed that the relationship between narcissism and age is mainly driven by a decline in vanity (r = −.15 for unidimensional forced-choice) and entitlement (r = −.27 for unidimensional forced-choice) whereas leadership was not related to age.

The correlation between sex and the NPI total score (.12) in our study was identical to the correlation reported in the meta-analysis by Grijalva et al. (2015) and also indicated that men reported slightly higher narcissism than women. Of the three NPI facets, leadership and entitlement showed the highest correlations with sex (between .14 and .21 for latent correlations). Piff (2014) reported that higher socioeconomic status was related to higher narcissism scores (r = .16) and higher entitlement scores (r = .17). Small to moderate correlations between socioeconomic status and NPI scores were also found in our study for the total score and the leadership and vanity facets (e.g., latent r = .22 for total NPI) in the unidimensional forced-choice format. For data based on the multidimensional forced-choice format and the true/false and rating scale format, socioeconomic status was positively related to all NPI facets, e.g., latent correlations between .20 and .26 for true/false data. These correlations demonstrate the differential validity of the NPI facets for predicting criteria. Future research could also investigate the incremental validity of one facet over the others or investigate in how far the specific facets predict criteria over and above a shared narcissism dimension in the context of a bifactor model (see for example Boldero et al., 2015).

In sum, the general pattern of several previously reported correlations was confirmed in our study, though our results indicate that correlations differ strongly across NPI facets – not only in magnitude but also in some cases in direction. These discrepancies were also found for correlations with the PTPI personality traits. Taking into account only the NPI total score therefore provides a distorted picture of the relationships of interest. In addition, correlations differed substantially across response formats for entitlement and-to a lesser degree-also for leadership, indicating that the findings reported previously are not generalizable to the application of other response formats.

8.2. Implications for the use of the NPI

This study applied a model-based approach to scoring the NPI that takes into account its forced-choice response format by modeling the dependencies between response options presented as a pair. The classical scoring practice of summing up the narcissistic responses (whether on the facet or total score level) treats the items as if they had been presented in a single-stimulus format and ignores the forced-choice nature of the items. Treating relative ratings (i.e., the preference of one response option over the other) as absolute scores may be less problematic if all item pairs are strictly unidimensional in the sense that the response options in the item pair both measure the same trait. However, as depicted above, there are a large number of multidimensional item pairs in the NPI (see also Ackerman et al., 2015). While this scoring practice did not appear to distort the relationships with external variables, estimates of participants' standings on the entitlement facet differed substantially between the model-based scoring and the observed mean scoring procedure. Since entitlement is the most maladaptive of the three NPI facets (Ackerman et al., 2011), this is potentially problematic. For example, if the goal in a selection context was to eliminate applicants with the highest levels of entitlement, applicants may incorrectly be classified as low or high on entitlement based on the observed scores. Thus, when individual trait estimates are of interest, the model-based scoring approach or an alternative instrument to assess narcissism should be applied.

The discrepancies between correlations based on the NPI total score and correlations computed separately for the NPI facets illustrate that using the NPI total score to investigate relationships between narcissism and other traits or criteria provides an inaccurate and muddled picture of the relationships. The computation of a total score is based on the assumption that the items measure one common trait. Numerous previous studies of the NPI's factor structure (e.g., Ackerman et al., 2011) as well as the ESEM analyses in this study demonstrate that the NPI comprises several different traits. In our study, three facets were found (leadership, vanity, entitlement) that were overall similar to the three facets reported in (Ackerman et al., 2015). Hence, it is important to conduct analyses at the facet-level.

8.3. Different response formats for the measurement of Narcissism

The NPI may partly have gained its immense popularity because it is assumed that the application of the forced-choice format

successfully eliminates socially desirable responding. Unfortunately, this is only the case when statements of very similar desirability are compared (i.e., the items in a pair are matched on desirability; Drasgow, Chernyshenko, & Stark, 2009). In the NPI, two opposing statements with very different desirability levels are compared in each pair since most item pairs contain one response option clearly identifiable as narcissistic and one response option clearly identifiable as non-narcissistic. Explicit comparison makes the more desirable statement rather obvious, no less than in the single-stimulus format, thus rendering the potential benefits of forced-choice non-existent (Feldman & Corah, 1960). A future direction for researchers concerned with socially desirable responding may therefore be to construct a more valid and reliable (multidimensional) forced-choice questionnaire for the assessment of narcissism. As Brown and Maydeu-Olivares (2011) have shown, model-based scoring of forced-choice data can achieve comparable construct validity, criterion validity, and reliability as rating scale data when certain guidelines (e.g., concerning the pairing of the items) are followed during test construction.

Modifying the NPI's response format and presenting the narcissistic response options with a single-stimulus response format as a number of studies have done is problematic because the meaning of the constructs can change and relationships with other traits (though not criteria) differ from those found for the original forced-choice format, making comparisons across studies difficult. If researchers want to use a single-stimulus response format, it seems preferable to apply a narcissism questionnaire that has been constructed explicitly for the single-stimulus response format such as the Pathological Narcissism Inventory (Pincus et al., 2009) or the Narcissistic Admiration and Rivalry Questionnaire (Back et al., 2013) rather than adapting the NPI's response format.

8.4. Limitations of the study

Limitations of this study include that the true/false format was not applied in the within sample and that the forced-choice and rating scale versions of the NPI were filled out within one testing session. Thus, correlations of the same constructs across response formats and correlations between NPI facets and PTPI traits could only be compared between the forced-choice and the rating scale format. Furthermore, the factor structure of the forced-choice format was imposed on the true/false and rating scale formats. This was necessary for purposes of comparison and differences across response formats were minor, but we recognize that using the appropriate factor structure for the true/false and rating scale formats would have yielded slightly different correlations (e.g., lower correlations of one construct measured with different response formats).

9. Conclusion

This study showed that the relationships of overall narcissism and the NPI facets leadership, vanity, and entitlement to criteria and personality traits were robust across the classical and model-based scoring approaches, but only in part across response formats. The scoring approaches and response formats achieved equivalent measurements of the vanity facet and in part of the leadership facet, but differed with respect to the entitlement facet.

Acknowledgments

This research was supported by a post-doc fellowship awarded to Eunike Wetzel by the German Academic Exchange Service (DAAD) and by NIA Grant 2AG21178 to Brent W. Roberts.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jrp.2015.12.002.

References

Ackerman, R. A., Donnellan, M. B., Roberts, B. W., & Fraley, R. C. (2015). The effect of response format on the psychometric properties of the Narcissistic Personality Inventory: Consequences for item meaning and factor structure. *Assessment*. http://dx.doi.org/10.1177/1073191114568113.

Ackerman, R. A., Witt, E. A., Donnellan, M. B., Trzesniewski, K. H., Robins, R. W., & Kashy, D. A. (2011). What does the Narcissistic Personality Inventory really measure? *Assessment, 18*(1), 67–87. http://dx.doi.org/10.1177/1073191110382845.

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23. http://dx.doi.org/10.1177/0146621697211001.

Asparouhov, T., & Muthen, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*(3), 397–438. http://dx.doi.org/10.1080/10705510903008204.

Back, M. D., Küfner, A. C., Dufner, M., Gerlach, T. M., Rauthmann, J. F., & Denissen, J. J. (2013). Narcissistic admiration and rivalry: Disentangling the bright and dark sides of narcissism. *Journal of Personality and Social Psychology, 105*(6), 1013–1037. http://dx.doi.org/10.1037/a0034431.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. http://dx.doi.org/10.1037/0033-2909.107.2.238.

Bogart, L. A., Benotsch, E. G., & Pavlovic, J. D. (2004). Feeling superior but threatened: The relation of narcissism to social comparison. *Basic and Applied Social Psychology, 26*(1), 35–44. http://dx.doi.org/10.1207/S15324834basp2601_4.

Boldero, J. M., Bell, R. C., & Davies, R. C. (2015). The structure of the Narcissistic Personality Inventory with binary and rating scale items. *Journal of Personality Assessment*, 1–12. http://dx.doi.org/10.1080/00223891.2015.103901.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*(3), 460–502. http://dx.doi.org/10.1177/0013164410375112.

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods, 44*(4), 1135–1147. http://dx.doi.org/10.3758/s13428-012-0217-x.

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18*(1), 36–52. http://dx.doi.org/10.1037/a0030641.

Brown, R. P., & Zeigler-Hill, V. (2004). Narcissism and the non-equivalence of self-esteem measures: A matter of dominance? *Journal of Research in Personality, 38*(6), 585–592. http://dx.doi.org/10.1016/j.jrp.2003.11.002.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Cain, N. M., Pincus, A. L., & Ansell, E. B. (2008). Narcissism at the crossroads: Phenotypic description of pathological narcissism across clinical theory, social/personality psychology, and psychiatric diagnosis. *Clinical Psychology Review, 28*, 638–656.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Erlbaum.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research, 34*(3), 315–346. http://dx.doi.org/10.1207/S15327906mbr3403_2.

Corry, N., Merritt, R. D., Mrug, S., & Pamp, B. (2008). The factor structure of the Narcissistic Personality Inventory. *Journal of Personality Assessment, 90*(6), 593–600. http://dx.doi.org/10.1080/00223890802388590.

De Ayala, R. J. (2009). *Theory and practice of item response theory*. New York: Guilford Publications.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49*(1), 71–75. http://dx.doi.org/10.1207/s15327752jpa4901_13.

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2009). Test theory and personality measurement. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 59–80). New York: Oxford University Press.

Egan, V., & Lewis, M. (2011). Neuroticism and agreeableness differentiate emotional and narcissistic expressions of aggression. *Personality and Individual Differences, 50*(6), 845–850. http://dx.doi.org/10.1016/j.paid.2011.01.007.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, N.J.: L. Erlbaum Associates.

Emmons, R. A. (1984). Factor analysis and construct validity of the Narcissistic Personality Inventory. *Journal of Personality Assessment, 48*(3), 291–300. http://dx.doi.org/10.1207/s15327752jpa4803_11.

Feldman, M. J., & Corah, N. L. (1960). Social desirability and the forced choice method. *J Consulting Psychology, 24*, 480–482.

Foster, J. D., Campbell, W. K., & Twenge, J. M. (2003). Individual differences in narcissism: Inflated self-views across the lifespan and around the world. *Journal of Research in Personality, 37*(6), 469–486. http://dx.doi.org/10.1016/S0092-6566(03)00026-6.

Gerbasi, M. E., & Prentice, D. A. (2013). The self- and other-interest inventory. *Journal of Personality and Social Psychology, 105*(3), 495–514. http://dx.doi.org/10.1037/A0033483.

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*(2), 261–310. http://dx.doi.org/10.1037/a0038231.

Hill, P. L., & Roberts, B. W. (2012). Narcissism, well-being, and observer-rated personality across the lifespan. *Social Psychological and Personality Science, 3*(2), 216–223. http://dx.doi.org/10.1177/1948550611415867.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling – A Multidisciplinary Journal, 6*(1), 1–55. http://dx.doi.org/10.1080/10705519909540118.

Jordan, C. H., Spencer, S. J., Zanna, M. P., Hoshino-Browne, E., & Correll, J. (2003). Secure and defensive high self-esteem. *Journal of Personality and Social Psychology, 85*(5), 969–978. http://dx.doi.org/10.1037/0022-3514.85.5.969.

Kubarych, T. S., Deary, I. J., & Austin, E. J. (2004). The Narcissistic Personality Inventory: Factor structure in a non-clinical sample. *Personality and Individual Differences, 36*, 857–872. http://dx.doi.org/10.1016/S0191-8869(03)00158-2.

Lee, S. Y., Gregg, A. P., & Park, S. H. (2013). The person in the purchase: Narcissistic consumers prefer products that positively distinguish them. *Journal of Personality and Social Psychology, 105*(2), 335–352. http://dx.doi.org/10.1037/A0032703.

Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research, 45*(6), 935–974. http://dx.doi.org/10.1080/00273171.2010.531231.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology, 77*, 531–551. http://dx.doi.org/10.1348/0963179042596504.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568–592. http://dx.doi.org/10.1037/0021-9010.93.3.568.

Morf, C. C., & Rhodewalt, F. (1993). Narcissism and self-evaluation maintenance – Explorations in object relations. *Personality and Social Psychology Bulletin, 19*(6), 668–676. http://dx.doi.org/10.1177/0146167293196001.

Muthén, L. K., & Muthén, B. O. (1998–2014). Mplus [Computer software]. Los Angeles, CA: Muthén & Muthén. Retrieved <http://www.statmodel.com>.

Piff, P. K. (2014). Wealth and the inflated self: Class, entitlement, and narcissism. *Personality and Social Psychology Bulletin, 40*(1), 34–43. http://dx.doi.org/10.1177/0146167213501699.

Pincus, A. L., Ansell, E. B., Pimentel, C. A., Cain, N. M., Wright, A. G., & Levy, K. N. (2009). Initial construction and validation of the Pathological Narcissism Inventory. *Psychological Assessment, 21*(3), 365–379. http://dx.doi.org/10.1037/a0016530.

Pozzebon, J., Damian, R. I., Hill, P. L., Lin, Y. C., Lapham, S., & Roberts, B. W. (2013). Establishing the validity and reliability of the Project Talent Personality Inventory. *Frontiers in Psychology, 4*. http://dx.doi.org/10.3389/Fpsyg.2013.00968.

Raskin, R. N., & Hall, C. S. (1979). Narcissistic personality inventory. *Psychological Reports, 45*(2), 590.

Raskin, R. N., & Terry, H. (1988). A principal-components analysis of the Narcissistic Personality Inventory and further evidence of its construct validity. *Journal of Personality and Social Psychology, 54*(5), 890–902.

Rhodewalt, F., & Morf, C. C. (1995). Self and interpersonal correlates of the Narcissistic Personality Inventory – A review and new findings. *Journal of Research in Personality, 29*(1), 1–23. http://dx.doi.org/10.1006/jrpe.1995.1001.

Roberts, B. W., Edmonds, G., & Grijalva, E. (2010). It is developmental me, not generation me: Developmental changes are more important than generational changes in narcissism – Commentary on Trzesniewski & Donnellan (2010). *Perspectives on Psychological Science, 5*(1), 97–102. http://dx.doi.org/10.1177/1745691609357019.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No.17), Retrieved <http://www.psychometrika.org/journal/online/MN17.pdf>.

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

Steiger, J. H. (1990). Structural model evaluation and modification – an interval estimation approach. *Multivariate Behavioral Research, 25*(2), 173–180. http://dx.doi.org/10.1207/s15327906mbr2502_4.

Trzesniewski, K. H., Donnellan, M. B., & Robins, R. W. (2008). Is "Generation Me" really more narcissistic than previous generations? *Journal of Personality, 76*(4), 903–918. http://dx.doi.org/10.1111/j.1467-6494.2008.00508.x.

Tucker, L. R., & Lewis, C. (1973). Reliability coefficient for maximum likelihood factor-analysis. *Psychometrika, 38*(1), 1–10. http://dx.doi.org/10.1007/Bf02291170.