Assessment
2021, Vol. 28(1) 29–43
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1073191120907967
journals.sagepub.com/home/asm
SAGE

Measurement Invariance of Three Narcissism Questionnaires Across the United States, the United Kingdom, and Germany

Eunike Wetzel[1,2,3], Felix J. Lang[1], Mitja D. Back[4], Michele Vecchione[5], Radoslaw Rogoza[6], and Brent W. Roberts[7]

Abstract

With a recent surge of research on narcissism, narcissism questionnaires are increasingly being translated and applied in various countries. The measurement invariance of an instrument across countries is a precondition for being able to compare scores across countries. We investigated the cross-cultural measurement invariance of three narcissism questionnaires (Brief Pathological Narcissism Inventory [B-PNI], Narcissistic Personality Inventory [NPI], and Narcissistic Admiration and Rivalry Questionnaire [NARQ]) and mean-level differences across samples from the United States (N = 2,464), the United Kingdom (N = 307), and Germany (N = 925). Overall, the B-PNI and NARQ functioned equivalently for the U.S. and U.K. participants. More violations of measurement invariance were found between Germany and the combined U.S. and U.K. samples, and for the NPI. In the B-PNI and NARQ, Americans scored higher than individuals from the United Kingdom regarding agentic aspects (self-sacrificing self-enhancement, admiration), while Germans scored lower than both Americans and U.K. individuals regarding antagonistic (entitlement rage, rivalry) and neurotic (hiding the self, contingent self-esteem) aspects. More inconsistent results were found for NPI facets. When noninvariance was present, observed means yielded biased results. Thus, the degree of measurement invariance across translated instrument versions should be considered in cross-cultural comparisons, even with culturally similar countries.

Keywords

narcissism, measurement invariance, measurement equivalence, cross-cultural differences, Brief Pathological Narcissism Inventory, Narcissistic Personality Inventory, Narcissistic Admiration and Rivalry Questionnaire

In recent years, there has been an increased interest in studying narcissism, which is justified by findings that narcissism predicts important outcomes such as counterproductive work behaviors (Penney & Spector, 2002), being unemployed (Leckelt et al., 2019), achieving a leadership position, and getting divorced (Wetzel et al., 2019). Narcissism questionnaires are increasingly being applied in different languages and countries than the ones they originated from. For example, the Narcissistic Admiration and Rivalry Questionnaire (NARQ; Back et al., 2013) was developed in German and English and now has been translated and validated in several other languages including Polish (Rogoza et al., 2016), Spanish (Doroszuk et al., 2020), and Italian (Vecchione et al., 2018). However, in most cases, instruments are adapted to other languages without checking the measurement invariance of the translated version with the original. Testing for measurement invariance is essential to ensure that the same construct is being measured and that different versions of the instrument function the same way. Measurement invariance is highly relevant to the generalizability of research on personality traits. For example, if the narcissism facet admiration does not have the same meaning between say, Americans and Germans, research on narcissism conducted with Americans cannot be generalized to Germans. In the same way, if Germans respond to certain

[1] University of Mannheim, Mannheim, Baden-Württemberg, Germany
[2] University of Vienna, Vienna, Austria
[3] Otto-von-Guericke University Magdeburg, Magdeburg, Germany
[4] University of Münster, Münster, Germany
[5] Sapienza University of Rome, Rome, Italy
[6] University of Cardinal Stefan Wyszynski, Warsaw, Poland
[7] University of Illinois at Urbana–Champaign, Champaign, IL, USA

Corresponding Author:
Eunike Wetzel, Department of Psychology, Otto-von-Guericke University Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany.
Email: eunike.wetzel@ovgu.de

narcissism items differently than Americans, despite having the same latent trait level, comparisons across the two groups need to take these differences into account to still be accurate. Measurement noninvariance can occur even between similar cultures speaking the same language (Doroszuk et al., 2020) as well as between ethnicities within one country (Wetzel et al., 2017). The goal of this study was to test the measurement invariance of three popular narcissism questionnaires across three countries, the United States, the United Kingdom, and Germany.

Establishing Measurement Invariance

There are many reasons why measures might not work equivalently across countries and cultures. The most critical reason for inequivalence would be that the construct does not exist in the same form in different cultures (i.e., a lack of conceptual equivalence). In addition to issues of conceptual equivalence, heterogeneity across countries in terms of the quality of item translations, the relevance of concepts expressed in items, and whether culturally specific knowledge is needed to fill out the items, may affect the measurement properties of the instrument.

Analyzing the equivalence of measures is a straightforward enterprise. In these analyses, it is first investigated whether the factor structure (number of factors and the pattern of salient and nonsalient loadings) is equivalent across countries. If this is the case, further restrictions representing stronger degrees of equivalence can be tested. Second, by constraining the factor loadings to equality across countries, it can be tested whether the items relate to the trait in the same way in the different countries. Third, by constraining the item intercepts to equality across countries, it can be tested whether the observed means conditional on the trait level are the same across countries. Fourth, by constraining the items' residual variances to equality across countries, it can be tested whether the amount of variance in the items not accounted for by the trait is the same across countries. Invariance in the general factor structure is referred to as configural invariance, invariance in factor loadings is referred to as metric invariance, invariance in factor loadings and item intercepts is referred to as scalar invariance, and invariance in factor loadings, item intercepts, and residual variances is referred to as strict invariance (Meredith, 1993). More detailed information on the process of testing for measurement invariance can be found in Vandenberg and Lance (2000) and Widaman and Reise (1997). Often, full invariance (i.e., invariance for all items) does not hold. In this case, equality constraints can be relaxed for the noninvariant parameters and partial invariance can be achieved (Byrne et al., 1989; Steenkamp & Baumgartner, 1998). Unbiased estimates of the latent mean differences between the countries can then be obtained from the final (partial) invariance model when there are few noninvariant items relative to the number of invariant items (Guenole & Brown, 2014).

Some prior research has examined the measurement equivalence of specific narcissism measures across several countries. For example, Leckelt et al. (2018) found that metric invariance held across German and combined U.S. and U.K. samples in a short version of the NARQ (NARQ-S), though they did not test for scalar invariance. Doroszuk et al. (2020) found partial scalar invariance between Spain, Chile, and Colombia for the Spanish NARQ. Zemojtel-Piotrowska et al. (2018) investigated measurement invariance across samples from the United Kingdom, Japan, and Poland in a short version of the Narcissistic Personality Inventory [NPI], the NPI-13 (Gentile et al., 2013), but their scalar model did not converge, which according to additional analyses may have been due to noninvariant items on the entitlement/exploitativeness facet. Meisel et al. (2016) found that metric invariance did not hold for the 40-item NPI between U.S. and Chinese university students. Thus, prior research on the measurement equivalence of narcissism has been somewhat unsystematic as different forms of invariance have not been consistently tested across countries (e.g., several studies only investigated metric, but not scalar, invariance). Furthermore, previous studies only applied a pass/fail decision rule for whether measurement invariance existed and did not allow for partial invariance or considered the effect size of the noninvariance.

Cross-Cultural Differences in Narcissism

Americans tend to be perceived as more narcissistic than people from other countries (Campbell et al., 2010; Miller et al., 2015). For example, in a study on perceptions of national character, Miller et al. (2015) found that people from Basque Country, England, China, and Turkey rated Americans as more narcissistic than members of their own world region. It is unclear whether these perceptions are accurate or whether they are just stereotypes stemming from the portrayal of Americans in movies and the media (e.g., celebrities). Previous research on self-reported narcissism supports the perception that people in the United States are more narcissistic than people from other countries. For example, Foster et al. (2003) compared scores on the Narcissistic Personality Inventory (NPI; Raskin & Hall, 1979; Raskin & Terry, 1988) across five world regions and found that participants from the United States reported the highest NPI scores, followed by Europe, Canada, Asia, and the Middle East. Jonason et al. (2017) found that Americans showed higher narcissism scores than participants from Australia, Russia, Hungary, Brazil, and Japan, with Japanese participants showing the lowest narcissism scores. Fukunishi et al. (1996) conducted an analysis of variance of NPI scores between students from the United States, Japan, and China and found the highest scores for Chinese students. While some studies appear to confirm the perspective that people in the United States are more narcissistic than people in other countries, results are still inconclusive. In addition, these previous studies failed to consider a fundamental issue necessary for making valid and comprehensive comparisons: establishing whether the narcissism measure was being used equivalently across countries. Thus, in the present study, we investigated whether Americans on average are more narcissistic than individuals from two other Western countries, the United Kingdom and Germany, after establishing measurement invariance.

Differentiating Distinct Aspects of Narcissism

Existing research on cross-cultural differences in narcissism has focused on single measures of narcissism, which fails to reflect the often complex and multifaceted ways in which narcissism is currently defined and operationalized. Research on narcissism has emerged out of at least two traditions in psychology reflected in clinical understandings of the concept (Pincus et al., 2009) and more normal-range, personality-trait based conceptualizations of narcissism (Back et al., 2013; Robins et al., 2001). These different perspectives have led to a number of different questionnaires with heterogeneous narcissistic content. More recent research across traditions and measurement instruments shows that a three-dimensional distinction of agentic, antagonistic, and neurotic narcissism is more appropriate and allows one to disentangle functionally distinct aspects of narcissism with different correlates and outcomes (Back, 2018; Back & Morf, in press; Crowe et al., 2019; Krizan & Herlache, 2018; Miller et al., 2016). Here, we focus on three questionnaires that collectively capture all relevant narcissistic aspects: the NPI, the NARQ, and the Pathological Narcissism Inventory (PNI).

The NPI was developed to assess the *Diagnostic and Statistical Manual of Mental Disorders–Third Edition* criteria of narcissistic personality disorder (American Psychiatric Association, 1980), though factor analyses suggest that the content is mainly adaptive with a preponderance of items referring to agentic aspects such as leadership (Ackerman et al., 2011; Wetzel, Roberts, et al., 2016), and only some antagonistic content (exploitativeness/entitlement/manipulation). The NARQ was developed to assess the two components of the Narcissistic Admiration and Rivalry Concept (Back et al., 2013): *admiration* (agentic self-promotion) and *rivalry* (antagonistic self-defense).

The PNI (Pincus et al., 2009) was designed to assess clinical forms of narcissism and therefore contains content more directly related to distress. Its vulnerability domain contains facets with neurotic content such as contingent self-esteem and hiding the self, while its grandiosity domain is more mixed and contains a facet with mostly antagonistic content (exploitativeness), a facet with mostly agentic content (grandiose fantasies), and a facet with mostly communal content (self-sacrificing self-enhancement). To provide comprehensive coverage of the different forms and operationalizations of narcissism, we investigated measurement invariance and mean differences across three countries on these three different narcissism measures.

The Present Study

In our study, we aimed at examining whether three narcissism questionnaires were equivalent across the United States, the United Kingdom, and Germany. If at least partial scalar measurement invariance was established, we additionally investigated whether perceptions of Americans as more narcissistic than natives of other countries are true by comparing Americans' mean levels with those of people from the United Kingdom and Germany. We analyzed data from the Brief Pathological Narcissism Inventory (B-PNI; Schoenleber et al., 2015), the NPI (Raskin & Hall, 1979; Raskin & Terry, 1988), and the NARQ (Back et al., 2013), allowing us to capture potential differences in agentic, antagonistic, and neurotic aspects of narcissism. The data were collected in five countries (the United States, the United Kingdom, Germany, Italy, and Poland), but because only the U.S., U.K., and German samples filled out all three questionnaires, we focus our analysis on these three countries. In the supplementary online material (SOM; https://osf.io/53amg/), we additionally report analyses of measurement invariance and mean differences for samples from Italy (NPI and NARQ) and Poland (B-PNI and NARQ). Based on prior research we expected to find varying levels of measurement invariance and comparability of these measures across the United States, United Kingdom, and Germany.

Method

Samples

For all samples, only participants aged between 18 and 50 years were included in the analyses to make the age distributions across countries more similar. The purpose of this was to avoid confounding potential noninvariance across countries with noninvariance due to age or cohort (Wetzel et al., 2017).

German Sample.

The German sample consisted of 925 participants (72% female) who filled out an online survey. Their mean age was 26.33 (*SD* = 6.41). Participants were recruited by posting the study link on multiple student mailing lists at universities in Germany as well as on Psytests (http://www.psytests.de), a large online panel formerly hosted by the Humboldt-University Berlin and now hosted by the University of Göttingen. Participants could win one of six vouchers worth 50€. The data from this sample were also analyzed in Study 1 in Back et al. (2013), Study 2 in Wetzel, Leckelt, et al. (2016), Leckelt et al. (2018), and Grosz et al. (2017). There is no overlap with the research questions and analyses of the current study.

Of the 925 participants, 251 dropped out before the end of the survey. We conducted attrition analyses between dropouts and completers at a Bonferroni-corrected α of .004. Dropouts did not differ significantly from people who completed the survey in terms of their gender composition, $\chi^2(1) = 6.15, p = .013$, their mean scores on the B-PNI facets: $t(84) = -0.88, p = .381$, for exploitativeness; $t(84) = -1.03, p = .308$, for self-sacrificing self-enhancement; $t(86) = 0.66, p = .511$, for grandiose fantasy; $t(87) = -0.10, p = .920$, for contingent self-esteem; $t(87) = 1.24, p = .217$, for hiding the self; $t(86) = 1.78, p = .078$, for devaluing; and $t(89) = 1.29, p = .202$, for entitlement rage, their mean scores on the NPI facets: $t(311) = -0.85, p = .397$, for leadership; $t(315) = -1.12, p = .264$, for vanity; and $t(307) = -2.06, p = .040$ for entitlement, or their mean scores on the NARQ facets: $t(432) = -0.25, p = .799$, for admiration; and $t(402) = -1.14, p = .253$, for rivalry. The only significant difference found between dropouts and completers was in their average ages, $t(499) = 3.27, p = .001$, with dropouts on average being slightly younger than completers.

U.S. and U.K. Samples. The U.S. and U.K. samples were collected together in an online survey. The survey was hosted on www.yourpersonality.net and was available for people from all over the world. There was no specific recruitment strategy, but anyone who searched the Internet for personality tests or narcissism could have come across this website. For the purposes of this study, we only extracted data from participants who reported that their country of residence was the United States or the United Kingdom. The U.S. sample originally consisted of 2,954 participants aged between 18 and 50 years. We removed 210 participants who had participated more than once and 280 participants who failed one or both instructed response items. The final U.S. sample therefore consisted of 2,464 participants (76% female, $M_{\text{age}} = 30.32$, $SD_{\text{age}} = 9.31$). The same data quality checks were applied to the U.K. sample, which reduced the sample size from 417 to 307. In the U.K. sample, 70% were female and the average age was 33.16 ($SD = 10.01$; Table 1). Data were only saved at the end of the survey. Therefore, we do not have any information on how many people may have started the survey and dropped out before the end. Furthermore, participants were required to provide a response in order to move on to the next page. Thus, there are no missing data on any of the survey items. Participants in the U.S. and U.K. samples received feedback on their narcissism scores in the three questionnaires at the end of the survey.[1,2]

Measures

Means scores, standard deviations, and omega reliabilities for facet scores by questionnaire and sample are depicted in Table S1 in SOM 1. Table S2 shows observed score correlations among all facets for the U.S. sample. Descriptive statistics for the Italian and Polish samples, which were used for supplementary analyses, can be found in SOM 2.

Table 1. Descriptive Statistics for the Three Samples.

Country	*N*	% Female	Age, *M* (*SD*)
United States	2,464	76	30.32 (9.31)
United Kingdom	307	70	33.16 (10.01)
Germany	925	72	26.33 (6.41)

B-PNI. The B-PNI (Schoenleber et al., 2015) was developed as a short version of the PNI (Pincus et al., 2009), which assesses pathological narcissism. The B-PNI retained the facet structure of the PNI, thus distinguishing between the subscales *exploitativeness* (e.g., "I find it easy to manipulate people"), *self-sacrificing enhancement* (e.g., "Sacrificing for others makes me the better person"), and *grandiose fantasy* (e.g., "I often fantasize about performing heroic deeds"), which together form the grandiosity composite. The vulnerability composite is made up of the subscales *contingent self-esteem* (e.g., "When people don't notice me, I start to feel bad about myself"), *hiding the self* (e.g., "It's hard to show others the weaknesses I feel inside"), *devaluing* (e.g., "Sometimes I avoid people because I'm afraid they won't do what I want them to do"), and *entitlement rage* (e.g., "I get annoyed by people who are not interested in what I say or do"). In the B-PNI, each of these facets is assessed with four items using a 6-point rating scale ranging from 1 (*not at all like me*) to 6 (*very much like me*). In the German sample, the full German PNI (preliminary version by Back et al. [2013]; final version see Morf et al., 2017) was applied and we scored the B-PNI from this full version. Translation and back-translation by a native speaker were applied to create the German version (see Morf et al., 2017).

NPI. The NPI (Raskin & Hall, 1979; Raskin & Terry, 1988; in German by Schütz et al., 2004) assesses grandiose, nonpathological narcissism with 40 item pairs. Each item pair consists of a narcissistic option (e.g., "I like to look at myself in the mirror") and a nonnarcissistic option (e.g., "I am not particularly interested in looking at myself in the mirror"). Participants are instructed to select the option out of the pair that best describes their feelings and beliefs. The factor structure of the NPI has been a subject of debate. Here, we use the factor structure obtained by a Thurstonian item response analysis of NPI data, which takes the dependencies between items presented in a pair into account. This analysis yielded three facets (Wetzel, Roberts, et al., 2016): *leadership* (e.g., "I would prefer to be a leader" vs. "It makes little difference to me whether I am a leader or not"), *vanity* (e.g., "I like to look at myself in the mirror" vs. "I am not particularly interested in looking at myself in the mirror"), and *entitlement* ("I will never be satisfied until I get all that I deserve" vs. "I take my satisfactions as they come"). Leadership is assessed with 22 item pairs, vanity with 13, and entitlement with 11 item pairs. Several item pairs loaded on two facets: two on leadership and vanity, three on leadership and entitlement, and four on vanity and entitlement. The German version of the NPI was constructed using translation and back-translation.

NARQ. The NARQ assesses grandiose narcissism on two dimensions: admiration and rivalry. *Admiration* consists of agentic aspects of self-enhancement (e.g., "I am great"), whereas *rivalry* consists of antagonistic self-defense (e.g., "I want my rivals to fail"). Each dimension contains nine items, which participants respond to on a 6-point rating scale from 1 (*not agree at all*) to 6 (*agree completely*). The NARQ was simultaneously developed in German and English by Back et al. (2013). The team of authors collectively translated the German items to English and had a bilingual person back-translate them to German. The similarity to the original items was coded and minor adjustments made to the items.

In the U.S. and U.K. samples, the order in which the three questionnaires were presented was randomized. In the German sample, all participants filled out the questionnaires in the order NARQ, NPI, PNI. The data for all samples are available from https://osf.io/hbuqz/.

Analyses

We tested for cross-country measurement invariance in multigroup item response models. For the NPI, the underlying model was the Thurstonian item response model (Brown & Maydeu-Olivares, 2011), which is a two-parameter logistic (2PL) model for forced-choice data that takes the dependencies between the items presented in a pair into account. For the B-PNI and NARQ, we used the graded response model (Samejima, 1969), which is a 2PL model for data from ordered rating scales. In the graded response model, the probability of endorsing a certain response category or the ones above it is parameterized with thresholds and there are one fewer thresholds than response categories for each item (e.g., five thresholds for a 6-point rating scale as in the B-PNI and NARQ). We started with a fully constrained strict invariance model with factor loadings,[3] item thresholds, and residual variances constrained to equality across countries.[4] We implemented strict invariance (instead of scalar invariance) because allowing residual variances to vary across countries can obfuscate noninvariance in loadings and thresholds (Lubke & Dolan, 2003). Noninvariance was determined using the classification system developed by Educational Testing Service, which categorizes items into no or negligible, small to moderate, and moderate to large noninvariance (Zieky, 1993). Transformed into the metric of item response models, the cutoffs for moderate to large noninvariance are 0.25 for factor loadings and 0.375 for thresholds (Wetzel et al., 2017). We iteratively freed parameters (loadings and thresholds) with moderate to large noninvariance in the order of the size of their modification indices. That is, in the fully constrained model we determined which one of the parameters with moderate to large noninvariance had the largest modification index. We then estimated the first partial invariance model in which we freely estimated this parameter in the country in which it showed noninvariance while constraining it to equality across the other two countries. Next, we determined which one of the remaining parameters with moderate to large noninvariance now had the largest modification index and freed it for the second partial invariance model, and so on. The advantage of this procedure for testing measurement invariance is that it is based on an effect size criterion for noninvariance, rather than significance testing (Wetzel et al., 2017). Therefore, large sample sizes and differing sample sizes across countries should not affect the results.

All analyses were conducted using unweighted least squares with mean- and variance-corrected Satorra–Bentler goodness-of-fit tests (denoted ULSMV in M*plus*). ULSMV utilizes full information maximum likelihood estimation for item parameters such as intercepts or thresholds. Therefore, all available data are used for the estimation of these parameters. For the estimation of correlations and covariances, missingness is dealt with by pairwise deletion. ULSMV like full information maximum likelihood estimation yields consistent estimates when the data are missing at random (Asparouhov & Muthén, 2010). The narcissism facets were modeled simultaneously in one model and allowed to correlate. We used the estimates of latent mean differences between countries from the final partial measurement invariance model to investigate whether the countries differed on the narcissism facets. We divided the latent mean difference by the square root of the variance of the latent mean difference to obtain Cohen's *d* values for the differences between countries. In addition to these analyses at the facet level, we also conducted the same analyses at the level of higher-order domains: grandiosity and vulnerability for the B-PNI and overall narcissism for the NPI and NARQ. We used M*plus* version 7.4 (Muthén & Muthén, 1998-2017) to estimate the models and the R (R Core Team, 2013) package MplusAutomation (Hallquist & Wiley, 2018) to extract modification indices and determine noninvariant parameters. The analyses scripts are available from https://osf.io/gdfa6/.

Results

In the following, we report the results of our measurement invariance analyses across the United States, the United Kingdom, and Germany as well as mean differences between participants from these countries. Results from the analyses including Italy and Poland are available in SOM 2.

B-PNI

We first checked configural invariance using the theoretical factor structure from Schoenleber et al. (2015) across the United States, the United Kingdom, and Germany. The fit of the configural model was good according to the root mean square error of approximation (RMSEA) [0.035], 90% confidence interval [CI: 0.033, 0.037]) and standardized root mean residual [SRMR] (0.041) and acceptable according to the confirmatory fit index (CFI; 0.941) and Tucker–Lewis index (TLI; 0.932). The pattern of factor loadings indicated that the items loaded strongly on the facet they belonged to in all countries (see Table S3 in SOM 1).

The measurement invariance analyses of the B-PNI resulted in six noninvariant loadings for Germany while all loadings were invariant between the United States and the United Kingdom (see Table S4 in SOM 1). The largest difference in factor loadings occurred for Item 28 on the facet *hiding the self* ("When others get a glimpse of my needs, I feel anxious and ashamed"). Here the unstandardized factor loading was much larger in the German sample (2.18) compared with the combined U.S. and U.K. samples (1.35), indicating that the item was more strongly related to the trait for German participants. With 28 items and a 6-point rating scale, there were 140 thresholds in each country or 420 in total. Of these 420 thresholds, 32 (8%) were noninvariant, most of them (29) between Germany and the combined U.S. and U.K. samples. In many cases, multiple thresholds were noninvariant on the same item. For example, four (out of five) thresholds on Item 9 on the facet *devaluing* ("When others don't meet my expectations, I often feel ashamed about what I wanted") differed between the German sample and the combined U.S. and U.K. samples with all of them having higher values in the German sample. This indicates that Germans needed a higher trait level on devaluing to endorse a certain response category or the ones above compared with participants from the United States or the United Kingdom. For instance, U.S. and U.K. participants had a probability of 50% of endorsing Categories 5 or 6 at a trait level of 1.69, whereas for German participants, this was the case at a trait level of 2.52.

In sum, the United States and the United Kingdom largely showed full measurement invariance on the B-PNI. There were some violations of measurement invariance between Germany and the United States and the United Kingdom, but partial measurement invariance was achieved, allowing us to investigate mean differences across countries.

The latent mean differences between the United States and the other two countries derived from the final partial measurement invariance model showed that the United States and the United Kingdom did not show mean-level differences on most of the B-PNI's facets with the exception of self-sacrificing self-enhancement, which was lower in the United Kingdom, $d = -0.35$, 95% CI [−0.47, −0.23]; see Figure 1 and Table 2. German participants on average showed lower levels than the United States on several facets, most notably self-sacrificing self-enhancement with a large effect size, $d = -0.81$, 95% CI [−0.89, −0.73]. Mean differences on the other facets were small to moderate, for example hiding the self with $d = -0.32$, 95% CI [−0.40, −0.24], or contingent self-esteem with $d = -0.29$, 95% CI [−0.37, −0.20].

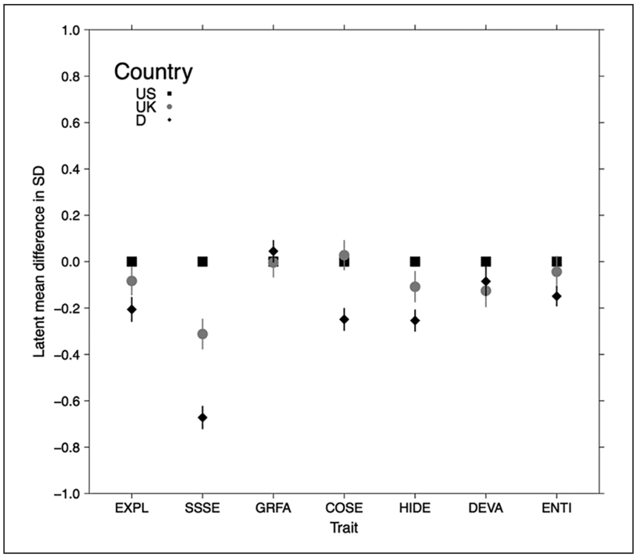

Figure 1. Latent mean differences on the B-PNI facets exploitativeness (EXPL), self-sacrificing self-enhancement (SSSE), grandiose fantasy (GRFA), contingent self-esteem (COSE), hiding the self (HIDE), devaluing (DEVA), and entitlement rage (ENTI) between the United States, the United Kingdom, and Germany (D).
Note. The means in the United States were fixed to 0 for identification and are included here only as a reference point. The mean estimates of the other countries indicate the difference to the United States. Error bars show ±1 standard error of the estimated mean difference.

We also checked whether there were notable fluctuations in the estimates of mean differences over the course of the measurement invariance models from strict invariance to the final partial invariance model. As Figures S1 to S7 in SOM 1 show, this was not the case. For facets with no noninvariant parameters such as self-sacrificing self-enhancement, the estimate stayed the same over the course of all models (Figure S2 in SOM 1). For facets with some noninvariant parameters such as exploitativeness for the German sample, the estimate changed when the noninvariance was taken into account by freeing parameters, resulting in an adjusted estimate of the mean difference. In the example of the mean difference between the United States and Germany on exploitativeness, the estimate from the strict invariance model would have underestimated the mean difference (d =−0.06 as opposed to −0.22 in the final partial measurement invariance model). A similar conclusion can be drawn when

Table 2. Latent Mean Differences and Effect Sizes for the Brief Pathological Narcissism Inventory.

	USA–UK			USA–Germany		
Trait	*M* [CI]	*SD*	Cohen's *d* [CI]	*M* [CI]	*SD*	Cohen's *d* [CI]
Exploitativeness	−0.08 [−0.2, 0.04]	0.87	−0.10 [−0.21, 0.02]	−0.21 [−0.31, −0.11]	0.93	−0.22 [−0.3, −0.14]
Self-sacrificing self-enhancement	−0.31 [−0.44, −0.19]	0.90	−0.35 [−0.47, −0.23]	−0.67 [−0.77, −0.58]	0.83	−0.81 [−0.89, −0.73]
Grandiose fantasy	0 [−0.13, 0.12]	0.97	0 [−0.12, 0.11]	0.05 [−0.05, 0.14]	0.91	0.05 [−0.03, 0.13]
Contingent self-esteem	0.03 [−0.1, 0.15]	0.94	0.03 [−0.09, 0.15]	−0.25 [−0.34, −0.16]	0.87	−0.29 [−0.37, −0.2]
Hiding the self	−0.11 [−0.24, 0.02]	0.94	−0.12 [−0.23, 0]	−0.25 [−0.34, −0.17]	0.79	−0.32 [−0.4, −0.24]
Devaluing	−0.13 [−0.26, 0.01]	0.94	−0.13 [−0.25, −0.01]	−0.09 [−0.21, 0.04]	1.17	−0.07 [−0.15, 0.01]
Entitlement rage	−0.04 [−0.17, 0.08]	0.99	−0.04 [−0.16, 0.08]	−0.15 [−0.23, −0.07]	0.76	−0.2 [−0.28, −0.12]

Note. CI =confidence interval.

comparing the observed mean difference on exploitativeness with the latent mean difference from the final partial invariance model (see Table S5 in SOM 1). Here, the observed mean difference between the United States and Germany resulted in a Cohen's d of −0.11 instead of the −0.22 found when noninvariance was corrected for. However, the bias in the observed mean differences can also lead to an overestimation, as seen when comparing the observed mean difference on hiding the self (d = −0.46) with the latent mean difference from the final partial invariance model (d = −0.32).

Finally, as an additional validity check, we compared the correlations between the B-PNI facets and age and gender across countries, both for the full invariance model (not adjusted for noninvariance) and the final partial invariance model (adjusted for noninvariance). The correlations were very similar across countries and models (see Table S6 in SOM 1). For example, self-sacrificing self-enhancement correlated −0.22 with age in the United States, −0.27 in the United Kingdom, and −0.23 in Germany in the final partial invariance model.

Thus, regarding the B-PNI facets, we found that Americans on average scored higher than individuals from the United Kingdom and Germany on the agentic facet self-sacrificing self-enhancement. Americans also scored higher than Germans on facets with neurotic content (hiding the self, contingent self-esteem).

At the level of the higher-order domains grandiosity and vulnerability, a similar number of parameters was noninvariant (3 loadings and 37 thresholds) as at the facet level, again mostly between Germany and the combined U.S. and U.K. samples (see Table S1 in SOM 3). Latent mean differences on grandiosity and vulnerability reflected those found for the facets and showed that Americans on average scored higher than participants from the United Kingdom and Germany on grandiosity. Americans also on average scored higher than Germans on vulnerability (see Figure S1 in SOM 3).

NPI

For the NPI, the configural invariance model showed a good fit according to the RMSEA, 0.021, 90% CI [0.020, 0.023], a just acceptable fit according to the SRMR (0.080), and a below acceptable fit according to the CFI (0.892) and the TLI (0.883). The pattern of factor loadings indicated that there were some items that showed poor factor loadings for all countries and that some factor loadings differed strongly across countries (see Table S7 in SOM 1). For example, the standardized factor loading on the vanity item 20 ("I try not to be a show off" vs. "I am apt to show off if I get the chance") was 0.43 in the U.S. sample, 0.14 in the U.K. sample, and 0.22 in the German sample. This indicates that there might be differences in the factor structure across countries. Nevertheless, since the model overall fit acceptably, and considering that the factor structure of the NPI has been a disputed topic even in homogeneous samples (Ackerman et al., 2011; Emmons, 1984), we proceeded with the measurement invariance analyses.

The measurement invariance analyses indicated that 21 parameters were noninvariant (8 loadings and 13 [12%] thresholds). With two exceptions, these noninvariant parameters pertained to the German sample, indicating that the United States and the United Kingdom were largely invariant while Germany required its own loading or threshold on a number of items. For example, Item 4 on vanity ("When people compliment me I sometimes get embarrassed" vs. "I know that I am good because everybody keeps telling me so") had an unstandardized loading of 1.51 in the combined U.S. and U.K. group and a lower loading in Germany (0.89). Thus, for eight NPI items, the relationship between the items and the underlying trait was different for the German sample compared with the combined U.S. and U.K. samples. With respect to the thresholds, a number of items also differed between Germany and the combined U.S. and U.K. samples. For example, Item 1 on leadership ("I have a natural talent for influencing people" vs. "I am not good at influencing people") had a threshold of −0.35 in the combined

U.S. and U.K. samples, but a threshold of −0.96 in the German sample. This indicates that—conditional on the trait level—German participants were more likely to choose the narcissistic option ("I have a natural talent for influencing people") compared with U.S. and U.K. participants. In fact, 73% of the German participants selected the narcissistic option compared with 59% in the United States and 52% in the United Kingdom. Table S8 in SOM 1 contains a list of the noninvariant NPI items for each country. In sum, for the United States and the United Kingdom, the NPI functioned largely equivalently. For Germany, there were some violations of measurement invariance.

In the final partial invariance model, mean differences relative to the United States as the reference group were found for the United Kingdom on leadership and vanity, with U.K. participants on average scoring lower than U.S. participants, $d = -0.37$, 95% CI [−0.49, −0.25], for leadership and $d = -0.25$, 95% CI [−0.37, −0.13], for vanity (see Figure 2 and Table 3). German participants showed slightly lower levels than the United States on leadership, $d = -0.28$, 95% CI [−0.36, −0.21], and higher levels on vanity, $d = 0.37$, 95% CI [0.30, 0.45], and entitlement, $d = 0.21$, 95% CI [0.13, 0.28]. The development of the estimates of mean differences from the full invariance model to the final partial invariance model is depicted in Figures S8 to S10 in SOM 1. Compared with the B-PNI, there were larger differences in mean estimates between the first and last model. For instance, in the strict invariance model, the German sample showed higher mean leadership levels than the U.S. sample ($d = 0.30$), which dropped to a negative mean difference over the course of the invariance models (final $d = -0.28$). Similarly, the observed mean difference also indicated a higher mean for the German sample on leadership ($d = 0.25$; see Table S5 in SOM 1). Correlations of the NPI facets with age and sex did not differ notably between the strict and final partial invariance models (Table S6 in SOM 1). Furthermore, correlations were very similar across countries, though in some cases smaller for the German sample compared with the U.S. and U.K. samples (e.g., the correlation of vanity with age was −0.16 in the United States, −0.15 in the United Kingdom, and −0.04 in Germany). In sum, the United States showed a higher leadership mean than the United Kingdom and Germany. On vanity, U.S. participants on average scored higher than U.K. participants, but lower than German participants. However, considering the larger number of items with noninvariance compared with the B-PNI, these results should be interpreted cautiously.

At the level of overall narcissism, fewer loadings were noninvariant (3 instead of 8). Furthermore, 15 thresholds were noninvariant, 13 of those for Germany (see Table S2 in SOM 3). The latent mean differences on overall narcissism indicated a higher mean for Germany compared with the United States and the United Kingdom (see Figure S2 in SOM 3).

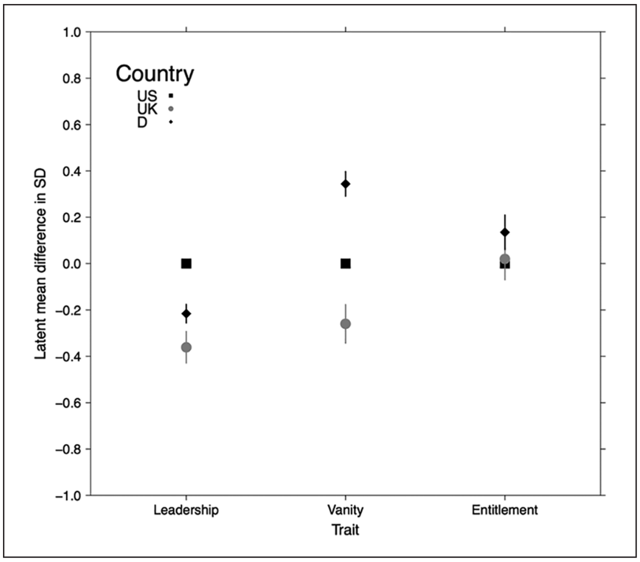

Figure 2. Latent mean differences on the Narcissistic Personality Inventory facets leadership, vanity, and entitlement between the United States (US), the United Kingdom (UK), and Germany (D).

Note. The means in the United States were fixed to 0 for identification and are included here only as a reference point. The mean estimates of the other countries indicate the difference to the United States. Error bars show ±1 standard error of the estimated mean difference.

NARQ

The fit of the configural invariance model with the factor structure from Back et al. (2013) was acceptable to good (RMSEA = 0.047, 90% CI [0.044, 0.049], SRMR = 0.051, CFI = 0.919, TLI = 0.905). All items showed substantial factor loadings on their respective facet in all countries (see Table S9 in SOM 1).

The measurement invariance analyses revealed 29 noninvariant parameters, all of them for the German sample, while the U.S. and U.K. samples were fully invariant (see Table S10 in SOM 1). Of these 29 noninvariant parameters, three were factor loadings. For example, Item 8 ("I deserve to be seen as a great personality") was more strongly related to the trait admiration for U.S. and U.K. participants (unstandardized loading 1.67) than for participants from Germany (unstandardized loading 1.05). Of the 270 thresholds (5 thresholds for each of the 18 items times 3 countries), 26 (10%) were noninvariant for the German sample. Eleven items in total were affected (five on admiration and six on rivalry), though only for five of these items three thresholds or more were noninvariant. For example, four out of the five thresholds of Item 18 on admiration ("Mostly, I am very adept at dealing with other people") were noninvariant for the German sample. These thresholds all had higher values in the German sample compared with the combined U.S. and U.K. samples, indicating that Germans

Table 3. Latent Mean Differences and Effect Sizes for the Narcissistic Personality Inventory.

	USA–UK			USA–Germany		
Trait	*M* [CI]	*SD*	Cohen's *d* [CI]	*M* [CI]	*SD*	Cohen's *d* [CI]
Leadership	−0.36 [−0.49, −0.23]	0.98	−0.37 [−0.49, −0.25]	−0.22 [−0.3, −0.14]	0.77	−0.28 [−0.36, −0.21]
Vanity	−0.26 [−0.42, −0.1]	1.03	−0.25 [−0.37, −0.13]	0.37 [0.27, 0.48]	1.00	0.37 [0.3, 0.45]
Entitlement	0.03 [−0.15, 0.2]	0.93	0.03 [−0.09, 0.15]	0.21 [0.07, 0.35]	1.02	0.21 [0.13, 0.28]

Note. CI =confidence interval.

needed a higher trait level to have the same probability of endorsing a certain response category as people from the other countries. All thresholds of Item 14 ("Other people are worth nothing") were noninvariant for the German sample, but this item had a very skewed response distribution in all samples (fewer than 15% of the participants responded in Categories 4, 5, or 6), which may have affected the estimation of the thresholds.

In sum, the NARQ was equivalent between the U.S. and U.K. samples. The German sample differed in their endorsement probabilities of some response categories on some items, but partial invariance with the U.S. and U.K. samples existed.

Figure 3 and Table 4 show latent mean differences on admiration and rivalry from the final partial invariance model. The U.K. and German participants had lower average trait levels than U.S. participants on admiration, $d = -0.29$, 95% CI [−0.41, −0.17], for the United Kingdom and $d = -0.29$, 95% CI [−0.37, −0.21], for Germany. Mean trait levels on rivalry did not differ between the United States and the United Kingdom, while Germany showed a lower average rivalry level, $d = -0.33$, 95% CI [−0.41, −0.26]. Figures S11 and S12 in SOM 1 show how the estimates of latent mean differences developed over the course of the partial invariance models. Mean estimates for Germany fluctuated slightly, but overall did not differ much from those in the full invariance model. The observed mean differences on admiration were very similar to those from the final partial invariance model (Table S5 in SOM 1). However, on rivalry, the observed means underestimated the difference between the United States and Germany ($d = -0.22$ vs. $d = -0.33$, in the final partial invariance model). Correlations between the NARQ facets and age and sex were very similar across models and countries (Table S6 in SOM 1).[5] In sum, the United States showed higher levels than the United Kingdom and Germany on the agentic facet admiration. The United States also showed higher levels than Germany on the antagonistic facet rivalry.

At the level of overall narcissism, the same number of noninvariant parameters was found (29). Twenty-eight of these pertained to thresholds (see Table S3 in SOM 3). According to the final partial invariance model for the NARQ, Americans scored higher on overall narcissism than individuals from the United Kingdom and Germany (see Figure S3 in SOM 3).

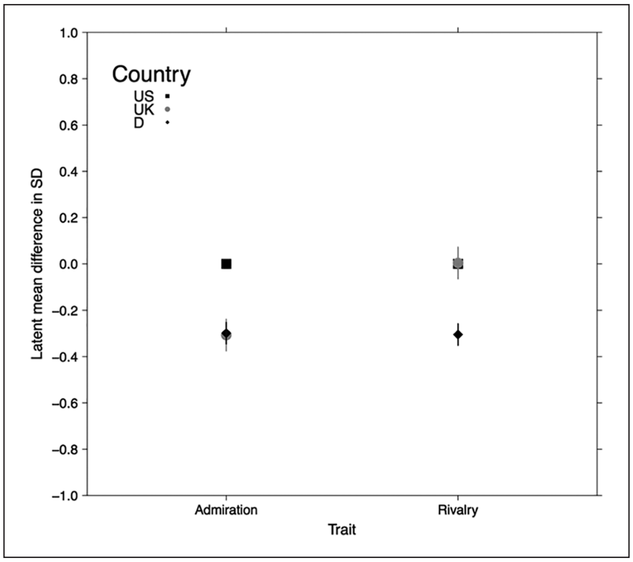

Figure 3. Latent mean differences on the Narcissistic Admiration and Rivalry Questionnaire facets admiration and rivalry between the United States (US), the United Kingdom (UK), and Germany (D).

Note. The means in the United States were fixed to 0 for identification and are included here only as a reference point. The mean estimates of the other countries indicate the difference to the United States. Error bars show ±1 standard error of the estimated mean difference.

Discussion

In this study, we investigated the measurement invariance of three narcissism questionnaires (B-PNI, NPI, NARQ) across the United States, the United Kingdom, and Germany. The three narcissism questionnaires functioned mostly equivalently in the U.S. and U.K. samples, indicating that mean comparisons between these two countries could be drawn without qualifications. Comparisons between the United States and Germany required adjustments for noninvariance. In the following, we first discuss the measurement invariance or lack thereof of the B-PNI, NPI, and NARQ across countries. Second, we discuss potential reasons for noninvariance

Table 4. Latent Mean Differences and Effect Sizes for the Narcissistic Admiration and Rivalry Questionnaire.

	USA–UK			USA–Germany		
Trait	*M* [CI]	*SD*	Cohen's *d* [CI]	*M* [CI]	*SD*	Cohen's *d* [CI]
Admiration	−0.31 [−0.44, −0.17]	1.05	−0.29 [−0.41, −0.17]	−0.3 [−0.39, −0.21]	1.03	−0.29 [−0.37, −0.21]
Rivalry	0 [−0.13, 0.14]	0.98	0 [−0.11, 0.12]	−0.31 [−0.4, −0.22]	0.91	−0.33 [−0.41, −0.26]

Note. CI =confidence interval.

in a cross-cultural context and the implications for translating and adapting measures. Third, we discuss the mean differences we found and the practical implications of measurement invariance for interpreting mean differences before noting some limitations of our study and future directions.

Measurement Invariance of the B-PNI, NPI, and NARQ Across Countries

We found the fewest violations of measurement invariance between the United States, the United Kingdom, and Germany in the B-PNI (7.5% of factor loadings and thresholds), followed by the NARQ (9%). The NPI showed the largest number of violations (13.3%), many of which pertained to factor loadings.[6] In fact, the NPI was the sole questionnaire with a preponderance of metric nonequivalent items. This is a more severe violation of measurement invariance than differences in item thresholds because it may indicate that items are relevant to a trait in one country, but not another (Huang et al., 1997). For example, the item pair "I insist on getting the respect that is due me" versus "I usually get the respect I deserve" loaded higher in the United States/the United Kingdom than in Germany. Thus, this item was less relevant to the entitlement facet in the German sample compared with the U.S. and U.K. samples.

In contrast to the NPI, in the B-PNI and the NARQ, we mainly found violations of measurement invariance at the level of the thresholds. Almost all of the noninvariant thresholds pertained to the German sample. This may indicate that the items measure aspects of the trait that are expressed to different degrees in Germany compared with the United States and the United Kingdom, leading to different probabilities of endorsing the items. For example, U.S. participants had a higher probability of endorsing a category stating agreement compared with German participants on Item 18 on admiration in the NARQ ("Mostly, I am very adept at dealing with other people"). Despite the noninvariance in loadings and thresholds found in all questionnaires, partial invariance across the United States, the United Kingdom, and Germany could be established for all traits.

Potential Reasons for Noninvariance

Why did some items show noninvariance across countries? According to *The ITC Guidelines for Translating and Adapting Tests* (International Test Commission, 2017), cross-cultural noninvariance can be due to 1) "translation nonequivalence that occurs from source to target language versions" and 2) "cultural contextual differences" (International Test Commission, 2017, p. 19). Of our three narcissism questionnaires, the B-PNI and the NPI were developed in English while the NARQ was simultaneously developed in German and English. Thus, the other versions of these questionnaires applied in this study were translations. Deviations from the original in translation, which can be translation errors or intentional cultural adaptations of the items, can have an impact on factor loadings and item thresholds. For example, Item 8 on leadership in the original English NPI reads "I will be a success" versus "I am not too concerned about success." In the German version by Schütz et al. (2004), the item reads *Ich will erfolgreich sein* versus *Mir ist Erfolg nicht besonders wichtig*. Importantly, the narcissistic response option is not an exact translation of "I will be a success" but rather says "I want to be successful," which is arguably a much weaker statement than "I will be a success." This might explain why 77% of German participants chose the narcissistic option of this item compared with only 57% in the U.S. sample, despite German participants overall scoring lower on leadership than American participants. Thus, this translation difference may have led to the noninvariance in the threshold of this item: it had a threshold of −0.05 in the combined U.S. and U.K. samples, but a threshold of −0.96 in the German sample, indicating that—conditional on the trait level—German participants were much more likely to select the narcissistic option compared with U.S. and U.K. participants.

Sometimes it is necessary to adapt the item content because the item content is not relevant in a different country or culture, it contains a cultural reference, it contains an expression that cannot be translated literally, or it contains an idiom (International Test Commission, 2017). For example, the narcissistic response option for Item 16 on the NPI is "I can read people like a book." This idiom may not exist in other languages or the same meaning might be expressed with a different analogy. In German, "to be an open book for someone" is a common expression, but "reading someone like a book" is not. In the German translation, this item was adapted to "I can read in others like in a book" (*Ich kann in anderen wie in einem Buch lesen*). Thus, the intricacies of language often make it necessary to adapt the item content, but this can come at the cost of reduced comparability, which can manifest itself in noninvariance. This implies that when instruments are translated and adapted to other languages or cultures, checking measurement invariance with the original version should be an integral part of the process. If a lack of measurement invariance is discovered at this stage, measures can be taken to improve the equivalence (e.g., by revising the translation). Later on, it might be too late. If a new instrument is developed with the goal of applying it in multiple languages, items can be tested for measurement invariance during preliminary studies, and only items that are equivalent can be selected for the final version of the instrument. For example, the instrument development process for the Programme for International Student Assessment involves a double translation design from two source languages, standardized procedures and guidelines for translating and adapting material, and an international verification of the national versions (Organisation for Economic Cooperation and Development [OECD], 2012). This extensive process applied in international large-scale assessments in the educational domain could be a model for the translation and adaption of instruments in personality or clinical psychology. Policy makers around the globe are looking for new indicators to describe social progress and quality of life beyond Gross Domestic Product (e.g., *well-being* as part of the OECD's Better Life Initiative; OECD, 2019). If these psychometric assessments are to be used in cross-country comparisons, they need to fulfil the highest possible standards of data quality, and measurement invariance will be one of the building blocks.

Mean Differences in Narcissism

Are Americans more narcissistic than people from other countries? Previous research using observed scores indicated that this was the case, with Americans, for example, scoring higher than Europeans on the NPI total score (Foster et al., 2003). According to our analyses, Americans consistently scored higher than participants from the United Kingdom and Germany on facets capturing agentic narcissistic content, such as self-sacrificing self-enhancement (B-PNI), leadership (NPI), and admiration (NARQ). Americans also scored higher than Germans, but not individuals from the United Kingdom, on B-PNI and NARQ facets capturing antagonistic and neurotic content (hiding the self, contingent self-esteem, rivalry). These results might be explained by differences between the countries on Hofstede's (2001) cultural dimension of individualism (vs. collectivism), arguably the most relevant cultural dimension to narcissism. According to Hofstede (2001), the United States had a score of 91 on individualism, while Great Britain had a score of 89 and Germany (the former West) only had a score of 67. Thus, values such as looking after oneself rather than relying on a group play a more important role in the United States compared with other countries and this might foster narcissistic tendencies. Results for two of the NPI facets, vanity and entitlement, were less consistent with this overall pattern with Germans scoring lower than Americans on these two facets. This is especially puzzling for B-PNI-entitlement rage and NPI-entitlement, which showed opposite results (d = −0.20 and d = 0.21, respectively) although they showed a strong (observed) correlation of 0.59 in the U.S. sample. This can be due to different conceptualizations of the entitlement facet in these two questionnaires since B-PNI-entitlement rage is part of vulnerability whereas the NPI measures grandiose narcissism. Also, 179 participants in the German sample had missing values on the B-PNI due to dropout in the course of the online survey, so the composition of the sample differed between the NPI and the B-PNI analysis. Furthermore, there were more issues with noninvariance in the NPI and the NPI has been criticized for a number of reasons, including its unclear factor structure and that some item pairs consist of items measuring different aspects of narcissism (Ackerman et al., 2016; Wetzel, Roberts, et al., 2016). Results at the higher-order level were consistent with the facet level for the B-PNI, with Americans scoring higher than Germans and participants from the United Kingdom on grandiosity and Americans scoring higher than Germans on vulnerability. Similarly, Americans scored higher than participants from Germany and the United Kingdom on overall narcissism in the NARQ. The NPI result was again inconsistent with the other two questionnaires with Germans showing the highest mean on overall narcissism. However, as the facet-level analysis showed, means did not differ on all facets and more differentiated patterns across countries were revealed. Thus, whether mean differences on narcissism exist depends on the specific aspect of narcissism. This implies that it is important to distinguish different components of narcissism, rather than comparing only total scores.

Practical Implications of Measurement Invariance for the Interpretation of Mean Scores

Our analyses also illustrate the importance of taking measurement noninvariance into account when interpreting mean scores and comparing them across countries: For those facets on which a number of items were noninvariant, observed means underestimated or overestimated the differences between countries. For example, for the B-PNI facet self-sacrificing self-enhancement, the observed mean difference was smaller than the latent mean difference from the final partial invariance model, whereas for the facet hiding the self, the observed mean difference was larger than the latent mean difference. Since there is no way of knowing in which direction the bias will go, researchers relying on observed means may draw incorrect conclusions about cross-country differences. Therefore, the measurement invariance of the instrument across countries should be investigated prior to drawing mean comparisons and potential noninvariance should be controlled for.

Limitations and Future Directions

Even though we tested for measurement invariance and controlled for noninvariance in the partial invariance models, we should be cautious in interpreting these mean differences in narcissism facets because there are other potential reasons that could have played a role in addition to true cross-country differences. We made the age distribution of the samples more similar by restricting the age range from 18 to 50 years, but, since mean-level changes in narcissism occur from young adulthood to middle age (Wetzel et al., 2019) and cross-cohort measurement noninvariance has been found for the NPI (Wetzel et al., 2017), it is still possible that differences between age groups/cohorts may have been confounded with differences between countries. Nevertheless, correlations between narcissism facets and age and gender were similar across countries, both for unadjusted trait estimates and adjusted trait estimates (taking measurement noninvariance into account). In addition, method biases such as differences in using the response scales (e.g., acquiescence, extreme response style) could have influenced the results though the impact of response styles appears to be less severe than is often assumed (Plieninger, 2017; Wetzel, Böhnke, et al., 2016). Last, taking measurement invariance into account when investigating mean differences across countries cannot control for the reference-group effect (Heine et al., 2008; Mottus et al., 2012).

Our samples were from mostly Western countries. Thus, future research could examine the equivalence of narcissism questionnaires and mean differences in narcissism in more diverse countries, including countries from Asia and Africa. With more diverse countries, different patterns of results might emerge with respect to the different components of narcissism (e.g., more individualistic countries scoring higher than more collectivistic countries on agentic aspects of narcissism). Our samples were all nonclinical samples. Since the B-PNI was developed for the assessment of pathological narcissism, it would be interesting to investigate the cross-cultural measurement invariance of the B-PNI in clinical samples and to include other questionnaires such as the Narcissistic Vulnerability Scale (Crowe et al., 2018).

Conclusion

Questionnaires are translated and adapted to other cultures with the intent to minimize or eliminate cultural differences in responding to the items. A test of measurement invariance can be used to check whether that goal was achieved. Configural invariance overall supports the use of the B-PNI and NARQ in the countries investigated here, though it is questionable for the NPI. All questionnaires showed some noninvariance across countries, indicating that caution needs to be exercised when investigating and interpreting mean differences. In line with stereotypical perceptions, we found that individuals from the United States on average scored higher on agentic facets of narcissism than individuals from the United Kingdom and Germany. For antagonistic and neurotic facets, there were largely no differences between the United States and the United Kingdom. The United States showed higher means than Germany on some, but not all, facets with antagonistic and neurotic content.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a grant from the Mentoring Program of the Zukunftskolleg at the University of Konstanz awarded to Eunike Wetzel and a grant from the National Science Center Poland awarded to Radoslaw Rogoza (2014/14/M/HS6/00919).

ORCID iDs

Eunike Wetzel https://orcid.org/0000-0002-4224-0366
Michele Vecchione https://orcid.org/0000-0002-8907-9872
Radoslaw Rogoza https://orcid.org/0000-0002-4983-9320

Supplemental Material

Supplemental material for this article is available online from https://osf.io/53amg/.

Notes

1. A subsample of the U.S. and U.K. data were analyzed in Study 3 of Wetzel, Leckelt, et al. (2016). This subsample consisted of the 971 participants who had completed the online survey by September 8, 2015. It also included participants from other English-speaking countries such as Canada and Australia. Data collection continued for the purposes of this study. Wetzel, Leckelt, et al. (2016) investigated whether latent classes of narcissists could be distinguished using the NARQ data. Thus, there is no overlap with the research questions or analyses of this study.
2. The Italian and Polish samples are described in SOM 2.
3. We used the M*plus* parameterization with factor loadings instead of item discrimination parameters.
4. For residual variances, this means that all of them were fixed to 1 because they have to be fixed to 1 in one of the groups for model identification.
5. To obtain an index of overall similarity across the correlations with age and sex between countries, we correlated the

Fisher-*Z*-transformed correlations (e.g., correlations of trait estimates on all facets from the partial invariance models with age) between pairs of countries and retransformed these into correlations. The correlations ranged from 0.63 to 0.73 ($M = 0.68$), indicating overall high similarity. In addition, we checked the effect sizes of the differences between pairs of correlations. For the U.S.–U.K. correlations and the U.S.–Germany correlations, the vast majority showed a negligible difference (i.e., $<|.10|$) and only 6 (U.S.–U.K.) or 7 (U.S.–Germany) showed a small difference (i.e., $>|.10|$), though none of the differences was $>|.20|$. About half of the U.K.–Germany correlations showed a small difference, but again none was $>|.20|$. Thus, overall, correlations of the narcissism facets with age and gender were similar across countries.

6. The NPI was also the only questionnaire in which configural invariance was questionable when additionally including data from Italy, with the pattern of factor loadings differing substantially across countries.

References

Ackerman, R. A., Donnellan, M. B., Roberts, B. W., & Fraley, R. C. (2016). The effect of response format on the psychometric properties of the Narcissistic Personality Inventory: Consequences for item meaning and factor structure. *Assessment*, *23*(2), 203-220. https://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. https://doi.org/10.1177/1073191110382845

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Author.

Asparouhov, T., & Muthén, B. (2010). *Weighted least squares estimation with missing data*. https://www.statmodel.com/download/GstrucMissingRevision.pdf

Back, M. D. (2018). The Narcissistic Admiration and Rivalry Concept. In A. D. Hermann, A. Brunell, & J. Foster (Eds.), *The handbook of trait narcissism: Key advances, research methods, and controversies* (pp. 57-67). Springer. https://doi.org/10.1007/978-3-319-92171-6_6

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. https://doi.org/10.1037/a0034431

Back, M. D., & Morf, C. C. (in press). Narcissism. In V. Zeigler-Hill, & T. K. Shackelford (Eds.), *Encyclopedia of personality and individual differences*. Springer.

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

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456-466. https://doi.org/10.1037/0033-2909.105.3.456

Campbell, W. K., Miller, J. D., & Buffardi, L. E. (2010). The United States and the "Culture of Narcissism": An examination of perceptions of national character. *Social Psychological and Personality Science*, *1*(3), 222-229. https://doi.org/10.1177/1948550610366878

Crowe, M. L., Edershile, E. A., Wright, A. G. C., Campbell, W. K., Lynam, D. R., & Miller, J. D. (2018). Development and validation of the Narcissistic Vulnerability Scale: An adjective rating scale. *Psychological Assessment*, *30*(7), 978-983. https://doi.org/10.1037/pas0000578

Crowe, M. L., Lynam, D. R., Campbell, W. K., & Miller, J. D. (2019). Exploring the structure of narcissism: Toward an integrated solution. *Journal of Personality*, *87*(6), 1151-1169. https://doi.org/10.1111/jopy.12464

Doroszuk, M., Kwiatkowska, M. M., Torres-Marin, J., Navarro-Carrillo, G., Wlodarczyk, A., Blasco-Belled, A., Martínez-Buelvas, L., Newton, J. D. A., Oviedo-Trespalacios, O., & Rogoza, R. (2020). Construct validation of the Narcissistic Admiration and Rivalry Questionnaire in Spanish-speaking countries: Assessment of the reliability, structural and external validity and cross-cultural equivalence. *International Journal of Psychology*, *55*(3), 413-424. https://doi.org/10.1002/ijop.12595

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

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. https://doi.org/10.1016/S0092-6566(03)00026-6

Fukunishi, I., Nakagawa, T., Nakamura, H., Li, K., & Hua, Z. Q. (1996). Relationships between Type A behavior, narcissism, and maternal closeness for college students in Japan, the United States of America, and the People's Republic of China. *Psychological Reports*, *78*(3), 939-944. https://doi.org/10.2466/pr0.1996.78.3.939

Gentile, B., Miller, J. D., Hoffman, B. J., Reidy, D. E., Zeichner, A., & Campbell, W. K. (2013). A test of two brief measures of grandiose narcissism: The Narcissistic Personality Inventory-13 and the Narcissistic Personality Inventory-16. *Psychological Assessment*, *25*(4), 1120-1136. https://doi.org/10.1037/a0033192

Grosz, M. P., Lösch, T., & Back, M. D. (2017). The narcissism-overclaiming link revisited. *Journal of Research in Personality, 70*(October), 134-138. https://doi.org/10.1016/j.jrp.2017.05.006

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology*, *5*, 980. https://doi.org/10.3389/fpsyg.2014.00980

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in M*plus. Structural Equation Modeling*, *25*(4), 621-638. https://doi.org/10.1080/10705511.2017.1402334

Heine, S. J., Buchtel, E. E., & Norenzayan, A. (2008). What do cross-national comparisons of personality traits tell us? The case of conscientiousness. *Psychological Science*, *19*(4), 309-313. https://doi.org/10.1111/j.1467-9280.2008.02085.x

Hofstede, G. (2001). *Culture's consequences*. Sage.

Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits: Differential item functioning in the NEO personality inventory. *Journal of Cross-Cultural Psychology*, *28*(2), 192-218. https://doi.org/10.1177/0022022197282004

International Test Commission. (2017). *The ITC guidelines for translating and adapting tests*. http://www.intestcom.org/

Jonason, P. K., Foster, J., Oshio, A., Sitnikova, M., Birkas, B., & Gouveia, V. (2017). Self-construals and the Dark Triad traits in six countries. *Personality and Individual Differences, 113*(July), 120-124. https://doi.org/10.1016/j.paid.2017.02.053

Krizan, Z., & Herlache, A. D. (2018). The narcissism spectrum model: A synthetic view of narcissistic personality. *Personality and Social Psychology Review*, *22*(1), 3-31. https://doi.org/10.1177/1088868316685018

Leckelt, M., Richter, D., Wetzel, E., & Back, M. D. (2019). Longitudinal associations of narcissism with interpersonal, intrapersonal, and institutional outcomes: An investigation using a representative sample of the German population. *Collabra: Psychology*, *5*(1), 26. https://doi.org/10.1525/collabra.248

Leckelt, M., Wetzel, E., Gerlach, T. M., Ackerman, R. A., Miller, J. D., Chopik, W. J., Penke, L., Geukes, K., Küfner, A. C. P., Hutteman, R., Richter, D., Renner, K.-H., Allroggen, M., Brecheen, C., Campbell, W. K., Grossmann, I., & Back, M. D. (2018). Validation of the Narcissistic Admiration and Rivalry Questionnaire Short Scale (NARQ-S) in convenience and representative samples. *Psychological Assessment*, *30*(1), 86-96. https://doi.org/10.1037/pas0000433

Lubke, G. H., & Dolan, C. V. (2003). Can unequal residual variances across groups mask differences in residual means in the common factor model? *Structural Equation Modeling*, *10*(2), 175-192. https://doi.org/10.1207/S15328007sem1002_1

Meisel, M. K., Ning, H., Campbell, W. K., & Goodie, A. S. (2016). Narcissism, overconfidence, and risk taking in U.S. and Chinese student samples. *Journal of Cross-Cultural Psychology*, *47*(3), 385-400. https://doi.org/10.1177/0022022115621968

Meredith, W. (1993). Measurement invariance, factor-analysis and factorial invariance. *Psychometrika*, *58*(4), 525-543. https://doi.org/10.1007/Bf02294825

Miller, J. D., Lynam, D. R., McCain, J. L., Few, L. R., Crego, C., Widiger, T. A., & Campbell, W. K. (2016). Thinking structurally about narcissism: An examination of the Five-Factor Narcissism Inventory and its components. *Journal of Personality Disorders*, *30*(1), 1-18. https://doi.org/10.1521/pedi_2015_29_177

Miller, J. D., Maples, J. L., Buffardi, L., Cai, H., Gentile, B., Kisbu-Sakarya, Y., Kwan, V. S. Y., LoPilato, A., Pendry, L. F., Sedikides, C., Siedor, L., & Campbell, W. K. (2015). Narcissism and United States' culture: The view from home and around the world. *Journal of Personality and Social Psychology*, *109*(6), 1068-1089. https://doi.org/10.1037/a0039543

Morf, C. C., Schurch, E., Kufner, A., Siegrist, P., Vater, A., Back, M., Mestel, R., & Schröder-Abe, M. (2017). Expanding the nomological net of the Pathological Narcissism Inventory: German validation and extension in a clinical inpatient sample. *Assessment*, *24*(4), 419-443. https://doi.org/10.1177/1073191115627010

Mottus, R., Allik, J., Realo, A., Pullmann, H., Rossier, J., Zecca, G., Ah-Kion, J., Amoussou-Yéyé, D., Bäckström, M., Barkauskiene, R., Barry, O., Bhowon, U., Björklund, F., Bochaver, A., Bochaver, K., de Bruin, G. P., Cabrera, H. F., Chen, S. X., Church, A. T., . . . Tseung, C. N. (2012). Comparability of self-reported conscientiousness across 21 countries. *European Journal of Personality*, *26*(3), 303-317. https://doi.org/10.1002/per.840

Muthén, L. K., & Muthén, B. O. (1998-2017). M*plus* [Computer software]. Author. www.statmodel.com

Organisation for Economic Cooperation and Development. (2012). *PISA 2009 technical report*. http://dx.doi.org/10.1787/9789264167872-en

Organisation for Economic Cooperation and Development. (2019). *OECD better life initiative: Measuring well-being and progress*. https://www.oecd.org/sdd/OECD-Better-Life-Initiative.pdf

Penney, L. M., & Spector, P. E. (2002). Narcissism and counterproductive work behavior: Do bigger egos mean bigger problems? *International Journal of Selection and Assessment*, *10*(1-2), 126-134. https://doi.org/10.1111/1468-2389.00199

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. https://doi.org/10.1037/a0016530

Plieninger, H. (2017). Mountain or molehill? A simulation study on the impact of response styles. *Educational and Psychological Measurement*, *77*(1), 32-53. https://doi.org/10.1177/0013164416636655

R Core Team. (2013). R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing. http://www.R-project.org/

Raskin, R. N., & Hall, C. S. (1979). Narcissistic Personality Inventory. *Psychological Reports*, *45*(2), 590-590. https://doi.org/10.2466/pr0.1979.45.2.590

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. https://doi.org/10.1037/0022-3514.54.5.890

Robins, R. W., Tracy, J. L., & Shaver, P. R. (2001). Shamed into self-love: Dynamics, roots, and functions of narcissism. *Psychological Inquiry*, *12*(4), 230-236.

Rogoza, R., Rogoza, M., & Wyszyńska, P. (2016). Polska adaptacja modelu narcystycznego podziwu i rywalizacji [Polish adaptation of the narcissistic admiration and rivalry model]. *Polskie Forum Psychologiczne*, *21*(3), 410-431. https://doi.org/10.14656/PFP20160306

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

Schoenleber, M., Roche, M. J., Wetzel, E., Pincus, A. L., & Roberts, B. W. (2015). Development of a brief version of the Pathological Narcissism Inventory. *Psychological Assessment*, *27*(4), 1520-1526. https://doi.org/10.1037/pas0000158

Schütz, A., Marcus, B., & Sellin, I. (2004). Die Messung von Narzissmus als Persönlichkeitskonstrukt: Psychometrische

Eigenschaften einer Lang- und einer Kurzfrom des Deutschen NPI (Narcissistic Personality Inventory) [Measuring narcissism as a personality construct: Psychometric properties of a long and a short version of the German NPI]. *Diagnostica*, *50*(4), 202-218. https://doi.org/10.1026/0012-1924.50.4.202

Steenkamp, J. B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78-90. https://doi.org/10.1086/209528

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4-70. https://doi.org/10.1177/109442810031002

Vecchione, M., Dentale, F., Graziano, M., Dufner, M., Wetzel, E., Leckelt, M., & Back, M. D. (2018). An Italian validation of the Narcissistic Admiration and Rivalry Questionnaire (NARQ): Further evidence for a two-dimensional model of grandiose narcissism. *Applied Psychology Bulletin*, *66*(281), 29-37. https://doi.org/10.1037/t68091-000

Wetzel, E., Böhnke, J. R., & Rose, N. (2016). A simulation study on methods of correcting for the effects of extreme response style. *Educational and Psychological Measurement*, *76*(2), 304-324. https://doi.org/10.1177/0013164415591848

Wetzel, E., Brown, A., Hill, P. L., Chung, J. M., Robins, R. W., & Roberts, B. W. (2017). The narcissism epidemic is dead; long live the narcissism epidemic. *Psychological Science*, *28*(12), 1833-1847. https://doi.org/10.1177/0956797617724208

Wetzel, E., Grijalva, E., Robins, R. W., & Roberts, B. W. (2019). You're still so vain; Changes in narcissism from young adulthood to middle age. *Journal of Personality and Social Psychology*, Advance online publication. https://doi.org/10.1037/pspp0000266

Wetzel, E., Leckelt, M., Gerlach, T. M., & Back, M. D. (2016). Distinguishing subgroups of narcissists with latent class analysis. *European Journal of Personality*, *30*(4), 374-389. https://doi.org/10.1002/per.2062

Wetzel, E., Roberts, B. W., Fraley, R. C., & Brown, A. (2016). Equivalence of Narcissistic Personality Inventory constructs and correlates across scoring approaches and response formats. *Journal of Research in Personality, 61*(April), 87-98. https://doi.org/10.1016/j.jrp.2015.12.002

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, & M. Windle (Eds.), *The science of prevention: Methodological advance from alcohol and substance abuse research* (pp. 281-324). American Psychological Association. https://doi.org/10.1037/10222-009

Zemojtel-Piotrowska, M., Piotrowski, J., Rogoza, R., Baran, T., Hitokoto, H., & Maltby, J. (2018). Cross-cultural invariance of NPI-13: Entitlement as culturally specific, leadership and grandiosity as culturally universal. *International Journal of Psychology*, *54*(4), 439-447. https://doi.org/10.1002/ijop.12487

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Lawrence Erlbaum.