UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

To know personality is to measure it

Introducing a Dutch brief form of the Multidimensional Personality Questionnaire

Eigenhuis, A.

Publication date
2017
Document Version
Other version
License
Other

Link to publication

Citation for published version (APA):
Eigenhuis, A. (2017). *To know personality is to measure it: Introducing a Dutch brief form of the Multidimensional Personality Questionnaire*. [Thesis, fully internal, Universiteit van Amsterdam].

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:13 Jun 2024

Chapter 3

Personality differences between the United States and the Netherlands

The influence of violations of measurement invariance

Annemarie Eigenhuis

Jan H. Kamphuis

Arjen Noordhof

An excerpt of this chapter is published as:
Eigenhuis, A., Kamphuis, J. H., & Noordhof, A. (2015). Personality differences between the United States and the Netherlands: The influence of violations of measurement invariance. *Journal of Cross-Cultural Psychology*, *46*(4), 549–564. http://doi.org/10.1177/0022022115570671

Abstract

Cross-cultural comparisons of personality have yielded inconsistent results, which might be partly due to poor model-fit and disregard of Differential Item Functioning (DIF). Here, the brief Dutch form of the Multidimensional Personality Questionnaire (MPQ-BF-NL) was tested for cross-cultural measurement invariance, using Multiple Group Confirmatory Factor Analysis (MGCFA) for categorical outcomes. Representative Dutch and U.S. samples (both *Ns* = 1,055) were used for model exploration, and student samples (both *Ns* = 410) for cross-validation. The MPQ-BF-NL appeared partially strict invariant. 19% of the items contained DIF when scales were treated separately, while in the full model, allowing for 150 cross-loadings, 40% of the items contained DIF. The majority of DIF was observed in thresholds (i.e. difficulty parameters). Not accounting for DIF would yield invalid inferences for 4 out of 11 primary scales. Higher corrected factor scores were evident in the U.S. sample for Social Closeness and Stress Reaction, whereas lower for Achievement and Aggression, respectively, suggesting that the U.S. society, relative to the Dutch, fosters more community and less agency.

Introduction

"…the study of personality and culture is no longer a matter of documenting how culture shapes personality; instead, it asks how personality traits and culture interact to shape the behavior of individuals and social groups" (Hofstede & McCrae, 2004 p. 57)

A prerequisite for studying the interaction of personality traits and culture is the ability to distinguish the two. While Hofstede and McCrae appear confident that our measures are able to make that distinction, notable others are not (see for example Chen, 2008). Two widely held conceptions about personality traits exist that carry different implications for the relation between culture and personality (Borsboom, Mellenbergh, & van Heerden, 2003). One can hold a realist conception, believing that personality traits are 'real', and refer to psychobiological structures that bring about behavior. Or, alternatively, one can hold a constructivist conception, assuming that personality traits serve to organize observations, but not necessarily reflect nature. While the former stance, at least in principle, implies that culture and personality are clearly separable, the latter implies a more permeable relation between personality and culture.

From the realist viewpoint, it is often assumed that there should be a general, universal personality structure, applying to everyone regardless of cultural context. Depending on the specific criteria used, previous studies suggest that the content of two to six personality factors is cross-culturally replicable (Ashton & Lee, 2010; De Raad et al., 2010; McCrae, 2002; Saucier et al., 2014). However, imposing a specific personality model on different cultures by translating the original measure generally resulted in suboptimal model fit (Aluja, García, & García, 2004; García, Aluja, & García, 2004; Trull & Geary, 1997). Findings like these suggest that there is something universal about the content of personality, but that specific cultural and linguistic influences can obscure the commonality. In other words, instruments that are currently used are likely not adequately distinguishing between personality and culture.

A further implication of these findings is that inferences about cross-cultural differences cannot straightforwardly be made from observed differences in scale scores. Distributional properties can only be meaningfully compared between cultures when the scales measure the same construct in the same way in the different groups (i.e. cultures, languages). Whether this is the case can be investigated by testing for measurement invariance. Lack of cross-cultural measurement invariance indicates either non-universality of the constructs,

or influences of non-intended cultural or linguistic factors on the responses. In the former case quantitative comparisons are unwarranted because of qualitative differences in the constructs. In the latter case, one may still be able to compare the groups. However, the comparison of raw scale scores might yield invalid conclusions about distributional differences. Differentially weighing item influences for deducing trait scores or removing the unfair items altogether can make the cross-cultural comparison valid.

Although numerous studies on cross-cultural differences in distributional properties of personality measures have been published (see for example Schmitt, Allik, McCrae, & Benet-Martinez, 2007), most do not convincingly show that observed differences between cultures indeed reflect proper trait differences as opposed to biases in measurement (for exceptions see Church et al., 2011; Huang, Church, & Katigbak, 1997; Johnson et al., 2008). The majority of cross-cultural comparisons within the personality literature do not formally test for measurement invariance (Chen, 2008), although some recent studies do (e.g., Bou Malham & Saucier, 2014; Cieciuch, Davidov, Vecchione, Beierlein, & Schwartz, 2014; Joshanloo et al., 2014; Supple, Su, Plunkett, Peterson, & Bush, 2013). Generally when measurement properties of a measure are compared across different cultures, loading patterns derived from Exploratory Factor Analysis (EFA) or Principal Components Analysis (PCA) are informally compared across samples, often using the congruence coefficient to decide on their similarity (Lorenzo-Seva & ten Berge, 2006). However, this congruence coefficient has its limitations, because when loadings are uniformly high in both samples (as is to be expected with translated personality measures), differences in the ordering of item loadings are not very informative. What would be informative, is finding out that a specific item has a loading that is significantly different in one sample compared to another. Moreover, the congruence coefficient is not informative about differences in intercepts or difficulties of the items or about the error terms of the models. These parameters also need to be the same over cultures to be able to meaningfully compare distributional properties of the construct at hand (Meredith, 1993).

In order to be able to make a meaningful cross-cultural comparison of mean levels of personality traits, we conducted a measurement invariance study using general population samples from Minnesota (U.S.) and the Netherlands. The Multidimensional Personality Questionnaire (MPQ; Tellegen & Waller, 2008) in its Dutch brief form (MPQ-BF-NL; chapter 2) was used. The MPQ is a comprehensive hierarchical measure of normal personality variation. It provides coverage of a range of traits encompassing the domains of

temperament, interpersonal and imaginative style, and behavioral regulation. The underlying model of the MPQ follows from a realist trait perspective making it conceptually suitable for studying the universality of personality structure. From this realist perspective it is expected that with the MPQ cultures can be quantitatively compared on qualitatively the same traits. Violations of measurement invariance should be corrected for by adjusting item parameters for Dutch and U.S. samples.

Cross-cultural studies on the MPQ have been conducted with Hebrew (Ben-Porath et al., 1995) and German (Johnson et al., 2008) versions of the measure. In Israel, the pattern of loadings from Hebrew data was informally compared to loadings from U.S. data (Ben-Porath et al., 1995). These patterns appeared similar, and also the internal consistencies of the scales were commensurate. Formal tests of measurement variance were not carried out. In Germany there was searched for Differential Item Functioning (DIF) using Item Response Theory (IRT). Around 30% of the items contained DIF, and most differences in mean scale scores found between the U.S. and the German samples turned out to be due to differences in item-difficulties (Johnson et al., 2008).

Several studies, using a variety of measures, have examined trait differences between the U.S. and the Netherlands. Overall, these studies yielded highly inconsistent results. For example, with regard to the domain of the Five Factor Model (FFM), a large scale cross-cultural comparison of the Big Five Inventory (BFI; Schmitt et al., 2007) found that a Dutch student sample scored lower on Agreeableness and Conscientiousness than U.S. student samples. In a study that used the OPQ32 to operationalize the FFM, the U.S. sample also scored higher on Conscientiousness, but lower on Openness as well. No difference was observed for Agreeableness (Bartram, 2011).

The study described here is, to the authors' knowledge, the first cross-cultural comparison of distributional properties of broadband personality traits in which (a) a well-fitting personality model is used, and (b) absence of measurement invariance is controlled for. Moreover, this paper examines personality differences between the U.S. and the Netherlands by applying a design that resolves some of the limitations of previous studies. First, most cross-cultural comparisons to date were limited to students, or samples characterized by limited background information. We employ large samples that are representative for the general population. Secondly, in previous U.S.-Netherlands comparisons, no formal tests of measurement invariance were carried out on item level, or

(partial) measurement invariance higher than the configural level was not shown. We aim for (partially) strict measurement invariance in the measure, thus targeting more meaningful distributions of the traits. Finally, most previous studies could not rule out sample characteristics as no cross-validation data were provided. To guard against such sampling influences, we cross-validated our results in independent samples.

Method

Samples

Dutch primary sample

We used a representative Dutch sample, stratified by gender, age, educational level, and county (also used in the development of the MPQ-BF-NL; Centraal Bureau voor de Statistiek, 2009; chapter 2) for this study. It consisted of 1,060 participants who took part in a panel survey conducted by Flycatcher, a full-service online research company. Five participants were removed due to random responding or adhering to a strong response set as indicated by very high scores on respectively the Variable Response Inconsistency (VRIN) or True Response Inconsistency (TRIN) scales.[1] The final sample (N = 1,055) consisted of 510 (48%) men and 545 women, with a mean age of 45.83 years (SD = 15.74).

U.S. primary sample

1,055 records from the Minnesota Twin Registry were used as U.S. comparison sample. The records were randomly selected from the Registry, with the following restrictions: (a) all records needed to be valid (in terms of VRIN and TRIN); (b) the resulting sample needed to correspond to the Dutch sample's stratification in terms of gender and age in order to control for effects of these variables. Moreover, as few members of the same family as possible were included, though some multiple inclusions were inevitable because of stratification purposes. The resulting sample was identical to the Dutch primary sample in terms of gender - 510 (48%) men and 545 women, and age (M = 45.60 years; SD = 13.35).

[1] Specifically, protocols adhering to the following criteria were excluded: (a) a response pattern that was excessively inconsistent with respect to item content (i.e., score on VRIN > 3 *SD* above mean score), (b) a response pattern that was excessively polarized toward either true or false irrespective of item content (i.e., score on TRIN > |3.21| *SD* above mean score), and (c) a response pattern that was both inconsistent and polarized in direction (i.e., score is 2 *SD* above the mean for VRIN and |2.28| > *SD* from the mean for TRIN).

The 1,055 included protocols belonged to members of 762 families, and only four twin pairs were included.

Dutch cross-validation student sample

A second Dutch sample was used to determine whether the final measurement invariance model replicated over samples. This sample consisted of 438 undergraduate students from the University of Amsterdam who took part in a computerized mass testing session in 2012 in partial fulfillment of their first year course requirements. Six protocols exceeded the cut-offs for the VRIN/TRIN validity scales. 410 of the valid protocols were randomly selected, in order to match the number of protocols in the U.S. cross-validation student sample described below. The final sample (N = 410) consisted of 129 (31%) men and 281 women. Age ranged from 17 to 30 with a mean of 19.69 years (SD = 1.81).

U.S. cross-validation student sample

For replication purposes, protocols from 412 undergraduate psychology majors from Florida State University were used. Students filled out the questionnaire with paper and pencil in 1992, and for course credit. Two protocols were rejected because of invalid VRIN and/or TRIN scores as described above. The resulting sample (N = 410) consisted of 127 (31%) men and 283 women. Age ranged from 16 to 43 with a mean of 18.78 years (SD = 2.25).

Measure

The Dutch brief form of the MPQ (MPQ-BF-NL; chapter 2) was used. The items comprising the MPQ-BF-NL were selected to represent the constructs measured by the full length U.S. version as closely as possible. In Table 3.1, the item numbers for the MPQ-BF-NL are displayed, corresponding to the 276-item full length U.S. version.

The 11 MPQ-BF-NL primary trait scales are measured by 12 binary items each. Including three extra items needed for determining VRIN and TRIN scores, the full measure consists of 135 items. The primary trait scales coalesce into three higher-order factors: Positive Emotionality (PEM), Negative Emotionality (NEM), and Constraint (CON). PEM is comprised of primary trait scales Wellbeing (WB), Social Potency (SP), Achievement (AC), and Social Closeness (SC). NEM includes Stress Reaction (SR), Aggression (AG) and Alienation (AL), and CON includes Control (CO), Harm Avoidance (HA) and Traditionalism (TR). Absorption (AB) is not allocated to any of the three higher-order factors. Table 3.2 provides reliabilities and further descriptions of the measurement domain for each primary scale.

Table 3.1. Items comprising the MPQ-BF-NL and their cross-validated DIF

	WB	SP	AC	SC	SR	AG	AL	CO	HA	TR	AB
Item 1	17	1	10	4	3	7	27	24	31	9[b]	53
Item 2	32	15	34	16	14	20	52	38[c]	69	56	81
Item 3	42[a]	23[b]	71[a]	29	36	59	66	47[e]	94	78[ab]	123
Item 4	51	43[b]	87	67	117	72[c]	77	57[c]	114	109[e]	149
Item 5	61	83[a]	111	75[b]	127	100	119	92[bc]	134	118	156[b]
Item 6	120	93[c]	122	88	158[b]	113[ac]	146[b]	103	145	140	182
Item 7	167	105[a]	138	101	171	143	161[b]	136	166	151	189
Item 8	176[a]	128[b]	150	137	193	212[b]	187	147	186	160	197
Item 9	191	157	178	152	214[e]	226[e]	230	159[b]	217	210	215
Item 10	205	170[b]	194	202	248	239[ab]	246	172	228	240	238
Item 11	235[c]	213[a]	204	216[b]	258[b]	254	260	195[c]	237[b]	252	257
Item 12	272[bce]	255	271	241[b]	269	270	274	209	247	262	273[b]
#DIF items in scale	3	7	1	3	3	4	2	3	1	3	2
#DIF items	4	8	1	3	3	5	2	6	1	3	2

Note. Numbers correspond to the 276-item full length version of the MPQ. NL = the Netherlands; U.S. = United States; WB = Wellbeing; SP = Social Potency; AC = Achievement; SC = Social Closeness; SR = Stress Reaction; AG = Aggression; AL = Alienation; CO = Control; HA = Harm Avoidance; TR = Traditionalism; AB = Absorption.
[a]Item exhibits stable DIF in the loading. [b] Item exhibits stable DIF in the threshold. [c]Item exhibits stable DIF in a cross-loading on another scale. [e]Item exhibits stable DIF in the residual.

The MPQ scales have demonstrated good reliability in a variety of samples as well as theoretically predicted correlations with other instruments (Tellegen & Waller, 2008). Moreover, the scale scores have been shown to predict behavior (Kamp, 1986), to distinguish between different forms of psychopathology (see for example McGue et al., 1999, 1997; Miller et al., 2003) and to predict clinical variables better than most other personality scales (Grucza & Goldberg, 2007). The MPQ-BF-NL showed similar internal consistencies, with Cronbach's alpha in the range of .75 to .84 for most scales, and between .70 and .73 for Traditionalism, Harm Avoidance, and Aggression. Moreover, higher-order structure and correlational patterns of the MPQ-BF-NL are quite similar to the U.S. (chapter 2).

Testing measurement invariance

We tested for measurement invariance by means of Multiple Group Confirmatory Factor Analysis (MGCFA) for categorical outcomes, with a robust Weighted Least Squares (WLSMV) estimation procedure (Muthén, du Toit, & Spisic, 1997). All analyses were carried out in Mplus version 7.0 (Muthén & Muthén, 2012). The present study utilized the Theta parameterization as it provides the possibility to either restrict residuals to be equal over

groups or to freely estimate some of the residuals over groups. This practice most closely resembles measurement invariance practices in MGCFA for continuous outcomes.

Table 3.2. Cronbach's alphas and description of high scorers for the MPQ-BF-NL scales

	Cronbach's α		
Scale	NL (*N* = 1,055)	U.S. (*N* = 1,055)	Description of a high scorer
Wellbeing	.80	.80	Has a happy, cheerful disposition; feels good about self and sees a bright future
Social Potency	.84	.85	Is forceful and decisive; fond of influencing others; fond of leadership roles
Achievement	.76	.75	Works hard; enjoys demanding projects and working long hours
Social Closeness	.81	.82	Is sociable, likes people, and turns to others for comfort
Stress Reaction	.84	.82	Is nervous, vulnerable, sensitive, prone to worry
Aggression	.73	.75	Hurts others for own advantage; will frighten and cause discomfort for others
Alienation	.82	.81	Feels mistreated, victimized, betrayed, and the target of false rumors
Control	.75	.75	Is reflective, cautious, careful, rational, planful
Harm Avoidance	.72	.77	Avoids excitement and danger; prefers safe activities even if they are tedious
Traditionalism	.70	.74	Desires a conservative social environment; endorses high moral standards
Absorption	.80	.75	Is responsive to evocative sights and sounds; readily captured by entrancing stimuli

Note. NL = the Netherlands; U.S. United States.

A scale can be measurement invariant at different levels (Meredith, 1993). Usually, increasingly strict models are consecutively tested for tenability. Configural invariance holds when the same structural model fits in the different groups. However, all parameters are free to vary over groups. Weak invariance holds when all loadings can be said to be equal over groups; strong invariance when also thresholds are equal; and strict invariance is achieved when also residuals can be restricted over groups. When a certain level of measurement invariance does not hold, one or more of the fixed parameters could be freely estimated over groups in order to obtain the desired model fit. These differing parameters over groups can be said to contribute to the item's Differential Item Functioning (DIF), a term that is commonly used in the realm of Item Response Theory (IRT), but can also be used for describing non-invariance of item parameters in CFA (Kim & Yoon, 2011).

To evaluate the fit of the different models the Root Mean Square Error of Approximation (RMSEA), the Confirmatory Fit Index (CFI), and the Tucker-Lewis Index (TLI) were

evaluated. For the absolute fit of models RMSEA values smaller than .08 and smaller than .05 were considered to indicate acceptable respectively good fit (MacCallum, Browne, & Sugawara, 1996). For CFI and TLI values larger than .90 and .95 were used as indicators for acceptable and good fit respectively (Hu & Bentler, 2010). For comprehensiveness we also display the χ^2 statistic, but do not evaluate it, as it is well established that even well-fitting models are rejected in such large samples. For the evaluation of the relative fit of models, we considered reduction in fit to be significant when increase in RMSEA was larger than .015 or when decrease in CFI or TLI was larger than .01.

Testing measurement invariance in single scales and in the full model

We tested for measurement invariance in all scales of the MPQ separately as well as in the full multidimensional model. When strict invariance did not hold, partially strict invariant models were derived, in order to allow for cross-cultural comparisons. The strict invariant model was adjusted to derive a model that showed equal fit to that of the configural model. Because individual parameter estimates are sensitive for changes in the rest of the model, we hold this procedure for more efficient than sequentially securing weak measurement invariance, strong measurement invariance, etc. In the process of freeing parameters priority lay with freeing thresholds over loadings, and loadings over residuals of items, respectively. This specific rule was adopted because threshold DIF generally is more extensive than other forms of DIF (Johnson et al., 2008), and because loading DIF is easier to interpret than residual DIF. Exception to this hierarchical rule was reserved for parameters for which the modification index was notably (in practice always > 7) larger than for the parameter higher in the hierarchy. This procedure allows that a threshold is freely estimated across groups, while the respective loading remains fixed (or vice versa). Of note, there is some debate about separately fixing and freeing threshold and loading parameters, with some arguing against it because of interdependence (Muthén & Asparouhov, 2002), while others deem the separate treatment of these parameters appropriate (Millsap & Yun-Tein, 2004). Because of the conceptual likeness of loadings and thresholds on the one hand and discrimination and difficulty parameters from IRT on the other, we decided to treat the freeing of loadings and thresholds separately.

Results

Tests of measurement invariance and DIF in single scales

9 of the 11 primary trait scales showed acceptable to good fit for the configural model. For Social Potency and Control, fit indices pointed to poor to acceptable fit. For all scales except Social Closeness, strict invariant models showed reduced fit relative to the configural models. To achieve partially strict invariance across scales, a total of 26 measurement parameters out of a total of 396 was freely estimated over the primary samples: 22 thresholds (17%), 1 loading (.8%) and 3 residuals (2%). The freely estimated parameters belonged to 25 (19%) of the items. 85% of the observed DIF was replicated in the student samples, indicating that most DIF was stable and not sample specific. Full description of measurement invariance and DIF of single scales is provided in the chapter 3 supplementary results section, supplementary Tables S.3.1 through S.3.4 and supplementary Figure S.3.1.

Development of a well-fitting CFA model for the full MPQ-model

In line with previous investigations, the simple structure full model (with factors allowed to covary) did not adequately fit the data in the NL and U.S. samples (see Table 3.3). Therefore, we sought a stable set of cross-loadings that would bolster model fit. The Dutch primary sample was randomly split into two halves. 99 cross-loadings (> ±.2 in both sample halves) from Exploratory Structural Equation Models (ESEM) were selected. When these cross-loadings were added to the simple structure CFA model some of the cross-loadings got notably smaller. Cross-loadings not making the ±.2 criterion were removed. More cross-loadings were added by inspection of modification indices in both sample halves. A maximum of one cross-loading per item was selected and modification indices needed to be larger than 10 in both sample halves. We ran the model again, removed small cross-loadings, inspected modification indices and re-iterated the model run. We repeated this procedure until a well-fitting model was derived. Each time a model with more cross-loadings was estimated, criteria for retention were further relaxed until good model fit was achieved. The resulting model contained 150 cross-loadings that were all larger than ±.075. Although choices were made statistically, the content of cross-loadings generally made substantive sense to the authors. For example, four items of the Achievement scale loaded strongly on Stress Reaction. All of these items pertained to some extent to with perfectionism and striving. It is conceivable that these aspects of Achievement are

influenced by Stress Reaction. Likewise, some Control items that appeared to be related to (inverse) recklessness loaded on Harm Avoidance.

Full model measurement invariance and DIF

In Table 3.3, the results of tests for measurement invariance of the full MPQ model are displayed. The model including 150 cross-loadings (hereinafter referred to as CFA150) showed good fit in the Dutch primary sample, which reduced to acceptable fit in the Dutch cross-validation sample, indicating that some of the cross-loadings were sample dependent without imposing major problems for model fit in cross-validation. The model fit in the U.S. samples also was acceptable, indicating that the selection of cross-loadings was equally suited for both the U.S. and NL. 126 (84%) of the cross-loadings had the same direction in all four samples.

Table 3.3. Fit for (Multiple Group) CFA(150) models

Model	*df*	χ^2	RMSEA 90% CI	CFI	TLI
Primary samples (NL, *N* = 1,055; U.S., *N* = 1,055)					
CFA single group NL	8,459	14,572	[.025, .027]	.788	.784
CFA single group U.S.	8,459	13,727	[.024, .025]	.815	.810
CFA150 single group NL	8,309	9,581	[.011, .013]	.956	.954
CFA150 single group U.S.	8,309	10,999	[.017, .018]	.905	.901
CFA150 configural invariance	16,618	20,590	[.014, .016]	.931	.928
CFA150 strict invariance	17,010	23,153	[.018, .019]	.893	.891
CFA150 partially strict invariance 1[a]	16,984	22,252	[.017, .018]	.908	.906
CFA150 partially strict invariance final[b]	16,948	21,265	[.015, .016]	.925	.923
Cross-validation samples (NL, *N* = 410; U.S., *N* = 410)					
CFA single group NL	8,459	1,030	[.021, .025]	.792	.787
CFA single group U.S.	8,459	9,843	[.018, .022]	.836	.833
CFA150 single group NL	8,309	9,079	[.013, .017]	.913	.909
CFA150 single group U.S.	8,309	9,134	[.013, .018]	.902	.898
CFA150 configural invariance	16,618	18,213	[.014, .017]	.908	.904
CFA150 strict invariance	17,010	19,162	[.016, .019]	.876	.874
CFA150 partially strict invariance 1[a]	16,984	18,979	[.015, .018]	.885	.883
CFA150 partially strict invariance final[b]	16,948	18,822	[.015, .018]	.892	.890

Note. CFA150 = Full CFA model with 150 cross-loadings (see text for clarification on selection of cross-loadings); NL = the Netherlands; U.S. = United States; RMSEA = Root Mean Square Error of Approximation; CI = Confidence Interval; CFI = Comparative Fit Index; TLI = Tucker Lewis Index.
[a]The choice of parameters to free was made on account of the single scale analyses on the primary samples: 1 loading was freely estimated over samples as were 22 thresholds and 3 residuals. [b]As partially strict invariance 1, plus freed parameters over groups selected from modification indices (>12.5).

CFA150 model measurement invariance

Fit of the CFA150 strict invariance model was near acceptable levels, but was significantly less good than fit of the configural invariance model (lower bound 90% CI RMSEA +.004; higher bound 90% CI RMSEA +.003; CFI -.038; TLI -.037). Therefore, a partially strict invariance model was developed, allowing a selection of parameters to differ between

samples. First, all 26 parameters that were freed in the single scale analyses were also freely estimated in these analyses, which improved model fit significantly, but not enough for it to be as good as fit of the configural model. Second, additional parameters were freely estimated over groups when the modification index was greater than 12.5. A maximum of one parameter per item was freed. Accordingly, another 36 measurement parameters were freely estimated over samples, of which 7 thresholds, 11 primary loadings, 14 cross-loadings, and 4 residuals. The final partially strict invariance model included 62 measurement parameters that were freely estimated over samples, of which 29 thresholds (22%), 12 primary loadings (9%), 14 cross-loadings (9%), and 7 residuals (5%). This final partially strict invariant model fitted the data as well as the configural model. Detailed information on the final partially strict invariance model is provided in supplementary Tables S.3.5 and S.3.7.

Cross-validation of the CFA150 model measurement invariance

In terms of model fit results were replicated in the cross-validation samples. The configural model yielded acceptable fit and the reduction in fit from configural to strict invariance was comparable to that observed in the primary samples (lower and higher bound 90% CI RMSEA +.002; CFI -.032; TLI -.030). The final partially strict invariant model fitted slightly less well as the configural model (lower and higher bound 90% CI RMSEA +.001; CFI -.016; TLI -.014), indicating replication of most, but not all DIF. Parameter estimates of the final partially strict invariance model in the replication samples can be found in supplementary Tables S.3.6 and S.3.7.

Of the 62 measurement parameters - belonging to 53 of the items - that were freely estimated over samples, 17 (27%) did not replicate in the cross-validation samples (parameter estimates can be found in supplementary Table S.3.4), which is 12% more than in the single scale analyses, indicating that DIF for the full-model was more sample dependent than for single scales. Most DIF pertained to thresholds, followed by cross-loadings, while very little DIF was apparent in loadings or residuals (see Table 3.1). 38 items (29% of the total questionnaire) contained some form of stable DIF over primary and cross-validation samples. Most scales contained three stable DIF items (ignoring cross-loadings) with a range of one (Achievement and Harm Avoidance), to seven (Social Potency).

Mean level trait differences between United States and Dutch samples

The above results show that strict invariance was not tenable. Therefore, raw scale score differences do not necessarily represent differences in underlying traits. Other cultural or linguistic constructs, not targeted by the scales, might have influenced the differences, resulting in DIF. For this reason, raw scale score differences between the NL and U.S. primary samples were compared to factor score differences. Both factor score differences from the (partially) strict invariance models obtained on the single scales, and factor score differences from the full model partially strict invariance model are displayed in Table 3.4. For the U.S. primary sample, mean raw scores were higher for Wellbeing, Social Closeness, and Stress Reaction than the respective Dutch scores. Mean raw scores for Social Potency, Aggression, Alienation and Control were lower in the U.S. sample than in the NL sample. When considering DIF-corrected factor scores, the pattern of cross-cultural differences shifted, with differences in single scale versus full model results being negligible. For 7 out of 11 scales, mean differences of corrected factor scores were only marginally different from raw mean differences. However, shifts for Wellbeing, Social Potency, Achievement, and Control seemed notable, because the adjusted estimates led to a substantially different conclusion regarding cross-cultural differences. Some of the observed differences in raw scale scores were not apparent in factor scores that were corrected for DIF. Raw scores suggested higher scores for Wellbeing and lower scores for Social Potency and Control for the U.S. sample, while their respective corrected factor scores did not show mean differences between samples. In contrast, while raw scale scores did not suggest a cross-cultural difference in Achievement, factor scores pointed to lower mean Achievement scores in the U.S. sample.

Sources of DIF and inferences about cross-cultural differences

For Wellbeing, Social Potency, Achievement, and Control the partially strict invariance models resulted in substantially different conclusions with regard to cross-cultural differences in comparison to their raw scale score counterparts. These different conclusions may indicate either differences in measurement between the Dutch and American MPQ or structural cultural differences in the constructs being measured. To understand which of these two alternatives was most plausible we explored DIF and other structural properties for each of these four scales.

Table 3.4. Comparison of mean differences for raw scale scores and unstandardized partially invariant factor scores

	Raw score *M*[a]		Standardized raw score[b]		Factor estimates single scale models[c]		Factor estimates full model[d]	
Scale	NL	U.S.	Δ	*Var*.U.S.	Δ	*Var*.U.S.	Δ	*Var*.U.S.
Wellbeing	8.09	8.76	.23**	0.93	**-.01**	0.85	**.04**	1.19
Social Potency	5.44	4.78	-.18**	0.98	**-.06**	1.16	**-.07**	0.94
Achievement	7.34	7.12	-.07	0.98	**-.20****	0.94	**-.24****	0.77
Social Closeness	6.70	7.83	.34**	0.98	.42**	1.22	.43**	1.03
Stress Reaction	4.97	5.87	.27**	0.99	.22**	0.84	.25**	0.60
Aggression	2.99	2.66	-.14*	1.02	-.15*	0.91	-.19**	0.89
Alienation	2.35	1.61	-.27**	0.70	-.18*	1.03	-.14	0.93
Control	8.46	7.81	-.23**	1.03	**-.10**	0.93	**-.01**	0.98
Harm Avoidance	8.45	8.42	-.01	1.11	.02	1.39	-.01	1.14
Traditionalism	8.84	9.02	.07	1.24	-.04	1.19	-.06	1.06
Absorption	5.88	5.97	.03	0.82	-.10*	0.75	-.12	0.74

Note. Bold-faced differences are notably different from standardized raw score differences. Var. = variance; NL = the Netherlands; U.S. = United States.

[a]Ranges of scales are 0-12. [b]Values are standardized such that the NL sample has a mean of zero and a standard deviation of one. [c]Estimates are from the single scale partially strict invariant models (see supplementary Table S.3.1). [d]Estimates are from the full partially strict invariant model (see Table 3.3).

*$p < .01$. ** $p < .001$.

Wellbeing

The mean raw score of the U.S. sample was significantly higher than for the NL sample. Correcting for DIF effectively removed this difference. The difference in the raw scale score appeared to be totally driven by DIF in items Wellbeing12 (272) and Wellbeing11 (235). When these items were disregarded in the calculation of the raw scale score, no difference was apparent between groups. Although DIF from item Wellbeing11 (235) was not replicated in the student samples, DIF from item Wellbeing12 (272) was. Also, item Wellbeing12 (272) showed DIF in three of its parameters, making it the item that was most severely affected by DIF of the whole questionnaire. Item Wellbeing12 (272) reads *"I always seem to have something exciting to look forward to"*, and was more readily endorsed by the U.S. samples than by the Dutch samples (i.e. threshold DIF). It was also positively related to Stress Reaction scores in the Dutch samples, while it was negatively related to Stress Reaction scores in the U.S. samples (i.e. cross-loading DIF). Lastly residuals differed between the U.S. and NL, although this did not translate into highly different proportions of explained variance in the item (see supplementary Table S.3.6). The DIF in item Wellbeing12 (272) could be due to differences in the meaning of the word *'exciting'*. The Dutch translation - *´spannend´* - can have a negative connotation, making it harder to endorse for the Dutch, and making it indicative for Stress Reaction as well as for Wellbeing.

Social Potency

When using raw scale score means, U.S. participants had lower scores on Social Potency than the Dutch. When using corrected factor scores however, there was no mean difference on this scale. Pertinent to this observation is that Social Potency contained most DIF of all scales, and that all of the eight parameters that were freely estimated over groups in the full model were replicated. With such substantial DIF, one may expect more fundamental differences in the construct at hand. Accordingly, the structural one-factor model of Social Potency as used here might not be optimal for a cross-cultural comparison. To clarify this issue, additional analyses were carried out and can be found in supplementary Table S.3.8.

A two-factor ESEM configural model showed significantly better fit than the one-factor configural CFA model, indicating that a multidimensional approach to the construct of Social Potency might be meaningful. The first factor contained items that revolved around seeking leadership, while the second factor contained items that revolved around persuasiveness. When we tried to achieve partially strict invariance, we found most of the threshold DIF in the items still to be present, but the multidimensionality made the loading DIF disappear. On the leadership factor, U.S. participants showed significantly lower scores than the Dutch (Δ = -.20 *SD*, $p < .001$), while they did not score differently (Δ = .05 *SD*, $p = .469$) on the persuasiveness factor. Consequently, differences observed in raw scale scores between the samples seemed to be driven by U.S. participants seeking out leadership less than the Dutch participants. Apparently, the cross-cultural difference in only one aspect of the construct was compensated for in the one-factor solution by introducing loading DIF.

Achievement

U.S. participants showed lower scores on Achievement than Dutch participants after controlling for DIF, while in the raw score comparison no difference was observed. Most notable when inspecting the parameter estimates and fit indices for this scale is that most of the DIF observed in the primary samples was not replicated in the student samples (parameter estimates can be found in supplementary Table S.3.4). Also, the strict invariance model relatively to the configural invariance model fitted much better for the cross-validation samples than for the primary samples, indicating that for students there is less DIF than for the general population when comparing U.S. and Dutch samples. The interpretation of DIF for AC cannot be said to be robust, and should therefore be cautiously taken.

Notwithstanding the issues of robustness, the appearance of lower scores for the U.S. primary sample when controlling for DIF seems due to threshold DIF in item Achievement2 (34): *"I play hard and I work hard"*. When this item was removed before determining the raw scale score difference, a significant lower score was also present in the U.S. primary sample than in the Dutch primary sample (Δ M = -.17 SD, p < .001). The item refers to a common English expression that does not as such exist in Dutch. Because of its relative unfamiliarity, the primary Dutch sample might have been more reluctant endorsing it.

Control

The mean raw score of the U.S. sample was significantly lower than of the NL sample. Correcting for DIF made this difference disappear. The difference in the raw scale score appeared to be totally driven by DIF in items Control5 (92) and Control9 (159). Disregarding these items from the raw scale resulted in no observed difference between the groups. The authors hold no plausible theory about the origin of these DIF findings, especially since highly similar items did not exhibited DIF.

Discussion

The aim of the present study was to make a cross-cultural comparison of the structure and distribution of personality traits as operationalized by the MPQ-BF-NL. Two prerequisites for making valid cross-cultural comparisons were satisfied: (a) a well-fitting model was obtained, and (b) Differential Item Functioning (DIF) was controlled for. In a comparison between representative samples from Minnesota (U.S.) and the Netherlands 19% of items showed DIF when scales were considered separately. When attending to the MPQ model as a whole, 40% of the items contained some form of DIF. Robust DIF, i.e. DIF that replicated in the student samples from both countries, was found in 16% (for single scales) and 29% (for the full model) of the items, respectively: 22% of thresholds, 9% of loadings and 5% of residuals contained DIF in the full model. These results indicate that items differed mainly in their 'difficulty' or endorsement probabilities, and not so much in their discrimination or relevance to the constructs. The observed amount of non-cross-validated DIF in the single scales (19%) is somewhat lower than the proportion (30%) reported in a study comparing U.S. and German samples using the MPQ (Johnson et al., 2008) and firmly lower than the 40-60% that is observed in cross-cultural FFM comparisons (see Church et al., 2011; Nye, Roberts, Saucier, & Zhou, 2008). It should be noted that these FFM comparisons included cultures that are presumably more 'distant' from each other than the cultures selected in

our comparison. Specifically, Church et al. (2011) compared data from the U.S. with data from the Philippines and Mexico, and Nye et al. (2008) compared samples from the U.S., Greece and China. The pattern of discrepancies is in line with the expectation that the degree of DIF is a (positive) function of the cultural dissimilarity.

U.S. participants scored about half a standard deviation higher than Dutch participants on Social Closeness, indicating they generally reported being more sociable, and enjoying the company of others more (mean Social Closeness $\approx {}^{1}/_{2}$ *SD* higher; for precise values and notes on these see Table 3.4). Moreover, U.S. participants generally scored higher on Stress Reaction, i.e. they were more prone to feelings of nervousness and worry, and described themselves as more sensitive (mean Stress Reaction $\approx {}^{1}/_{4}$ *SD* higher). Perhaps somewhat surprisingly, they endorsed less inclination for working hard and enjoying long hours (mean Achievement $\approx {}^{1}/_{5}$ *SD* lower). U.S. participants were less inclined to seek leadership (mean Social Potency leadership $\approx {}^{1}/_{5}$ *SD* lower), although they reported being equally persuasive as their Dutch counterparts. They also reported less alienation, in that they indicated less feeling deceived, or victimized (mean Alienation $\approx {}^{1}/_{6}$ *SD* lower), and less instrumental aggression, i.e. the inclination towards hurting others for personal advantage, or enjoying frightening or harming people (mean Aggression $\approx {}^{1}/_{6}$ *SD* lower). No cross-cultural differences were found for mean levels of Wellbeing, Control, Harm Avoidance, Traditionalism and Absorption.

Controlling for DIF did not substantially alter the conclusions about cross-cultural mean differences for most trait scores. However, mean differences for Wellbeing, Social Potency, Achievement, and Control did diverge from those observed when inspecting unadjusted raw scale scores. Based on the latter, one might mistakenly infer that the U.S. and Dutch participants, on average, equally enjoyed working long hours (equal mean Achievement), or that U.S. participants had happier dispositions (higher mean Wellbeing), were less persuasive than the Dutch (lower mean Social Potency), or reported being less planful and more impulsive than the Dutch participants (lower Control).

Of note, these adjustments cannot by themselves speak to whether a realist or universalist stance towards personality is untenable, or not. Adjusted scores that reflect a limited degree of DIF may still be regarded as reflecting actual quantitative differences between cultures

on qualitatively the same construct, and such differences may well be interpreted by specific cultural and linguistic differences.[2]

Some speculative theorizing may be ventured. On average, U.S. participants were lower on the leadership aspect of Social Potency and on Achievement, while higher on Social Closeness than Dutch participants. Social Potency and Achievement represent the agentic, and Social Closeness represents the communal aspect of the higher-order construct Positive Emotionality (Tellegen & Waller, 2008). These higher-order dimensions of agency and communion are thought of as basic dimensions of social judgments (Wojciszke, Abele, & Baryla, 2009), and are associated with the cross-cultural replicable Big Two (Saucier et al., 2014). Moreover, agency and communion have also been mapped onto the two poles of (one-dimensional) individualism-collectivism (Hofstede & McCrae, 2004), with agency being predominantly associated with concern with the self, and communion with concern for others (Wojciszke et al., 2009). Departing from the premise that after controlling for DIF cross-cultural differences reflect societal influences on personality, the U.S. societal environment might foster communal personality traits that serve more collectivistic values, while Dutch society might encourage agentic personality traits that serve more individualistic values. Of course, these conjectures are relative statements (i.e. comparing U.S. and Dutch samples) in need of further empirical testing, but we hold that unbiased cross-cultural comparisons using broadband personality questionnaires provide a good starting point for deriving such hypotheses.

Several limitations of the present study deserve mentioning. First, the samples that were used for the U.S.-Netherlands comparison might have differed in other respects besides culture and personality. Because the U.S. samples were collected several years before the Dutch samples, cross-cultural differences might be confounded by cohort differences. This cohort difference may serve as an alternative hypothesis for the counter intuitive lower score on Achievement for the U.S. primary sample. Also, the cross-validation samples consisted of students. Although we do not deem it probable, it cannot be ruled out that

[2] Linguistic differences constitute the Achilles heel of cross-cultural comparisons of translated measures. Consequently, it has been suggested by some researchers to abstain from merely translating questionnaires, but to instead develop measures from within each of the specific languages (Saucier et al., 2014) or let development of the measure be informed by different specific languages (Boehnke et al., 2014). We value both approaches, but as we have illustrated, a translated instrument can provide information about both cross-cultural generalizable aspects of personality and cultural – or language – specific aspects.

personality structure for students differs from the general population. If so, reduction in fit of the tested MPQ model with 150 cross-loadings might be due to this systematic sample difference. Second, while a well-fitting multidimensional CFA model was derived that was replicated in independent samples, the simple structure MPQ model (i.e., with items loading exclusively onto factors they are presumed to measure) did not fit the data. This finding is a well-documented problem in the literature on personality, and with recently developed techniques, such as ESEM (Marsh et al., 2009) or Bayesian SEM (Muthén & Asparouhov, 2012), one should be able to solve this issue by being able to estimate the full loading matrix. Indeed, for the present data one of these techniques would have been the preferred method, if it were not for problems of convergence and partial invariance of loadings. A final limitation worthy of mention is that although a confirmatory strategy was chosen, exploratory procedures (i.e. relaxing restrictions by inspecting modification indices) were employed to gain partial measurement invariance. Mitigating this concern is that the majority of DIF replicated across samples, and their interpretation – albeit post hoc – generally made intuitive sense and in fact helped in understanding U.S.-Dutch differences in culture and personality.

Conclusion

We demonstrated that inferences about distributional properties of traits in different cultural groups might be incorrect when relying on raw scale scores without accounting for lack of measurement invariance. Depending on whether one opts for a scale by scale, unidimensional, or, multidimensional approach, 19% respectively 40% of the items of the MPQ-BF-NL contained DIF. Inspection of these DIF items provided a window to cultural differences in the expression and structure of personality, as well as on the actual mean level differences in traits. The data provided here add to the distinction between personality and culture, and might even add to the knowledge about their interrelation.