To know personality is to measure it

Introducing a Dutch brief form of the Multidimensional Personality Questionnaire

Eigenhuis, A.

Chapter 4

Personality in general and clinical samples

Measurement invariance of the Multidimensional Personality Questionnaire

Annemarie Eigenhuis
Jan H. Kamphuis
Arjen Noordhof

An excerpt of this chapter is published as:
Eigenhuis, A., Kamphuis, J. H., & Noordhof, A. (in press). Personality in general and clinical samples: Measurement invariance of the Multidimensional Personality Questionnaire (MPQ). *Psychological Assessment.*
http://dx.doi.org/10.1037/pas0000408

Abstract

A growing body of research suggests that the same general dimensions can describe normal and pathological personality, but most of the supporting evidence is exploratory. We aim to determine in a confirmatory framework the extent to which responses on the Multidimensional Personality Questionnaire (MPQ) are identical across general and clinical samples. We tested the Dutch brief form of the Multidimensional Personality Questionnaire (MPQ-BF-NL) for measurement invariance across a general population subsample (*N*=365) and a clinical sample (*N*= 365), using Multiple Group Confirmatory Factor Analysis (MGCFA) and Multiple Group Exploratory Structural Equation Modeling (MGESEM). As an omnibus personality test, the MPQ-BF-NL revealed strict invariance, indicating absence of bias. Unidimensional per scale tests for measurement invariance revealed that 10% of items appeared to contain bias across samples. Item bias only affected the scale interpretation of Achievement, with individuals from the clinical sample more readily admitting to put high demands on themselves than individuals from the general sample, regardless of trait level. This formal test of equivalence provides strong evidence for the common structure of normal and pathological personality and lends further support to the clinical utility of the MPQ.

Introduction

A growing body of evidence suggests that normal and pathological personality can be described by the same general dimensions, and differ only in magnitude, not in kind. (Clark, 2005; Markon et al., 2005). This notion has two implications. The first is fundamental: personality pathology can be described by maladaptive combinations of extreme standings on common traits. Reflecting this recognition, the alternative (Section III) model for personality disorders of the Diagnostic and Statistical Manual, fifth edition (DSM-5; American Psychiatric Association, 2013) describes personality pathology by five dimensional constructs. The second implication of the commonality in structure of normal and pathological personality is practical: measures developed to describe variation in normal personality are, at least in principle, suited for use within clinical populations.

Although numerous studies have been carried out that support the common structure of normal and pathological personality (see for example Andersen & Bienvenu, 2011), research that specifically examined the structure of responses on personality questionnaires has two limitations. First, the evidence is rather unevenly distributed across possible comparisons. It is instructive to examine this issue as a 2x2 factorial problem, as graphically displayed in Table 4.1. In order to conclude that the basic dimensions of personality and personality pathology are the same, one should ascertain that this is the case for all comparisons in the marginals of Table 4.1. Most attention in research has been devoted toward the A/B marginal where normal and pathology personality instruments were compared in a general population sample (for example Blais, 2010; De Fruyt et al., 2013; Krueger et al., 2001; Schroeder, Wormworth, & Livesley, 1992; Sellbom & Ben-Porath, 2005) probably due to the ease of availability of general population and student samples. Although these comparisons shed light on structural similarities between measures of normal and pathological personality in the general population, they cannot demonstrate that the constructs are identical in clinical populations. Studies geared at this comparison (C/D) do exist, but are more scarce (DiLalla et al., 1993; Sellbom, Ben-Porath, & Bagby, 2008b). Studies examining the A/B and C/D marginals provide information about the structural similarities of the different measures in the same populations, but of course personality structure might still differ across general and clinical samples. Some studies have been carried out on pathology measures in general and clinical samples (B/D marginals; Livesley, Jackson, & Schroeder, 1992; Pukrop, Gentil, Steinbring, & Steinmeyer, 2001; Strack & Guevara, 1999). We know of only one study on the comparison between a

general and a clinical sample with a normal range broadband personality measure (A/C marginal; Bagby et al., 1999). In this study NEO-PI-R (Costa & McCrae, 1992b) responses from a psychiatric sample were factor analyzed, and loadings of a five-factor solution were informally evaluated for their congruence with the expected pattern. Lastly, several studies allowed for all (A/B/C/D) comparisons possible. However, these studies generally described interrelations between scale scores of the measures and did not directly compare structure on the item-level (Costa & McCrae, 1992a; Markon et al., 2005; O'Connor, 2002; van der Heijden, Rossi, van der Veld, Derksen, & Egger, 2013; Watson, Stasik, Ro, & Clark, 2013). Of note, the compositional structure of the clinical samples employed in the studies also differs substantially, each posing specific challenges to the generalizability of the obtained results. Samples vary for example in severity and specific nature of psychopathology..

Table 4.1. The common structure of normal and pathological personality as a factorial 2x2 problem

		Measure: Normal	Measure: Pathology	
Sample: General		**A** **Normal measure in general sample**	**B** **Pathology measure in general sample**	A/B Normal and pathology measure in general sample
Sample: Clinical		**C** **Normal measure in clinical sample**	**D** **Pathology measure in clinical sample**	C/D Normal and pathology measure in clinical sample
		A/C Normal measure across samples	B/D Pathology measure across samples	A/B/C/D Normal and pathology measure across samples

A second concern with regard to the evidence in support of the common structure of normal and pathological personality is that none of the aforementioned studies conducted formal structure comparisons over the responses across the different measures or samples. Whereas informal (i.e. exploratory) comparisons can be informative about the similarity in the number of the factors that best describe normal and pathological personality and about the similarity of their content, they do not show whether the relation between the indicators (items) and the constructs is the same over groups. To examine this, one needs to conduct formal tests of measurement invariance. This issue is particularly pressing when one uses a measure that is developed in one population to measure the same constructs in another population. Measurement properties of the instrument may not generalize to this

second population. When a measure lacks measurement invariance (i.e., when the measurement properties do not generalize to the new population), bias or Differential Item Functioning (DIF) is present and one cannot make straightforward, valid comparisons of scores from the different groups. An item contains DIF when the probability of endorsing the item differs across samples, without this difference being attributable to differences in the latent variable. DIF can pose a problem when a measure is used in different populations because scores might be influenced by non-intended constructs, and therefore scores might be incomparable between samples.

Normal range broadband personality instruments can be valuable for clinical practice. For example, knowledge about personality standings of an individual can guide hypothesis formulation regarding the nature of the presented complaints (e.g. reactive and transient versus more ingrained and long-term) and help the matching of treatment to the person (Harkness & Lilienfeld, 1997). Indeed, several personality traits have been shown to moderate response to treatment for Major Depressive Disorder (Bagby et al., 2008; Bulmash, Harkness, Stewart, & Bagby, 2009; Quilty et al., 2008). Studies in which normal personality is assessed in clinical samples often use Five-Factor Model (FFM) instruments. An alternative model worthy of attention of both researchers and clinicians is operationalized by the Multidimensional Personality Questionnaire (MPQ; Tellegen & Waller, 2008). The MPQ measures 11 primary trait scales that coalesce into three[1] higher-order dimensions (see the Method section for a more elaborate description of the model).

The theoretical value of the MPQ can be illustrated by the influential neurobiological personality model of Depue (Depue & Lenzenweger, 2001) in which the MPQ constructs are central. Furthermore, MPQ primary trait scales (i.e. Stress Reaction and Harm Avoidance) have shown utility in examining aspects of fear learning (Gazendam et al., 2015). Clinical relevance is exemplified in the MPQ's utility in distinguishing between internalizing and externalizing conditions, or in delineating subtypes of disorders (Krueger et al., 2001; Miller et al., 2003). Finally, the MPQ has shown incremental predictive utility in comparison to other operationalizations of personality with regard to clinical indicators (Grucza &

[1] The three-factor higher-order model consists of Positive Emotionality (PEM), Negative Emotionality (NEM) and Constraint (CON). An alternative conceptualization of the higher-order structure of the MPQ includes four higher-order factors in which PEM is split into Agentic Positive Emotionality (PEM-A) and Communal Positive Emotionality (PEM-C; Church & Burke, 1994; Tellegen & Waller, 2008)

Goldberg, 2007). This strong predictive capacity stems primarily from the primary trait scales, which have been developed in the tradition of trait-realism on the basis of deductive-inductive cycles of construct-development (see Tellegen & Waller, 2008). Accordingly, the MPQ features 11 primary trait scales, a number of which assess constructs not specifically measured by other personality instruments, and that demonstrated clinical relevance (e.g. Wellbeing, Stress Reaction, Alienation, Harm Avoidance, Absorption; Arseneault, Moffitt, Caspi, Taylor, & Silva, 2000; DiLalla et al., 1993; Krueger, Caspi, & Moffitt, 2000; Miller, 2003; Sellbom & Ben-Porath, 2005; Sellbom, Ben-Porath, & Bagby, 2008a)

Here, we describe a study into measurement invariance (Meredith, 1993) of the Dutch Brief Form of the Multidimensional Personality Questionnaire (chapter 2; Tellegen & Waller, 2008) across a general and a clinical sample from a treatment facility specialized in the treatment of personality disorders, examining the least researched marginal (i.e. A/C) of Table 4.1. For examining measurement invariance we used both Confirmatory Factor Analysis (CFA) and Exploratory Structural Equation Modeling (ESEM; Asparouhov & Muthén, 2009). Measurement invariance for each single scale of the MPQ was tested with CFA. For testing measurement invariance in the full multidimensional model encompassing all 11 primary traits we used ESEM. For these latter analyses we selected ESEM over CFA because CFA models applied to broadband personality questionnaires have been shown not to fit response structures. Broadband personality models do not adhere to simple structure (i.e. items only load on the factor they were intended to measure), which means that numerous cross-loadings need to be defined for the models to fit well[2]. Exploratory Factor Analysis (EFA) might seem to solve this issue, because within EFA the full loading matrix is estimated. However, EFA does not provide a means to test for measurement invariance. ESEM provides a way to estimate groups of EFA factors within a CFA framework, making it possible to formally test measurement invariance while also estimating the necessary cross-

[2] In fact, for broadband personality models, simple structure is not expected nor particularly sought for. First, the hierarchical structure of personality, with higher-order constructs explaining the relation between lower-order constructs, makes that there is some overlap across indicators. Second, although scales might be clearly unidimensional, indicators generally are not. Responses on specific trait indicators can therefore be influenced by more than one trait. For example, a particular MPQ Traditionalism item states "I am disgusted by dirty language". As can be derived from supplementary Table 4.2, this item has a notable negative cross-loading on Aggression. The explanation for this observation has face validity. Uttering dirty language speaks against the moral values of some, but can also be viewed as an aggressive act.

loadings in the models. For a more elaborate description of testing measurement invariance with ESEM see Marsh et al. (2010).

In sum, we present an examination of measurement invariance of the MPQ-BF-NL across a general population and a clinical sample in order to determine to what extent the response structures are identical across these samples. When a measure is invariant, one can conclude that the same constructs are assessed in the same way in the different samples. Also, absence of measurement invariance points to item bias or DIF, and provides researchers and assessors with knowledge about pitfalls in employing the measure in populations on which the measure was not developed. We therefore had two specific aims: (1) to determine the extent to which responses on the MPQ-BF-NL are structured the same across a general and a clinical sample, and (2) to determine the influence of any item bias or DIF on scale scores.

Method

Samples

Clinical sample

The clinical sample consisted of patients (*N* = 365) who were referred for an intake assessment to ´De Viersprong´, a specialized clinic for the assessment and treatment of personality disorders, between 2010 and 2011. Protocols were excluded when variable and/or fixed item-response patterns exhibited marked content inconsistencies, as indicated by a very high score on the Variable Response Inconsistency (VRIN) scale and/or by either a very high or a very low score on the True Response Inconsistency (TRIN) scale. Accordingly, protocols meeting any of the following criteria were excluded: (a) a VRIN score > 3 *SD* above the mean score, (b) a TRIN score > 3.21 *SD* above the mean score or < 3.21 *SD* below the mean score, and (c) a response pattern reflecting both VRIN and TRIN inconsistencies indicated by a VRIN score > 2 *SD* above the mean and a TRIN score > 2.28 *SD* above the mean or < 2.28 *SD* below the mean. Of note, cut-off values were selected to yield the same percentage of rejected protocols for VRIN and TRIN (see also Patrick et al., 2002). In the clinical sample no records were excluded because all VRIN and TRIN scores were within normal range. The sample consisted of 131 (35.9%) men and 234 women, with a mean age of 34.17 years (*SD* = 11.04). SCID-II interview data (Gibbon, Spitzer, & First, 1997) were available for 337 (92%) patients. 130 (39%) of the patients met criteria for a DSM-IV

Personality Disorder (PD). Of these 4 (3%) met criteria for Cluster A PD, 56 (43%) for Cluster B PD, 67 (52%) for Cluster C PD, and 14 (11%) for PD Not Otherwise Specified.

General sample

An equal size subsample (*N* = 365) from a larger representative Dutch sample (*N* = 1,055) was used as the reference group. Two considerations guided our decision to use a subset instead of the full representative sample for our analyses. First, in both MGCFA and MGESEM, the estimation of parameters depends on total sample size over groups. With (greatly) unequal sample sizes, parameter estimates would have been substantially biased towards the larger general sample. Second, by using a subset we were able to match the general sample to the clinical sample in terms of gender and age in order to control for effects of these sample characteristics. Within this constraint, records were randomly selected from the larger general population. The full sample is described in chapter 2, and consisted of 1,060 participants, who took part in a panel survey conducted by *Flycatcher*, a full-service online research company. In order to achieve a good representation of the adult Dutch population (Centraal Bureau voor de Statistiek, 2009), we stratified the sample on gender, age, educational level, and county. Five participants were removed due to high VRIN and/or TRIN scores as defined above. The subsample used here (*N* = 365) consisted of 131 (35.9%) men and 234 women, with a mean age of 34.06 years (*SD* = 11.19).

Measure

The Dutch brief form of the MPQ (MPQ-BF-NL; chapter 2) was used. The 11 MPQ-BF-NL primary trait scales are measured by 12 binary items each. Including 3 extra items to assess VRIN and TRIN scales, the full MPQ-BF-NL consists of 135 items. The primary trait scales coalesce into three higher-order factors: Positive Emotionality (PEM), Negative Emotionality (NEM), and Constraint (CON). PEM comprises Wellbeing (WB), Social Potency (SP), Achievement (AC), and Social Closeness (SC). NEM comprises Stress Reaction (SR), Aggression (AG) and Alienation (AL), and CON comprises Control (CO), Harm Avoidance (HA) and Traditionalism (TR). Absorption is not specifically allocated to any of the three higher-order factors.

The MPQ-BF-NL has demonstrated adequate to good psychometric properties in diverse samples, that are generally quite similar to those of the U.S. brief form of the MPQ (MPQ-BF; Patrick et al., 2002); both in terms of reliability as well as in terms of its higher-order

structure. As can be inferred from Table 4.2, reliabilities were somewhat higher for the clinical sample employed in this study than for the representative sample.

Table 4.2. Cronbach's alphas and description of high scorers for the MPQ-BF-NL scales

	Cronbach's α		
Scale	General (*N* = 1,055)	Personality pathology (*N* = 365)	Description of a high scorer
Wellbeing	.80	.83	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	.79	Works hard; enjoys demanding projects and working long hours
Social Closeness	.81	.84	Is sociable, likes people, and turns to others for comfort
Stress Reaction	.84	.74	Is nervous, vulnerable, sensitive, prone to worry
Aggression	.73	.77	Hurts others for own advantage; will frighten and cause discomfort for others
Alienation	.82	.85	Feels mistreated, victimized, betrayed, and the target of false rumors
Control	.75	.84	Is reflective, cautious, careful, rational, planful
Harm Avoidance	.72	.78	Avoids excitement and danger; prefers safe activities even if they are tedious
Traditionalism	.70	.67	Desires a conservative social environment; endorses high moral standards
Absorption	.80	.79	Is responsive to evocative sights and sounds; readily captured by entrancing stimuli

Testing for measurement invariance

Multiple Group CFA and multiple group ESEM with categorical outcomes

Measurement invariance was tested by using Multiple Group CFA (MGCFA) and Multiple Group ESEM (MGESEM). When using categorical outcomes within Structural Equation Modeling (SEM), parameter estimates can be biased (Wirth & Edwards, 2007). With binary or ordinal indicators, this problem can be resolved by viewing the categorical responses as discrete representations of continuous latent responses. In the dichotomous case of the MPQ this means that below a certain point of the latent response continuum a participant would reject the statement, while above the cut-point the participant would endorse it. Tetrachoric correlations between responses are used and several estimators for model parameters are available. We used a robust method for estimating Weighted Least Squares (WLSMV; Muthén et al., 1997). Furthermore, target rotation was used allowing for factor covariances (items within a scale had target loadings for the same factor, and non-target loadings for the other factors). All analyses were carried out in Mplus version 7.2 (Muthén & Muthén, 2012). The present study utilized the Theta parameterization because it provides

the possibility to either restrict residuals to be equal across groups or to freely estimate some of the residuals across groups. This practice most closely resembles measurement invariance practices in MGCFA and MGESEM with continuous outcomes.

Levels of measurement invariance

A scale can be measurement invariant at different levels (Meredith, 1993). As convention dictates, we tested increasingly restrictive models. In the least restrictive model, the same structural model is fitted in the different groups, allowing loadings and thresholds (or intercepts when continuous outcomes are used) to vary across groups. This type of variance is called 'configural invariance', indicating that if it holds the configural structure of the models is the same (i.e. same indicators are measures of the constructs). Configural invariance leaves open whether DIF is present across samples. In order to examine DIF, more restrictive models need to be tested. The 'weak invariance' and 'strong invariance' models test whether the loadings (discrimination) and thresholds (difficulty) of the items can be said to be the same across the different groups. The most restrictive model, in which all measurement parameters are equal across groups is termed the 'strict invariance model'. In this model loadings, thresholds and residuals are fixed to be equal over groups. The structural parameters (i.e. factor means, variances and covariances) can be validly compared when strict invariance holds. When strict invariance (or any other level of measurement invariance) does not hold, one or more of the fixed parameters may be freely estimated across groups in order to obtain the desired model fit. Items for which one or more of the parameters need to be freely estimated can be said to contain DIF, a term commonly used in the realm of Item Response Theory (IRT), but that can also be used for describing non-invariance of item parameters in CFA (Kim & Yoon, 2011). For illustration, the chapter 4 supplementary material contains an example of how DIF may influence measurement.

Evaluating model fit

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 inspected. For the absolute fit of models RMSEA values smaller than .08 and smaller than .05 were considered to indicate acceptable and good fit, respectively (MacCallum et al., 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 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. Finally, we also display the χ^2 statistic, but we do not use it for model evaluation, as it is well established that even well-fitting models are rejected in large samples (Thompson, 1995).

Testing measurement invariance in single scales and in the full model

Whereas personality scales are generally evaluated on the single scale level, broadband models of personality like the MPQ claim to describe the full domain of personality. We therefore tested for measurement invariance in all MPQ primary trait scales separately, as well as for the full multidimensional model. When strict invariance did not hold, partially strict invariant models were derived. Although adjusting a model to secure partial invariance is inherently post-hoc and subjective to judgment, it can provide valuable information about DIF and about distortions in the interpretation of raw scale scores. Moreover, it has been shown that valid comparisons between groups are possible when only partial invariance applies (Byrne, Shavelson, & Muthén, 1989).

To derive a model with equivalent fit to the configural model, the strict invariant model was adjusted. As individual parameter estimates are sensitive to all other changes in the model, we hold this procedure for more parsimonious than sequentially securing weak measurement invariance, strong measurement invariance, etc. In freeing parameters we gave priority to freeing thresholds over loadings, and loadings over residuals of items, respectively. We adopted this decision rule because threshold DIF tends to be 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 markedly (in practice always > 4.5) larger than for parameters higher up 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 no consensus about the practice of separately fixing and freeing threshold and loading parameters (see for example Millsap & Yun-Tein, 2004, who support this, and Muthén & Asparouhov, 2002, who do not). Because of the conceptual likeness of loadings and thresholds on the one hand, and discrimination and difficulty parameters from IRT on the other, we treated the freeing of loadings and thresholds separately.

Results

Tests of measurement invariance and DIF in single scales

8 of the 11 primary trait scales showed acceptable to good fit for the configural model. For Social Potency, Achievement and Control, fit was near acceptable. For 4 out of the 11 scales (i.e. Social Potency, Alienation, Control and Absorption), the strict invariant model fitted as well as the configural model, implying that there was no DIF apparent in these scales. As reduction in fit from the configural to the strict model was significant for the other 7 scales, DIF was present in these scales. To achieve partially strict invariance, across scales a total of 13 measurement parameters out of a total of 396 (3%) was freely estimated across samples: 10 thresholds (8%), 2 loadings (2%) and 1 residual (1%). The freely estimated parameters belonged to 13 (10%) of the items. Table 4.3 displays an overview of the number of specific DIF items within each of the MPQ scales. Fit information of the single scale models is provided in supplementary Table 4.1 and more information on DIF in the single scale models can be found in Table 4.4 and in the chapter 4 supplementary Results section.

Table 4.3. Number of DIF parameters for scale by scale partial strict invariant models

	DIF type			
Scale	Threshold	Loading	Residual	Total DIF items
Wellbeing	1	0	0	1
Social Potency	0	0	0	0
Achievement	2	0	0	2
Social Closeness	1	1	1	3
Stress Reaction	2	1	0	3
Aggression	1	0	0	1
Alienation	0	0	0	0
Control	0	0	0	0
Harm Avoidance	1	0	0	1
Traditionalism	2	0	0	2
Absorption	0	0	0	0
Total	10	2	1	13

To rule out multidimensionality as a confound, we fitted two-factor ESEM models on the responses on the scales that showed suboptimal fit for the one-factor models (i.e. for Social Potency, Achievement, and Control). For all three scales configural two-factor models fitted significantly better than the one-factor models, and strict invariance models fitted as well as the configural invariance models, indicating absence of DIF. The observation of absence of DIF was in line with the analyses on the one-factor models for Social Potency and Control, but for Achievement it was not. In the unidimensional case DIF was observed in two items.

The absence of DIF in the two-factor model may be understood in terms of the differential pattern of mean differences in factor scores between the general and clinical sample. While the mean score on the first factor was somewhat lower for the clinical sample than for the general sample (Z = -.35, SE = .11, p < .001), it was much higher on the second factor (Z = .84, SE = .15, p < .001). The two items that showed DIF in the one-factor analyses loaded on the second factor, which may have resulted in lower thresholds in the clinical sample when the scale was considered to be unidimensional. More elaborate description of these analyses can be found in the chapter 4 supplementary material.

Table 4.4. Unstandardized parameter estimates for measurement parameters that were freely estimated across samples for scale by scale partial strict invariant models

Scale	Item number[a]	Parameter	General	Clinical
Wellbeing				
	235	Threshold	0.05	-0.66
Achievement				
	71	Threshold	-0.02	-0.89
	150	Threshold	0.13	-0.78
Social Closeness				
	4	Residual	1.00	0.26
	29	Loading	0.46	0.87
	216	Threshold	0.09	-0.50
Stress Reaction				
	36	Threshold	-0.49	0.09
	193	Threshold	-0.71	0.22
	258	Loading	1.45	0.65
Aggression				
	239	Threshold	0.10	0.92
Harm Avoidance				
	228	Threshold	-1.06	-0.47
Traditionalism				
	118	Threshold	-0.22	-0.59
	240	Threshold	-1.24	-0.64

[a]Numbers correspond to the 276-item full length version of the MPQ.

Tests of measurement invariance and DIF in the multidimensional model

As expected, and in line with previous investigations (see for example Marsh et al., 2010), the simple structure full CFA-model did not adequately fit the data in either the general or clinical sample. However, ESEM models showed good fit in both the general and clinical samples (see Table 4.5). The lower part of Table 4.5 displays the results of tests for measurement invariance of the full MPQ. The MPQ appeared strict invariant across general and clinical samples (lower bound 90% CI RMSEA +.001; higher bound 90% CI RMSEA +.001; CFI -.010; TLI -.007 compared to the configural model). Consequently no significant amount of DIF was observed in the full model. Parameter estimates and other detailed information on the strict invariant ESEM model are provided in supplementary Tables S.4.2 and S.4.3.

Table 4.5. Fit for (Multiple Group) CFA(150) models

Model	*N*	*df*	χ^2	RMSEA 90% CI	CFI	TLI
Single group models						
CFA general original (full) sample	1,055	8,459	14,572	[.025, .027]	.788	.784
CFA general stratified sample	365	8,459	9,826	[.019, .023]	.816	.812
CFA clinical sample	365	8,459	10,148	[.022, .025]	.816	.812
ESEM general original (full) sample	1,055	7,249	7,970	[.008, .011]	.975	.970
ESEM general stratified sample	365	7,249	7,467	[.003, .013]	.971	.965
ESEM clinical sample	365	7,249	7,626	[.008, .015]	.959	.951
Multiple group models						
ESEM configural invariance	2x365	14,498	15,090	[.008, .013]	.964	.957
ESEM weak invariance	2x365	15,829	16,547	[.008, .013]	.957	.953
ESEM strong invariance	2x365	15,818	16,502	[.008, .013]	.959	.955
ESEM strict invariance	2x365	15,950	16,716	[.009, .014]	.954	.950

Note. CFA = Confirmatory Factor Analysis; ESEM = Exploratory Structural Equation Modeling; RMSEA = Root Mean Square Error of Approximation; CI = Confidence Interval; CFI = Comparative Fit Index; TLI = Tucker Lewis Index.

Distributional differences of traits in the general and clinical samples

Although strict invariance was tenable for the full model, interpretation of distributional differences of the trait scores between the samples may be not straightforward for two reasons: (1) at the single scale level DIF was present; (2) at the full multidimensional level the trait constructs may have drifted because the full loading matrix (i.e. all possible cross-loadings) is estimated in ESEM. To get an idea of the impact of DIF and cross-loadings on the scale scores, Figure 4.1 displays boxplots representing the distribution of scores for each of the primary trait scales in the clinical sample as compared to the general sample. The scores of the general sample are defined to have a mean of zero and a standard deviation of one. Boxplots consequently represent the distributions of the trait scores in the clinical sample as deviations from this standard normal distribution. For each trait three boxplots are given, including (a) raw scale score distributions (white); (b) score distributions for the final models from the single scale analyses (light gray), and (c) factor score distributions for the strict invariant model of the full model analyses (dark gray).

Mean differences in scores between general and clinical samples

Regardless of type of score, individuals within the clinical sample scored consistently much lower on Wellbeing (range M = -1.79 - -2.02) than individuals within the general sample, indicating that people in the clinical sample were less happy, had less cheerful dispositions and felt less good about the self and the future. Also, in general, individuals within the clinical sample scored substantially higher on Stress Reaction (range M = 0.69 – 1.25) and lower on Traditionalism (range M = -0.60 – -1.04), indicating that people in the clinical sample experienced more stress and anxiety than people from the general population

sample and that this group was less conservative than the general population. Smaller, but consistently significant differences were observed for Social Closeness (range *M* = -0.29 – -0.46), Alienation (range *M* = 0.57 – 0.63), and Control (range *M* = -0.34 – -0.41). For Harm Avoidance a moderately lower mean score was observed within the single scale analyses (*M* = -.53), while significant but only small differences were observed for the other types of scores (range *M* = -.11 - -.18). However, the differences between raw scale scores

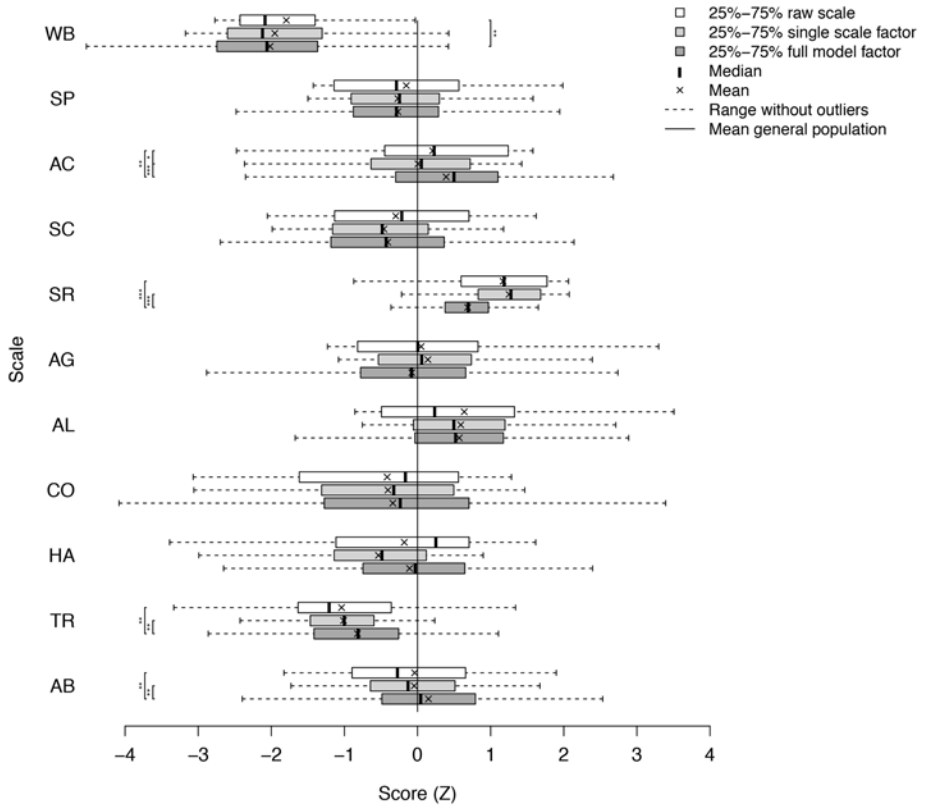

Figure 4.1. Score distributions for the clinical sample compared to the general sample (with *M* = 0; *SD* = 1). WB = Wellbeing; SP = Social Potency; AC = Achievement; SC = Social Closeness; SR = Stress Reaction; AG = Aggression; AL = Alienation; CO = Control; HA = Harm Avoidance; TR = Traditionalism; AB = Absorption. Differences between mean differences across type of score are tested with ANOVAs and post-hoc tests.
* false discovery rate (fdr) corrected $p < .01$. ** fdr corrected $p < .01$; *** fdr corrected $p < .001$.

and factor scores were not significant. For Achievement raw scale and multidimensional factor score means were larger for the clinical sample than for the general sample (range *M*

= 0.21 – 0.39), while no significant differences were observed for single scale factor scores (*M* = 0.01). No, or minimal mean differences were observed for Social Potency (range *M* = -0.15 – -0.27), Aggression (range *M* = -0.08 – 0.14) and Absorption (range *M* = -0.05 – 0.15) across samples.

Variation in mean differences for raw scale scores compared to single factor scores was negligible. Only for Achievement the difference between means was significantly smaller when single scale factor scores were considered, than when raw scale scores were considered, meaning that DIF might have affected raw scale score interpretation for Achievement. Variation in mean differences for raw scale scores compared to full model factor scores was more extensive. A more elaborate description of this variation can be found in the chapter 4 supplementary Results section. Because of the unpredictable influence of cross-loadings in the full multidimensional ESEM model, we advise against interpreting factor scores of this model.

Discussion

The current study involved a formal comparison of the structural equivalence of the MPQ-BF-NL across a general and a clinical sample. As was shown, the MPQ-BF-NL measured its comprising constructs nearly the same way in the general and clinical sample. Strict invariance pertained to the full multidimensional model, while some degree of DIF was present at the single scale level. These results lend further support to the notion of a common structure of normal and pathological personality.

With adequate measurement invariance established, group averages on trait standings can be meaningfully compared. As might be expected, the clinical sample was characterized by very low Wellbeing and very high Stress Reaction scores. Also, these individuals were generally low on Traditionalism, moderately low on Social Closeness, Control and Harm Avoidance, and moderately high on Alienation. Patients exhibited normal range scores for Social Potency, Aggression, and Absorption. Scores on Achievement could not be straightforwardly be compared, as a result of DIF. The differences in scores on this trait depended on the specific aspect of the trait. Individuals from the clinical sample generally reported less ambitious striving while at the same time reporting to put higher demands on themselves.

To our knowledge, this is the first study that tested measurement invariance of a broadband personality inventory across general and clinical samples. Within the extensive body of work on the structural similarity of normal and pathological personality this comparison is a relatively neglected one (indicated by the A/C marginal in Table 4.1). Future studies are needed in order to replicate these findings with other clinical samples and other personality instruments. Similarly, we recommend more measurement invariance studies with regard to instruments directly aimed at pathological aspects of personality traits (i.e. the B/D marginal in Table 4.1). Such instruments – for example the Personality Inventory for DSM-5 (PID-5; Krueger et al., 2012) and the Computerized Adaptive Test of Personality Disorder (CAT-PD; Simms et al., 2011) – are of specific importance for assessing the dimensional structure proposed in Section III of DSM-5.

Our findings have relevance for clinical practice by psychometrically establishing that scores on MPQ primary trait scales can be adequately used in a clinical sample. Moreover, a number of MPQ scales tap clinically relevant constructs that are not specifically captured by other personality instruments (e.g. Stress Reaction, Wellbeing, Alienation, Harm Avoidance, Absorption; Tellegen & Waller, 2008). While the current trait literature shows increasing convergence with regard to higher-order structure (e.g. two, three and five-factor models being hierarchically related rather than opposed to each other; DeYoung, 2006; Digman, 1997; Saucier, 2008), no clear convergence can yet be observed regarding lower-order structure. As this more fine-grained level is of particular importance for clinical assessment, we believe this to be an important area for future research.

A number of limitations warrant specific discussion. First, a rather specific, albeit heterogeneous clinical sample was employed. Our clinical sample consisted of patients specifically referred for an evaluation of personality pathology. The roughly four out of ten patients who met diagnostic thresholds for DSM-IV were predominantly assigned cluster B (but not Antisocial PD) and/or cluster C diagnoses. It remains to be established whether the (partial) measurement invariance that we established will generalize to samples encompassing wider ranges of psychological complaints (for example other clinical disorders) and of different levels of severity. Likewise, the generalizability of the pattern of mean differences between our general and clinical sample will be limited. For example, the absence of cluster A patients in our sample may explain why no group differences in Absorption were observed; this may be quite different in clinical samples that also contain a substantial proportion of patients with thought disorder (Clancy, McNally, Schacter,

Lenzenweger, & Pitman, 2002; Sellbom & Ben-Porath, 2005). A second limitation pertains to the fitted multidimensional ESEM model, in which all possible cross-loadings were estimated. As a consequence of these cross-loadings, latent constructs exhibited some apparent drift from their intended interpretation. We therefore caution against the evaluation of ESEM factor scores derived from responses based on broadband personality measures.

Conclusion

The current study showed that normal personality as measured by the MPQ-BF-NL is near identically structured in a general and a rather heterogeneous clinical sample. Quantitative differences however were marked and in line with conceptual expectations. Our findings lend further support to the use of the MPQ-BF-NL in clinical settings.