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](#)

Chapter 7

General discussion

The aims of this thesis were twofold: (1) developing and validating a Dutch brief form of the Multidimensional Personality Questionnaire (MPQ-BF-NL), and (2) making substantive contributions to the field of personality. Both aims were simultaneously achieved by using the newly developed MPQ-BF-NL for inquiries into cross-cultural differences in personality, personality in general and clinical samples, and personality and its associations with brain structure. In the remainder of this thesis I will first discuss the outcomes of the development and validation of the MPQ-BF-NL and then the contributions this undertaking yielded to furthering knowledge in the field of personality. Included in these discussions are the limitations of the present body of research as well as related considerations for next steps in the ongoing refinement of the MPQ(-BF-NL), and for future more fundamental examinations into personality.

Development and validation of the MPQ-BF-NL

Chapters 2, 3 and 4 dealt directly with the development and the measurement properties of the MPQ-BF-NL. In chapter 2 we described the development of the MPQ-BF-NL and provided basic validation data on the resulting instrument. Internal consistencies of the MPQ-BF-NL were generally good, and comparable to the U.S. brief version. Also, correlations with the U.S. full length scales were high, and its higher-order structure converged with the structure of the original MPQ. Convergent and divergent patterns with scales from other measurement instruments were consistent with prediction, with Positive Emotionality relating to social and activating behavior, Negative Emotionality relating to anxiety, and Constraint to reversed impulsivity and externalizing behaviors.

Integration and extension of explorations into measurement invariance across different samples

Chapters 3 and 4 both presented studies into measurement invariance of the instrument across U.S. and Dutch samples (chapter 3) and general and clinical samples (chapter 4). Measurement invariance is important when one intends to use an instrument in different populations (e.g. cultures; general/patient populations), because it speaks to the question whether the same constructs are measured in the same way in the different populations. Also, it is only possible to validly compare trait standings of people from different populations when adequate measurement invariance is established. Furthermore, if one aims to generalize research results based on data from a specific sample to other samples (e.g. generalize knowledge on personality gathered by the original MPQ on U.S. data to the

Dutch context), measurement invariance should hold across these samples. Lack of measurement invariance is apparent through Differential Item Functioning (DIF) across populations. In chapter 3 we showed that across U.S. and Dutch general population samples DIF was present in all scales of the MPQ-BF-NL, affecting 19% of the items (evaluating the scales one by one). DIF was most extensive for Social Potency and Aggression. Most scale score interpretations were not affected by DIF but Wellbeing, Social Potency, Achievement and Control were. In chapter 4 we showed that DIF was only to a very limited extent present across a general and a clinical sample, with 10% of the items showing bias when the scales were evaluated one by one. Only the interpretation of Achievement was affected by DIF.

Throughout this thesis it is assumed that there is no DIF in the MPQ-BF-NL for gender: males and females are treated as belonging to the same population with the same response structures. However, this common treatment is unwarranted when adequate measurement invariance does not hold. We therefore tested whether measurement invariance holds for gender within our normative Dutch sample. These analyses are described in the supplementary material of chapter 7. In short, when considered multidimensionally, the MPQ-BF-NL was not affected by significant bias for different gender types. Unidimensional per scale tests revealed that 4% of items was affected by (threshold) DIF. DIF did not affect scale score interpretation as reflected in gender differences in mean scale scores.

In conclusion, the explorations into measurement invariance of the MPQ-BF-NL across different populations showed that least bias (in 4% of the items) was present in the instrument when males and females were compared. Somewhat more bias was evident in the measurement of the constructs for general as compared to clinical samples (in 10% of items). Most bias pertained to comparisons between U.S. and Dutch samples (in 19% of items). These results do not reflect the extent to which the different samples differ in their personalities; after all, differences in mean trait standings are most pronounced when general and clinical samples are compared (see further below), but they do reflect the extent to which the different populations seem to be alike in their interpretation of the items comprising the scales. The extent to which DIF was present does not warrant against raw scale score comparisons between males and females, or between individuals from general versus clinical samples. Raw scale score comparisons between people from the U.S. versus the Netherlands should be made with more caution.

Opportunities for model and instrument refinement

Although the MPQ-BF-NL generally showed good psychometric properties, development of the instrument does not (ever) cease. The results presented in chapters 2, 3 and 4 of this thesis provide the following opportunities for further model and instrument refinement: (1) the internal structure of some scales may be improved as indicated by suboptimal internal consistencies (chapter 2) and signs of multidimensionality (chapters 3 and 4); (2) the association between primary trait scales and the higher-order traits suggests that the hierarchical structure of the trait constructs do not adhere to simple structure (chapter 2). Simple structure is not expected nor particularly sought for in broadband personality models (as argued in chapter 4), but cross-loadings should make theoretical and empirical sense.

Optimizing reliability and (construct)validity

The most widely used indicator of the reliability of a scale is Cronbach's alpha, which is an estimator for internal consistency. More specifically, alpha estimates the ratio of true-score and total-score variances. The higher alpha the higher the mean inter-item correlation between the items is[1]. In the Netherlands, the Commissie Testaangelegenheden Nederland (COTAN) provides reviews of tests and questionnaires (Egberink, Janssen, & Vermeulen, 2016) as does the Buros Institute of Mental Measurements in the U.S. (Carlson, Geisinger, & Jonson, 2014). For important decisions (e.g. personnel selection, referral to special education, admission to inpatient care), the COTAN considers values of Cronbach's alpha exceeding .80 acceptable, and exceeding .90 good. For less important decisions (e.g. monitoring, general descriptive use, vocational guidance, therapy-indication), values over .70 and .80 indicate acceptable and good performance respectively. All of the primary trait scales of the MPQ-BF-NL show acceptable internal consistency for less important decisions. However, not all scales exceed the threshold for good reliability for less important decisions or acceptable reliability for important decisions. Based on our normative Dutch sample (see chapter 2) this concerns Achievement (α = .76), Aggression (α = .73), Control (α = .75), Harm Avoidance (α = .72), and Traditionalism (α = .70). For most of these scales internal

---

[1] Although Cronbach's alpha is the most widely used indicator for reliability, its usefulness is also questioned: it underestimates internal consistency and is uninformative about the internal structure of the scale (Sijtsma, 2009). A conceptually and statistically more appealing indicator is McDonald's omega which gives an estimate of the general factor saturation of a scale (Zinbarg, Revelle, Yovel, & Li, 2005). Throughout this thesis we limited ourselves to reporting Cronbach's alpha because of its popularity. In future writings we intend to report McDonald's omega as well.

consistencies of the U.S. brief form were also relatively weak, but for Harm Avoidance (U.S. α = .76) and Traditionalism (U.S. α = .78) they were notably stronger in the U.S. brief form.

There are several potential sources for relatively low values of alpha. First, internal consistencies increase when test length increases. Internal consistencies as low as .75 are not uncommon for scales of 12 items from tests that are considered to have good predictive ability (Thalmayer, Saucier, & Eigenhuis, 2011). Second, one or more items may show low inter-item correlations with the other items in the scale and consequently with the total scale. If this is the case, dropping these items should result in higher internal consistency estimates. Third, relatively little variance in the attribute limits the magnitude of alpha. Fourth, the size of alpha is restricted when overall inter-item correlations between items are relatively low: the broader a construct, the lower the inter-item correlations of the scale measuring it. Likewise, when a construct becomes narrower, its inter-item correlations become higher. In the extreme they may become so high as to a point that the items become redundant. Hence, internal consistencies should not be too high either, because the items then merely constitute synonyms. Conversely, when inter-item correlations become very low the assessed construct can become so broad that it incorporates such a large domain that the scale also loses much of its informativeness. For broader constructs it has been suggested that the inter-item correlations may be as low as 15-.20, while for more narrow constructs inter-item correlations around .40-.50 are more suitable (Clark & Watson, 1995). Fifth, a large variability in inter-item correlations may indicate multidimensionality which restricts internal consistency.

Multidimensionality of a scale is not only a possible source of restricted reliability. It is also a problem for (construct)validity. Multidimensionality of a scale hampers straightforward interpretation of the scores, because more than one construct is being measured. As mentioned above, a relatively large variance in inter-item correlations is a sign of multidimensionality. However, to see whether a multidimensional conceptualization of the association between the items is more suited than a unidimensional conceptualization, one-factor models should be tested against models with more than one factor.

As mentioned above Achievement (α = .76), Aggression (α = .73), Control (α = .75), Harm Avoidance (α = .72), and Traditionalism (α = .70) showed internal consistencies below the threshold for good reliability for less important decisions or acceptable reliability for important decisions in our normative Dutch sample. Also, chapters 3 and 4 provide

indications that Social Potency, Achievement, and Control may suffer from multidimensionality. In order to develop hypotheses about the sources of restricted reliability in some of the scales and to shed light on how to possibly improve unidimensionality, additional descriptives and analyses for the normative Dutch sample are provided in the supplementary material of chapter 7. Summaries on the inter-item correlations as well as descriptives of the primary trait scale scores of the MPQ-BF-NL in the normative Dutch sample were used to speculate about possible sources of suboptimal reliability. Exploratory Factor Analysis (EFA) was used to explore possible multidimensionality. The resulting suggestions for instrument refinement are summarized in Table 7.1.

Table 7.1. Suggestions for refinement of primary trait scales with suboptimal reliability or possible multidimensionality.

Scale	Concern α	dim.	Suggestions for refinement
Social Potency		x	Select/formulate cross-cultural bias free items
Achievement	x	x	Investigate structural connectedness between perseverance, demands and ambition
Aggression	x		Develop easier items
Control	x	x	Develop harder items
Harm Avoidance	x		Develop easier items / replace forced-choice items by True/False
Traditionalism	x		Improve conceptualization by distinguishing or choosing between liberal and conservative interpretations of morality

Note. dim. = dimensionality.

For the scales for which internal consistency was exclusively the concern (Aggression, Harm Avoidance, and Traditionalism) restriction of range has may have played a role in restricting Cronbach's alpha. For Aggression the distribution was skewed to the right (the modal score being low), while for Harm Avoidance and Traditionalism the distributions were skewed to the left (the modal score being high). We may consider developing items that are more readily endorsed (i.e. are easier) for Aggression, and that are less readily endorsed (i.e. are harder) for Harm Avoidance and Traditionalism in order to increase variance in scale scores and possibly enhance inter-item correlations. For Harm Avoidance a further suggestion is to replace forced-choice items by True/False items since the inter-item correlations of the forced choice items are notably weaker than those for the True/False items. For Traditionalism it seems worthwhile to reconsider the conceptualization of the construct. Although no signs of multidimensionality appeared in chapters 3 and 4, different meanings of the term *'morality'* may have engendered low inter-item correlations. Some items refer to more conservative values while other items refer to

more liberal values. In the future development of the scale we should distinguish between these two meanings of morality or possibly opt for the conservative interpretation since that is the interpretation of the construct that has been leading in the development of the scale.

Although the internal consistency of the Social Potency scale was strong, chapters 3 and 4 provided indications for multidimensionality. The additional analyses showed that multidimensionality was not a major problem for the interpretation of scale scores within the normative Dutch sample. However, it was reflected in cross-cultural DIF (with separate factors for leadership and persuasiveness). Items should be selected or developed that do not show DIF across U.S. and Dutch samples in order to ascertain bias free comparisons.

For both Achievement and Control relative weak internal consistencies were observed and indications in chapters 3 and 4 for multidimensionality were present. The signs of multidimensionality in Control could be attributed to one of the items being notably harder than all other items, thus forming a separate difficulty factor. For the internal structure, and also for discriminative purposes it may be worthwhile to develop more difficult Control items. Within the Achievement scale two factors with different content could be distinguished, both in the U.S. representative and Dutch normative sample. In both samples one of the factors could be termed perseverance, while the other factor represented demands and ambitious striving in the Dutch and U.S. data respectively. Future investigations in the (cross-cultural) structural connectedness between perseverance, demands, and ambition are warranted to decide how to conceptualize and further develop the Achievement scale.

The hierarchical structure of the MPQ(-BF-NL)

The MPQ model provides a hierarchical operationalization of personality, with the primary trait scales coalescing in higher-order, broader constructs. PCAs of the primary trait scales have yielded three or alternatively four factors (Tellegen & Waller, 2008). The three-factor solution consists of Positive Emotionality (PEM), Negative Emotionality (NEM) and Constraint (CON), while the four-factor solution splits PEM into its agentic and communal subcomponents (PEM-A and PEM-C respectively). Throughout this thesis we focused on the three-factor solution. In chapter 2 we presented three-factor solutions of EFAs on the primary trait scale scores of the MPQ-BF-NL for the normative Dutch and U.S. samples (see

Table 2.3). The patterns of loadings generally conformed to expectations, but simple structure was not achieved.

Notable cross-loadings were observed for Wellbeing (negative on NEM), Social Potency (negative on CON), and Aggression (negative on CON). These cross-loadings are consistent with earlier findings (Tellegen & Waller, 2008). Cross-loadings can be explained by (a) the hierarchical nature of constructs in the personality domain, and (b) choices made regarding the rotation of factors. The negative loading of Aggression on CON can be understood in terms of its place within the personality hierarchy (see Figure 1.1). As illustrated, PEM, NEM and CON coalesce in even broader constructs (i.e. Stability, Plasticity, and eventually into Evaluation). Covariance between NEM and CON can be explained by the construct Stability. Higher Aggression is indicative of lower Stability and therefore higher NEM and lower CON. The cross-loadings of Social Potency on CON and Wellbeing on NEM can be understood when one considers the issue of rotation of personality factors. Indicators of Wellbeing can be indicators of Stress Reaction at the same time, because of the pleasantness-unpleasantness dimension that is situated at a rotational angle of 45 degrees from the Positive and Negative Affect dimensions that can be thought to be the state counter parts of PEM and NEM respectively (Watson & Tellegen, 1985). The secondary loading of Social Potency on CON can be understood by considering the interpersonal circumplex of Wiggins where agreeableness and ambition/dominance are two orthogonal constructs. Social Potency can be thought of as a construct that is situated at a 45 degree angle of both these constructs with higher scores being indicative of higher ambition/dominance and lower agreeability (Wiggins, 1979). As shown in this thesis in chapters 1 and 2 and elsewhere (Tellegen & Waller, 2008), Social Potency has a negative correlation with Big Five Agreeableness which can be considered a lower-order factor of CON (Markon et al., 2005). In short, absence of simple structure in the higher-order structure of the MPQ(-BF-NL) is not a theoretical or measurement problem, but reflects the intricate nature of the domain of personality. It also provides opportunities to illustrate the interrelatedness of various personality models that have their own particular merits.

An observation that is not in accordance with previous findings is that a primary loading of Social Closeness on its higher-order construct PEM was absent. Table 2.3 shows that EFAs of the MPQ-BF-NL scales in the U.S. sample also resulted in a PEM factor that lacks a primary loading for Social Closeness, while it is present in solutions for the full length instrument. Therefore the absence of the loading in the MPQ-BF-NL cannot merely be

attributed to translation but seems to be due to characteristics of the specific item set comprising the MPQ-BF-NL. In contrast, the loading of Achievement on PEM is very strong, suggesting that the PEM factor of the MPQ-BF-NL rather represents the agentic aspect than the communal aspect of the domain. For future refinements of the MPQ-BF-NL it may be worthwhile to scrutinize the relationship between agentic and communal aspects of the PEM domain to guide the prospective substitution of items of the primary trait scales.

Other aspects to attend to in the refinement of the MPQ-BF-NL

In addition to modifications to the MPQ-BF-NL to improve its internal structure, some other aspects deserve attention for its further validation. First, to be able to make the instrument employable in a wide variety of applications norm data should be obtained for relevant populations. For use in clinical settings it is important to gather data from more diverse clinical samples. The clinical sample employed in the studies described in this thesis should be complemented by other clinical samples to ascertain the generalizability of the norms. Although our clinical sample was heterogeneous in the sense that it consisted of patients varying in nature and severity of their personality pathology, it was limited in that the patients came from only one assessment setting. Moreover, cluster A PD pathology and antisocial tendencies were underrepresented. It also remains to be established whether our norms are applicable to samples primarily suffering from clinical syndromes rather than personality pathology. Furthermore, the MPQ-BF-NL may also prove useful in industrial and recruitment settings, but norms for these domains remain to be established.

Second, so far we did not have data to establish the test-retest reliability of the scales. Test-retest reliability is informative about the stability of the rank-ordering of individuals over time. Because we assume to measure relatively stable individual differences, these differences should hold over time to a high degree (with lower stability coefficients when the time lag increases). Also, when a measure is applied to investigate change, for example to find out whether personality is affected by treatment, it is important to know whether the instrument is sensitive to change, i.e. whether observed differences can be attributed to real changes in personality or to random fluctuations. For calculating the test-retest reliability, the instrument must be administered to samples from relevant populations several times (for example once and then again after a year).

The further development and validation of the brief form of the MPQ-BF-NL will be advanced in an international effort. The intention is to simultaneously release brief forms of

the MPQ for use in the U.S. and in the Netherlands. Ideally these will be parallel versions, comprised of the same bias free items. That way, it should be possible to validly compare inferences made from the different forms. As a consequence, it is unlikely that the current item selection of the MPQ-BF-NL will be maintained. However, the studies presented in this thesis will inform item selection of future forms. To materialize these goals our research group has joined a collaboration comprising the author of the MPQ (i.e. Auke Tellegen - University of Minnesota) and other experts in the fields of test development and publishing (i.e. Richard Handel - Eastern Virginia Medical School, Martin Sellbom - University of Otago and Beverly Kaemmer - University of Minnesota Press).

Contributions to the understanding of personality

The explorations into the internal structure of the MPQ-BF-NL did not only inform about the validity of inferences made from scores obtained with the instrument, but they also added more generally to the understanding of personality. Taken together with the more fundamentally oriented studies presented in chapters 5 and 6 the following theoretical issues were attended to in this thesis: (a) personality differences between different populations. How does personality differ between gender types, cross-culturally and between general and clinical samples?; (b) the common structure of general and pathological personality. Is the notion of a personality disorder as a medical condition in which psychological functioning is qualitatively different from 'normal' function tenable, or should it be conceptualized in terms of quantitative differences as compared to 'normal' functioning?; (c) the realist basis of personality. Do personality dimensions refer to 'real' traits that exist independently of their measurement?

Personality differences across different populations
In our investigations into measurement invariance of the MPQ-BF-NL across various samples (i.e. males versus females, U.S. versus the Netherlands, and general versus personality pathology) we calculated scale scores corrected for extant item bias. This provided us the opportunity to validly compare mean trait standings between samples. In Figure 7.1 mean deviations from the Dutch general population mean for the various groups (females, males, U.S. and clinical) are depicted. The means were taken from the (partially) strict invariance models from the scale by scale measurement invariance analyses.

As can be inferred from Figure 7.1, the most substantial deviations occurred in the clinical sample. As might be expected, the clinical sample was characterized by very low Wellbeing and very high Stress Reaction scores. Also, people from the clinical sample were generally low on Traditionalism, moderately low on Social Closeness, Control and Harm Avoidance, and moderately high on Alienation. Although no deviation was observed for the total scale score of Achievement, the clinical sample was also characterized by putting high demands on themselves while reporting less ambitious striving than the general population sample (see chapter 4).
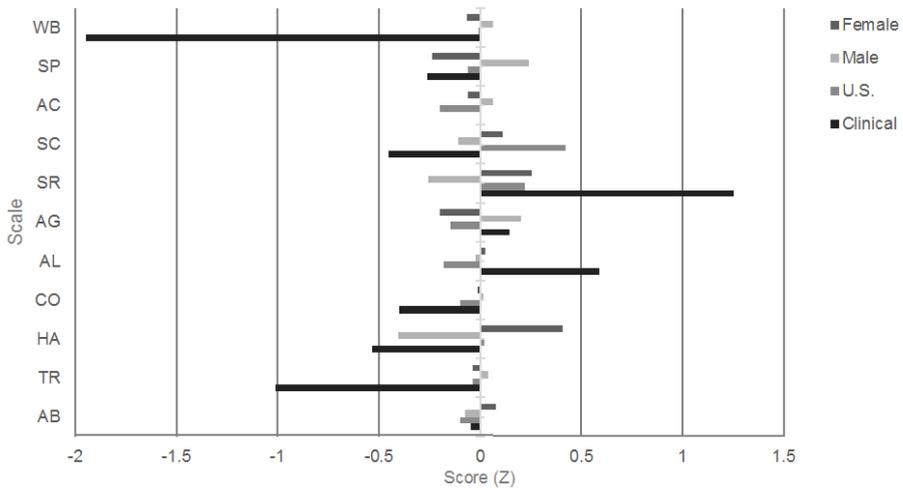

Figure 7.1. Mean deviations from the Dutch representative sample mean ($M = 0$) for males, females, the U.S. general sample and the clinical sample. Means are taken from the (partially) strict invariance models from the single scale measurement invariance analyses. WB = Wellbeing; SP = Social Potency; AC = Achievement; SC = Social Closeness; SR = Stress Reaction; AG = Aggression; AL = Alienation; CO = Control; HA = Harm Avoidance; TR = Traditionalism; AB = Absorption.

The deviations from the Dutch general population mean for the employed U.S. sample and for males and females were less pronounced. When one compares the deviations of gender and those for the U.S. sample it is striking that many of the U.S. deviations have the same direction as the Dutch female deviations, with both the U.S. and female mean scores being lower than the Dutch general mean for Social Potency, Achievement and Aggression. Both U.S. and female mean scores are higher for Social Closeness and Stress Reaction. As speculated already in chapter 3, our results suggest that the U.S. society (or at least Minnesotan society), relative to the Dutch, fosters more community and less agency. The

finding that women are higher on community traits and men higher on agentic traits conforms to the common stereotype. In terms of the gender differences U.S. society may be more feminine than Dutch society. These notions are of course in need for further testing, but show that even societies that are culturally fairly 'close' may still differ in subtle but fundamental ways.

The common structure of general and pathological personality
A growing body of research suggests that the same general dimensions can describe normal and pathological personality. In chapter 4 we added to the evidence of the common structure of general and pathological personality by carrying out a measurement invariance study on a relatively neglected relevant comparison, namely the invariance of a normal personality inventory (i.e. the MPQ-BF-NL) in general and clinical samples. In order for the conjecture of a common structure of normal and pathological personality to hold, one should show that the structure of normal personality also applies in clinical settings. When it does not, personality in clinical populations may be distorted in a sense that it differs qualitatively from normal personality. A conceptualization of personality pathology as maladaptive extreme trait standings is only tenable when personality structure is the same across these populations.

As an omnibus personality test, the MPQ-BF-NL revealed strict invariance across a general and a clinical sample. On a more fine-grained level (when each of the scales was considered in isolation) some, but very limited item bias was present in 7 of the 11 MPQ-BF-NL scales (i.e. Wellbeing, Achievement, Social Closeness, Stress Reaction, Harm Avoidance and Traditionalism). Like in the cross-cultural and gender comparisons, the majority of item bias reflected differences in item difficulty or endorsement probability, meaning that not so much the relevance of the different items for the traits differed, but the trait levels needed for endorsement. Item bias that could not be explained by general issues of the internal structure of the scales (as it could for Achievement, see above), may be explained by differences between people from general and clinical populations regarding the internal standards with which people compare their own emotions, cognitions and behaviors. People from the clinical sample endorsed some Positive Emotionality items more readily than people from the general sample regardless of trait-level, while they endorsed some of the Negative Emotionality items less easily regardless of trait-level. For example, the item *"Often I get irritated by little problems"* was 'easier' for people from the general sample than for people from the clinical sample. Possibly, because people who have personality

problems are confronted with greater challenges and larger setbacks than the general population, or at least they appraise life events in that way, they may have higher thresholds for reporting issues related with these setbacks.

Reflecting the recognition that general and pathological personality can be described by the same dimensions, the alternative (Section III) model for personality disorders of the DSM-5 (American Psychiatric Association, 2013) describes personality pathology by five dimensional constructs that loosely resemble the five-factor models or 'Big Five'. Within each of the traits a number of trait-facets is distinguished and the various categorical DSM-IV PDs can be described by different constellations of specific trait-facet scores. These specific constellations, or profiles, were rationally derived for use in the alternative model for personality disorders. A disadvantage of this rational derivation is that the profiles may not describe how scores actually systematically co-occur. In chapter 5 we show that personality pathology profiles could also be empirically derived. Moreover, we show that their derivation is not only useful for the description of personality pathology, but that their formulation also adds to the evidence of the common structure of general and pathological personality. We distinguished different personality profiles within a general and a clinical sample using multivariate normal mixture modeling. The personality profiles from the clinical sample could be traced back to a subset of the general population that was characterized by extreme scores on the MPQ primary emotional risk markers for psychopathology, Wellbeing and Stress Reaction. The profiles reflected internalizing and externalizing personality styles and had a low base rate in the general population sample. The group of people adhering to these profiles in the general population had markedly more adaptive mean scores on the various MPQ-BF-NL scales than people from the clinical sample: quantitative differences distinguished general from pathological functioning within qualitatively the same profiles.

In sum, we added to the evidence for the common structure of normal and pathological personality by establishing strict invariance across a general and a clinical sample. At the same time our results provided opportunity for speculation about specific pathological influences on personality expression. Our work on the empirical derivation of personality pathology profiles may also point to novel ways one might empirically derive hybrid dimensional-categorical conceptualizations of personality and personality pathology as exemplified in section III of DSM-5.

The realist basis of personality

The model underlying the MPQ departs from a realist trait perspective, which entails that common traits can be good approximations of individual traits that exist independently of their definition in the form of psychobiological structures. For this realist trait assumption to hold, the traits should apply universally: they should be able to describe individual differences occurring in all sorts of groups of people, regardless of their gender and cultural background (among other characteristics). Chapter 3, where we described cross-cultural measurement invariance of the MPQ-BF-NL, provides information about the extent to which the MPQ constructs are cross-culturally valid, or at least whether the constructs as conceptualized in a U.S. context translate to the Dutch context. Also, the measurement invariance analyses presented in the supplementary material of chapter 7 add to our knowledge about the generalizability of the constructs in providing evidence for applicability across gender types. In these studies we interpreted the aspects of the scales that provided bias free measurement as representing components of the constructs that were generalizable across the employed samples. Following the same line of reasoning we interpreted DIF as informative about sample specific expressions.

We showed that the 11 primary traits as operationalized by the MPQ-BF-NL pertain to both sexes and to both U.S. and Dutch citizens (i.e. configural invariance was established). This means that the trait constructs showed validity across the samples employed by us. However, the specific structure of the scales varied somewhat across samples (i.e. strict invariance was not established for all scales). Within both comparisons, most DIF was observed in item difficulties or endorsement probabilities. Loading DIF was much less extensive (even nonexistent in the gender comparison). This means that the relevance of the different items for the traits did typically generalize across samples while the trait levels needed for endorsement did more often differ. In the case of the U.S. Netherlands comparison, the differences in item difficulties could generally be explained by subtle linguistic or cultural differences. For example, the item *"I play hard and I work hard"* was more easily endorsed by people from the U.S. sample than people from the Dutch sample, regardless of level of Achievement. One could speculate that the word 'hard' is stronger in Dutch than in English. Alternatively, cross-cultural differences may exist in the norms about what working hard or playing hard entails. Within the comparison between gender types, the differences in item difficulties could generally be explained by the existence of heterotypic instability: regardless latent trait levels the manifestation of the trait may depend on gender type. For example, the item *"I would not like to try skydiving or bungie-*

jumping" is more readily endorsed by males regardless of their total trait score. It may well be that, because extreme sports are thought to be more appropriate for males than for females, males engage more easily in them without being more impulsive per se. The above examples illustrate how cultural and linguistic factors influence the expression of traits without disqualifying a realist trait perspective. Trait scores could be corrected for cultural and linguistic factors (as we did in chapters 3 and 4) in order to provide bias free measurement of traits that generalize across samples and can therefore be thought to have validity across the respective populations.

Universality of the measured constructs is a necessary but not sufficient condition for the realist assumption to hold. Even when the traits refer to universally shared concepts, they may still refer to social constructions, or folk-psychological concepts that result from shared processes in person perception (Block, 1995; Tellegen, 1991b, 1993). These concepts may well lack a psychobiological basis. Numerous studies have shown structural brain-personality associations, suggesting a direct basis of personality in brain structure. However, results are heterogeneous and inconsistent with single traits reported to be associated with many different brain structures bearing many different functions and specific associations being rarely replicated. In chapter 6 we hypothesized that many of the reported associations may be chance findings. We put this hypothesis to the test by searching for Gray and White Matter associations of the MPQ-BF-NL primary trait scales. Cutting a large Dutch representative young adult sample in an exploratory sample and a cross-validation sample, we obtained associations in the first subsample that we tested in the second subsample. A multitude of associations emerged from our exploratory analyses, but none of these were replicated in the cross-validation sample. Consequently, a robust direct link between the MPQ-BF-NL scores and brain anatomy indices could not be established, suggesting that many of the reported personality-brain associations also are not robust. We still deem it plausible that a great deal of individual differences in personality can be attributed to differences in brain structure and functioning. We speculate that instead of being the direct result of specific brain characteristics and processes personality is better conceptualized as an emergent construct that results from complex interactions these characteristics and processes. This conjecture is of course in need of further testing. Recently developed techniques for dynamic causal modeling could be employed for future investigations into this (Friston et al., 2003).

In short, we showed that the constructs measured by the MPQ-BF-NL are universal to the extent that they generalize across gender types and across U.S. and Dutch populations. In so doing, we were also able to formulate hypotheses about cultural and linguistic influences on expressions of personality. A direct link between personality and brain anatomy was not found, implying that if personality would be rooted in brain functioning, this would derive from a more complex relation between the two.

General conclusion

The MPQ-BF-NL is a broadband personality inventory that provides the Dutch-Flemish language area with an instrument that has favorable characteristics for use in both research and clinical practice. In developing and validating the MPQ-BF-NL, we simultaneously added to the knowledge base of personality by gaining insight into its cross-cultural differences, its psychopathological manifestations, and its neurobiological basis.

Pursuing the practical goal of developing a measurement instrument and at the same time exploring the theoretical basis of personality has been the approach taken by Auke Tellegen from the start of the development of the instrument. Or, in Tellegen's words: *"Scale construction clearly can be more than an ad hoc procedure or a mechanical routine. In its most interesting applications it is a method for comparing ideas of covariation with the facts. Used in that way it is one of the best methods for testing trait constructs, and the best for shaping them"* (Tellegen, 1991b, p. 30).

I hope this thesis lives up to this ambition.