Sex differences in the 16PF5, test of measurement invariance and mean differences in the US standardisation sample

RUNNING HEAD: Gender Differences in the 16PF5

Gender Differences in the 16PF5; Test of Measurement Invariance and Mean Differences in the US Standardisation Sample

Tom Booth & Paul Irwing

Psychometrics at Work Research Group,

Manchester Business School, University of Manchester, UK

(Word Count: 4,964; 20 Pages; 2 Tables; 1 Figure)

Key Words: Personality, Gender Differences, Measurement Invariance, 16PF.

Correspondence concerning this article should be addressed to Tom Booth, Psychometrics at Work Research Group, Manchester Business School East, The University of Manchester, Booth Street West, Manchester, M15 6PB. Electronic mail may be sent via Internet to Thomas.booth@postgrad.mbs.ac.uk

Abstract

Gender differences in personality, though widely commented on, have rarely been investigated either outside of the Five Factor Model, or using the most sophisticated methodologies. The current article looks to address this gap in research by applying Multi-Group Covariance and Means Structural Analysis (MG-CMSA) to the US standardisation sample (n=10,261) of the 16PF5. The results indicated that the assumptions of measurement invariance do not hold for the global scales of the 16PF5. Consequently, mean differences were only investigated in the 15 primary personality scales. Substantial mean differences were found in the scales of Sensitivity (d=2.29) and Warmth (d=.89), with moderate differences located in the scales of Emotional Stability (d=-.53), Dominance (d=-.54), Apprehension (d=.60) and Vigilance (d=-.36). These differences were shown to be systematically larger than estimates of mean differences in the same scales from observed scores.

1.0 Introduction

In recent years, there has been an increase in the number of studies of gender differences in personality. However, few of these studies investigate differences in large omnibus measures of personality, and fewer still adopt the most sophisticated methods of analysis. In the current article, we address this issue by applying multi-group covariance and mean structure analysis (MG-CMSA) to the standardisation sample of the 16 Personality Factor Questionnaire, Version 5 (16PF5).

Feingold's (1994) meta-analysis was one of the first studies to consider gender differences in broad personality characteristics. In the second of two analyses presented by Feingold (1994), the scales from thirteen different personality inventories were categorized according to their relationship with the facet scales from the NEO-PI-R. Males were found to be more assertive than females (Cohen's d-score= .50), whilst females scored more highly on anxiety (-.25), trust (-.28) and tender-mindedness (-.97). Importantly for the current studies, these results suggest a number of gender differences for 16PF scales. The Assertiveness grouping contained the 16PF Dominance (E) scale; Anxiety contained the Emotional Stability (C) scale; and Tender-Mindedness contained the 16PF scale of the same name, which in the 16PF5 is labelled Sensitivity (I).

Two meta-analyses present cross-cultural gender differences based on measures of the five factor model (FFM). Costa, Terracciano and McCrae (2001) analysed responses to the NEO-PI-R from 26 different cultures (N=23,031). In terms of the broad factors of the NEO-PI-R, the authors noted significant gender effects for the scales of Neuroticism (-.51), Agreeableness (-.59) and Extraversion (-.29), with women scoring more highly in all instances. They also compared the facet scale mean differences within the US to the mean facet scale differences in the remaining 25 countries. The authors note that the gender differences in the US are highly comparable in size and significance to the mean differences in the rest of the countries. In the US sample, gender differences were located in all facets of Extraversion and Agreeableness, and all but one of the facets in both Neuroticism and Openness to Experience. However, only Competence from Conscientiousness displayed a significant gender difference. Overall these results were consistent with past findings with women showing higher mean scores on all facets of Neuroticism and Agreeableness, as well as the facet scales of Warmth, Gregariousness and Positive Emotions. Men reported higher mean scores for Assertiveness and Excitement Seeking.

Schmitt, Realo, Voracek and Allik (2008) investigated gender differences in the Big Five Inventory (BFI) in 55 nations (N=17,637). Across all nations, the authors found the largest gender differences to be in the Big Five factors of Neuroticism (-.40), Agreeableness (-.15), Conscientiousness (-.12) and Extraversion (-.10). Specifically within the studies from the USA, Neuroticism had the largest gender difference (-.53), followed by Openness (.22), Conscientiousness (-.20), Agreeableness (-.19) and lastly Extraversion (-.15).

The meta-analytic research on gender differences in personality has a number of drawbacks. Firstly, they are primarily limited to measures of the Five Factor Model (FFM). Secondly, they assume that the structures of the focal personality tests are accurate. Yet there is evidence to suggest that this is not the case (Church & Burke, 1994; Aluja, Garcia, Garcia & Seisdedos, 2005). Thirdly, a majority of the studies use computed scores on facets or factors and compare the differences in these observed scores. This process, synonymous with classical test theory, logically implies that the observed score represents an *actual* score on a construct (personality trait), and thus treats the observed score, true score and construct score as equivalent. This is an inappropriate assumption. Personality constructs are accurately theorised as latent variables, which are not equivalent to either true scores or construct scores (*see* Borsboom & Mellenbergh, 2002).

Our focus in the current study is the 16PF5. The technical manual for the 16PF5 (Conn & Rieke, 1994) lists gender differences with medium effect sizes in the primary scales of Warmth, Sensitivity and Dominance. The mean score differences for these scales were -.83, -1.89 and .72 respectively, with females reporting higher mean scores on Warmth and Sensitivity, while males reported higher scores on Dominance.

Outside of the technical manual, we located only one study which considered gender differences in the 16PF5. In a study of male and female clergymen, Musson (2001), found men scored higher on the primary scales of Warmth, Rule-Consciousness, Sensitivity and Apprehension. These findings are almost entirely in the opposite direction to what would be expected from past research and theory. It therefore seems sensible to conclude that these results are highly sample specific.

Though the use of observed scores is common place in studies of group differences, methodological advances in recent years offer a far more comprehensive suite of analyses for investigating group differences. Collectively known as multi-group covariance and mean structures analysis (MG-CMSA), these methods adopt structural equation modelling to first test for the equivalence of the covariance structure within a given measure, and then use this robust structure to compare latent mean differences in the constructs of interest.

Measurement invariance tests the assumption that the construct being measured is the same in both groups. If invariance does not hold, then decisions based on group differences may be inaccurate (French & Finch, 2006). If invariance does hold, then precise estimates of group mean differences can be made.

Invariance can be assessed at multiple levels. Most commonly, the pattern of factor loadings (configural), degree of factor loadings (metric) and the intercepts of indicators (scalar) are assessed for invariance (Widaman & Reise, 1997). In second order factor models, configural and metric invariance may also be estimated. If the conditions of scalar invariance are met, then it is possible to compare between group means within latent factors.

Despite suggestions that tests of invariance are an important step forward in both personality research and assessment of group differences (Finch & West, 1997), there are few published examples of studies which consider gender differences in personality within this framework. Three studies have applied measurement invariance analyses to investigate gender differences in the Five Factor Model. However, once again none of these studies utilise omnibus measures of personality. Gomez (2006) investigated gender differences in a sample of adolescents, using a 25 item abridged Big Five measure developed by Scholte, van Aken and van Lieshout (1997). Gomez (2006) reported that the assumptions of metric invariance were violated, but goes on to conclude that the mean scores in the broad five factors are invariant across gender. Given the lack of metric invariance, these results must be interpreted with caution.

Gustavsson, Eriksson, Hilding, Gunnarsson and Ostensson (2008) provided an invariance and mean difference analysis of the 20 item Health Relevant Personality 5 questionnaire (HP5) in a sample of 5,700 individuals from the Stockholm Diabetes Prevention Programme. The results indicated that assumptions of configural, metric and scalar invariance were met in the HP5 across gender. Further, the authors found mean differences on all five factors, with women scoring more highly on Impulsivity, Hedonic Capacity and Negative Affectivity, whilst men scored more highly on Antagonism and Alexithymia.

Finally, Ehrhart, Roesch, Ehrhart and Kilian (2008) conducted similar analyses using the 50-item International Personality Item Pool (IPIP) measure of the Big Five in a sample of 1,727 college students. The findings of this study were consistent with the meta-analyses of the FFM. Women were found to be more agreeable (-.36) and conscientious (-.13), but less open (.36) and emotionally stable (.11). Our review of the literature indicated that there are no papers which explore gender differences in the 16PF using MG-CMSA.

The degree to which men and women differ in levels of personality traits and the substantive importance of these differences has been questioned. Hyde (2005) has argued in favour of the gender similarity hypothesis. Put briefly, this position suggests that the differences between men and women are not that large, with most d-score effect sizes falling in the small to moderate range. Further, Hyde (2005) argues for the dangers of over-emphasising the differences between males and females and the negative consequences which may arise as a result.

The current study contributes to this debate in two ways. Firstly, we provide a comparison of mean differences and effect sizes estimated using MG-CMSA and observed scale scores. It is suggested that MG-CMSA will provide more robust and accurate estimates of gender differences, and thus the results will offer important insight into the true magnitude of gender differences in personality traits. Secondly, this study represents the only example of MG-CMSA being applied to a full omnibus measure of personality, and also provides the most rigorous assessment of gender differences in the 16PF5.

2.0 Methodology

2.1 Participants

The current study used the American standardisation sample of the 16PF, 5th Edition (N= 10,261). The sample is structured to be representative of the general population of the USA with respect to a number of demographic variables. The sample is approximately equal across gender, with 50.1% (N=5,137) female and 49.9% (N=5124) male, and consists primarily of white (77.9%; N=7994) and black (10.8%; N=1113) respondents. The majority of respondents (77.9%; N=7996) are below the age of 44. The sample is proportionally geographically distributed and on average, the educational level and years in education of the sample is greater than that of the US population. In all analyses, the sample was split into a male (n=5124) and a female (n=5137) sample.

2.2 Measures

The 16PF 5th Edition (16PF5) contains 185 items organised into 16 primary factor scales containing between 10 and 15 items each. The response format for each of the items consists of a choice from three; "No", "?" and "Yes". The responses are scored as 0, 1 and 2 respectively. The "?" response is intended to provide *"... a uniform response choice that can cover several different reasons for not selecting either ... alternative" (Conn & Rieke, 1994, p.8).* The 16PF5 contains 15 primary personality scales, a 15 item Reasoning scale, and a 12 item Impression Management Scale. The current analysis utilises only the 15 personality scales, which are further organised into 5 global scales; namely Extraversion (Warmth, Liveliness, Social Boldness, Privateness & Self-Reliance), Anxiety (Emotional Stability, Vigilance, Apprehension & Tension), Tough-Mindedness (Warmth, Sensitivity, Abstractedness & Openness to Change), Independence (Dominance, Social Boldness, Vigilance & Openness to Change) and Self-Control (Liveliness, Rule-Consciousness, Practical & Perfectionism).

2.3 Analysis

As an initial step, item parcels were created using the Single Factor method (Landis, Beal & Tesluk, 2000). Item loadings from single factor confirmatory factor analyses of each of the 15 primary personality scales of the 16PF5 were used to create three item parcels per primary scale. The use of three parcels ensured model identification (Bollen, 1989).

Prior to estimating mean differences, measurement invariance was investigated. All invariance and mean structure models were estimated using robust maximum likelihood (RML) in LISREL 8.72. RML was preferred over DWLS as simulation studies have shown that within invariance analyses, the chance of Type I error increased with sample size when using DWLS (French & Finch, 2006).

Following the suggestions of Widaman and Reise (1997), we estimated a series of models to assess the degree of measurement invariance across gender in the 16PF. Measurement invariance was first established in the first order measurement model, before constraints were placed on the second order factor model. In models M1-M3, configural, metric and scalar invariance were tested in the primary scale structure. If the assumptions of scalar invariance were met, then it could be assumed that all differences between the groups were accounted for by differences in the first order latent variables. Next, we tested for configural (Model S1) and metric (Model S2) invariance within the second-order, global structure of the 16PF5.

To establish the overall fit of each of our models, we rely primarily on the simulations of Hu and Bentler (1998, 1999). We adopted cut-off points of .05 for the SRMSR, about .06 for the RMSEA, and ≥ .95 for the NNFI and CFI, which conform to recent recommendations based on Monte Carlo simulation (Hu & Bentler, 1998, 1999) and the review of Schemelleh-Engel, Moosbrugger, and Muller (2003). However, within invariance analysis, the difference in fit is of greater importance than the absolute values. Decline in model fit at a given stage of the analysis indicates that the assumptions of invariance do not hold in the constrained parameters (French & Finch, 2006). To assess possible decline in model fit, we rely on the conclusion of Cheung and Rensvold (2002), who suggest changes of equal to or less than -0.01 for CFI indicate that invariance holds. Further, we suggest comparable cut-off values of 0.013 for the RMSEA and -0.008 for the NNFI, based on the findings of Cheung and Rensvold (2002).

Once measurement invariance has been established, it is possible to estimate mean differences between groups in both the global and primary scales. Conventionally, the significance of mean differences can be estimated by placing invariance constraints on each mean individually, and noting the change in model fit (Fan & Sivo, 2009). In the current study, our primary focus was on the effect sizes of mean differences. Therefore, mean scores between the two groups were compared using Cohen d-scores (Cohen, 1988). Cohen d-scores were converted into an r^2 statistics, in order to estimate the amount of variance in scores explained by group membership. Given the large sample size utilised in the current study, the power of d-score estimates is high (Cohen, 1988).

2.4 Missing Data

An evaluation using PRELIS in LISREL 8.72 indicated that .075% of the data were missing. Once again, there is little consensus as to the most appropriate methods of dealing with missing data. In this instance, we used Schafer's NORM package to compute single imputations at the parcel level.

3.0 Results

The fit statistics for the invariance analysis are presented in Table 1. The baseline configurally invariant model, M1, shows excellent fit to the data. Similarly, both the absolute model fit and the difference in fit statistics for model M2 indicate that the assumptions of metric invariance hold. In model M3a, the ΔNNFI (-.02) is large in combination with the ΔCFI (-.01), indicating that the assumptions of scalar invariance do not hold across groups. Examination of the modification indices suggested that releasing the invariance constraints on parcel 1 of the Warmth primary scale (MI=1588.45) and parcel 2 of the Emotional Stability primary scale (MI=1497.34), would improve model fit. These two constraints were sequentially lifted, resulting in satisfactory model fit (Model M3b Table 1). These results suggest that with the exception of the two item parcels noted above, measurement invariance holds in the primary scales of the 16PF5.

(Insert Table 1 About here)

Using model M3b as a baseline, we next tested configural invariance in the second order factors. The fit for model S1 suggested that the assumptions of configural invariance were violated in the second order factors. The absolute fit values for the SRMR are high in the male (.076) and female (.072) groups. The ΔNNFI (-.02) in combination with the ΔCFI (-.01) and the proportionally large increase in SBχ^2 all support the conclusion that a degree of misspecification is present in the structure of the global factors of the 16PF5.

Given that it was not possible to clearly demonstrate invariance in the global factors of the 16PF5, mean differences were only examined within the 15 primary personality scales (Figure 1).

(Insert Figure 1 About Here)

The first three columns of Table 2 detail the mean difference, Cohen's d-score and r^2 value for each of the primary scales estimated using MG-CMSA. In the MG-CMSA, the means of the male group were fixed to zero, whilst the means of the female group were freely estimated. Therefore, the mean values represent the degree to which the female group means deviate from the male group means. Positive values indicate that the male mean was higher.

(Insert Table 2 About Here)

Non-significant mean differences were found for Abstractness (t = 0.24) and Perfectionism (t = -1.93). The primary scale Liveliness (t = 2.20) was significant at .05 level, whilst all other primary scales showed significant mean differences between gender at the .001 level.

The gender differences for Sensitivity and Warmth show large effect sizes according to the d-score criteria, with women showing higher mean scores on both scales. Consideration of the r^2 value suggests that 56.8% of the variance in Sensitivity mean scores is attributable to sex. In the case of Warmth, women once again scored significantly higher than men with 16.4% of mean score variance attributable to sex. A moderate gender difference, with women scoring more highly, was also observed in Apprehension (d=.60). A number of moderate gender differences are observable in the opposing direction. Men have higher mean scores on the primary scales of Emotional Stability (d=-.53), Dominance (d=-54), Rule-Consciousness (d=-.39) and Vigilance (d=-36). The variances in these mean scores attributable to gender range from 3.2% to 6.8%.

The last three columns of table 2 provide the mean difference, Cohen d-scores and r^2 values calculated from observed scores of the primary scales of the 16PF. As can be seen in table 2, with the exception of Abstractness, the mean difference, d-score and r^2 calculated from the observed scores are systematically lower than the estimates from the MG-CMSA analysis. This leads to important differences in interpretation of d-scores. For example, taking Cohen's criteria for significance, Warmth has a large effect size (.887) when calculated using MG-CMSA, but only a moderate effect size (.500) from observed scores. Similarly, Vigilance shows a small to moderate effect size (-.363) within the MG-CMSA analysis, but a trivial difference (.025) is indicated by the observed scores.

4.0 Discussion

The results of the current study offer some confirmation of prior results, but also a number of extensions. Firstly, the results support prior findings of gender differences in the 16PF5. Conn and Rieke (1994) report differences in the primary scales of Warmth, Sensitivity and Dominance. The current findings support these conclusions. In the current analysis, gender differences with moderate effect sizes also exist on the primary scales of Emotional Stability, Rule-Consciousness, Vigilance and Apprehension.

Feingold (1994) found the largest gender differences were related to the Tender-Mindedness facet scale of the NEO-PI-R. The 16PF scale of Sensitivity was one of the collection of primary scales Feingold grouped under Tender-Mindedness. Our results support this conclusion, as Sensitivity clearly demonstrated the largest difference between genders.

The results of Feingold's study also suggested that gender differences are to be found in the 16PF scales of Dominance and Emotional Stability. Once again these results were supported in the current study.

Across the meta-analytic studies which have investigated gender differences in the FFM (Costa et al. 2001; Schmitt et al 2008), Neuroticism has been consistently highlighted as demonstrating the largest gender differences. The 16PF5 equivalent global scale of Anxiety, encompasses the primary scales of Emotional Stability, Vigilance, Apprehension and Tension. Our findings indicated that each of these scales showed a significant gender difference, with moderate effect sizes for Emotional Stability, Vigilance, Apprehension and a small to moderate effect size for Tension.

The largest differences found in the current study related to Warmth and Sensitivity. These primary scales belong to the global scale of Tough-Mindedness, which is most strongly correlated with the NEO-PI-R factor of Openness to Experience (-.56). Only in the analysis of Schmitt et al (2008) was Openness identified as a factor with large gender differences.

Though the pattern of gender differences identified in the current study is broadly consistent with prior findings, there are two notable differences. Firstly, mean differences, and subsequent effect sizes, were systematically larger when estimates were made from the MG-CMSA models rather than observed scores. The MG-CMSA analysis also produced larger gender differences than have been reported in previous research. The MG-CMSA model is distinct from observed score procedures as it requires the fitting of a measurement invariant latent factor model. Our results suggest that ignoring these issues in the investigation of gender differences leads to under-estimation of true mean differences.

Given the possibility of systematic under-estimation, how should one consider Hyde's (2005) gender similarity hypothesis? Clearly our results suggest that on some narrow primary scales of the 16PF5, very large differences between men and women can be observed.

Further, given the large sample size, we can also be sure that differences with small to moderate effect sizes are substantive. Though we agree with Hyde's conclusions that over emphasising gender differences may have negative consequences, the results of the current study suggest that the opposite may also be true. That is, men and women do fundamentally differ to a large extent on specific characteristics of personality. To ignore such differences in favour of a gender similarity hypothesis may also have a negative impact on individuals.

In the current study, we were only able to reliably estimate mean differences in the primary scales of the 16PF5 as the assumptions of measurement invariance were violated in the global scales. Of the studies reviewed in the production of this paper, only the study of Costa et al (2001) considered gender differences in narrow scales. Our results suggest this may be a critical step in the analysis of gender differences for a number of reasons.

Firstly, the broad scales of the 16PF were not invariant across groups. Prior studies which show the poor fit of FFM inventories in confirmatory factor analyses (e.g. Church & Burke, 1994) suggest a similar situation may exist in other measures. As such, assessment of gender differences in these broad scales may be unreliable as it is not clear that they measure the same construct across groups. Secondly, specific information on the components of broad factors is lost when only a broad factor observed score is used. For example, in the current analysis, Warmth demonstrated a large effect size with women scoring more highly than men. Warmth is one component of the 16PF global scale of Extraversion. However, all other primary scales within the Extraversion factor (Liveliness, Social Boldness, Privateness & Self-Reliance) show small to neglible gender differences. If gender differences had been investigated in the observed score of the Extraversion broad factor, it is likely inappropriate conclusions would have been drawn as to the presence and source of gender differences.

Overall the results of the current study indicate the utility of MG-CMSA for investigating gender, and more broadly, group differences in psychological constructs. The results suggest that the use of observed scores may lead to under-estimates of true differences. Further, the results provide further evidence that the conditions of measurement invariance are not supported in some personality instruments. As such, reliable calculations of group differences are not possible. Finally, the results suggest the utility of investigating group differences in narrow primary or facet scales of personality measures. Failure to do so may lead to the loss of specific information on large gender differences on these narrow scales.

5.0 References

Aluja, A., Garcia, O., Garcia, L.F., & Seisdedos, N. (2005). Invariance of the NEO-PI-R factor structure across exploratory and confirmatory factor analyses. *Personality and Individual Differences*, *38,* 1879-1889.

Bollen, K.A. (1989). *Structural equations with latent variables*. New York: Wiley.

Borsboom, D., & Mellenbergh, G.J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence, 30,* 505-514.

Church, A.T., & Burke, P.J., (1994). Exploratory and Confirmatory Tests of the Big 5 and Tellegan's 3-Dimensional and 4-Dimensional Models. *Journal of Personality and Social Psychology, 66,* 93-114.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 235-255.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences. 2nd ed*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Conn, S. R., & Rieke, M. L. (1994). (Eds.) *The 16PF fifth edition technical manual.* Champagne, IL: Institute for Personality and Ability Testing, Inc.

Costa, P. T., Terraciano, A., & McCrae, R. R. (2001). “Gender differences in personality across culture: Robust and surprising findings.” *Journal of Personality and Social Psychology*, 81, pp.322-331.

Ehrhart, K.H., Roesch, S.C., Ehrhart, M.G., & Kilian, B. (2008). A Test of the Factor Equivalence of the 50-Item IPIP Five-Factor Model Measure Across Gender and Ethnic Group. *Journal of Personality Assessment, 90,* 507-516.

Fan, X. & Sivo, S.A. (2009). Using ΔGoodness-of-fit Indexes in Assessing Mean Structure Invariance, *Structural Equation Modeling, 16,* 54-69.

Feingold, A. (1994). Gender Differences in Personality: A Meta-Analysis. *Psychological Bulletin, 116,* 429-456.

Finch, J. F., & West, S.G. (1997). The Investigation of Personality Structure: Statistical Models. *Journal of Research in Personality, 31,* 439-485.

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural equation Modeling*, *13*, 378-402.

Gomez, R. (2006). Gender invariance of the five-factor model of personality among adolescents: A mean and covariance structure analysis approach. *Personality and Individual Differences, 41,* 755-765.

Gustavsson, J.P., Eriksson, A., Hilding, A., Gunnarsson, M., & Ostensson, C. (2008). Measurement invariance of personality traits from a five-factor model perspective: multi-group confirmatory factor analyses of the HP5 inventory. *Scandinavian Journal of Psychology, 49,* 459-467.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structural modelling: Sensitivity to underparametrized model misspecification. *Psychological Methods*, *3*, 424-453.

Hu, L. T., & Bentler, P. M. (1999). Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Hyde, J.S. (2005). The Gender Similarity Hypothesis. *American Psychologist, 60,* 581-592.

Landis, R. S., Beal, D. J., & Tesluk, P. E. (2000). A comparison of approaches to forming composite measures in structural equation models. *Organizational Behavioral Research*, *3,* 186–207.

Musson, D.J. (2001). Personality of Male Anglican Clergy in England: Revisited using the 16PF5. *Mental Health, Religion and Culture, 5,* 195-206.

Schermelleh-Engel, K., Moosbrugger, H., & Muller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8,* 23-74.

Schmitt, D.P., Realo, A., Voracek, M., & Allik, J. (2008). Why Can't a Man Be More Like a Woman? Sex Differences in Big Five Personality Traits Across 55 Cultures. *Journal of Personality and Social Psychology, 94,* 168-182.

Scholte, R.H.J., van Aken, A.G., & van Lieshout, C.F.M. (1997). Adolescent personality factors in self-ratings and peer nominations and their prediction of peer acceptance and peer rejection. *Journal of Personality Assessment, 69,* 534-554.

Widaman, K. F., & Reise, K. F. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, & M. Windle (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-324). Washington, DC: American Psychological Association.

Footnote

[1] All values in parentheses are Cohen’s d-scores. Cohen (1988) argues that d-scores of around .20 represent small effects, .50 moderate effects and .80 large effects. In keeping with convention, positive values indicate a male advantage, whilst negative values indicate a female advantage.

Table 1

Model Fit Statistics for Invariance Analysis of Male and Female Samples from the US Standardisation Sample of the 16PF5

Model	χ^2_{SB}	df	RMSEA	SRMR	CFI	NNFI
M1: First Order Configural Invariance	14636.52	1680	.039	.047 .045	.97	.97
M2: First Order Metric Invariance	16646.22	1710	.041	.055 .050	.97	.97
Δ Fit Statistics M2 vs M1	*5200.10* *(p<.0001)*	*30*	*.002*	*.008* *.005*	*0*	*0*
M3a: First Order Scalar Invariance	24097.21	1740	.050	.067 .053	.96	.95
Δ Fit Statistics M3a vs M2	*-4879.84*	*30*	*.009*	*.012* *.003*	*-.01*	*-.02*
M3b: First Order Scalar Tau-X 1 & 5 Released	20807.87	1738	.046	.059 .051	.96	.96
Δ Fit Statistics M3b vs M2	*-6413.69*	*28*	*.005*	*.009* *.001*	*-.01*	*-.01*
S1: First Order Scalar and Second Order Configural Invariance	29572.58	1886	.053	.076 .072	.95	.94
Δ Fit Statistics S1 vs M3b	*3252.43* *(p<.0001)*	*148*	*.007*	*.017* *.021*	*-.01*	*-.02*

Table 2

Comparison of mean differences, Cohen's d-score and r^2 between MG-CMSA and observed scores for the primary scales of the 16PF5.

	MG-CMSA			Observed Scores		
	ΔMean	*d*-score	r^2	ΔMean	*d*-score	r^2
A: Warmth	2.61	.887	.164	1.85	.500	.059
C: Emotional Stability	-2.91	-.525	.065	-1.29	-.316	.024
E: Dominance	-1.65	-.539	.068	-0.98	-.272	.018
F: Liveliness	.17	.049	.000	0.10	.025	.000
G: Rule-Consciousness	-1.37	-.385	.036	-0.40	-.108	.003
H: Social Boldness	-1.00	-.180	.008	-0.27	-.055	.001
I: Sensitivity	5.54	2.293	.568	5.02	1.342	.310
L: Vigilance	-1.46	-.363	.032	0.10	.025	.000
M: Abstractness	.02	.005	.000	0.19	.044	.001
N: Privateness	-1.29	-.146	.005	-0.87	-.195	.009
O: Apprehension	3.38	.603	.083	2.7	.587	.079
Q1: Openness to Change	-.84	-.212	.011	0.13	.036	.000
Q2: Self-Reliance	.86	.117	.003	0.26	.058	.001
Q3: Perfectionism	-.24	-.043	.001	-0.10	-.025	.000
Q4: Tension	1.65	.267	.018	1.07	.240	.014

Figure Caption