LUNDS
UNIVERSITET

Department of Psychology

Does the Gender-Equality-Personality-Paradox Depend on the Scoring Method?
The Implications of Measurement Invariance

Julie Sophie von Westerman

Master's Thesis

2024

Supervisor: Petri Kajonius

Abstract

With increasing country-level gender equality, country-level gender differences in personality are found to increase. This is known as the Gender-Equality-Personality-Paradox (GEPP). Possibly, the GEPP is a methodological artefact. So far, the personality trait scoring method has not been investigated as a potential source of the GEPP. Three kinds of scores were investigated in the present study: Sum-Scores add all questionnaire-items into a total score; Factor-Scores are extracted from a global personality model; and measurement invariance (MI) Factor-Scores are extracted from MI-models. MI holds if traits are measured equally across groups (i.e., if groups are comparable). If MI does not hold, MI-Factor-Scores can also be extracted from partial MI-models accounting for non-comparability between groups. The present study investigated the impact of these scoring methods on GEPP-estimates: For each scoring method, country-level gender differences in single personality traits and in overall personality were computed and then correlated with country-level gender equality. All three scoring methods suggest that the GEPP is small to medium and insignificant for most traits but large and significant for Extraversion. Differences between scoring methods were subtle and did not change inferences regarding the GEPP. However, when ranking countries based on the size of gender differences, scoring methods changed the rank positions of some countries (e.g., China). In sum, this study suggests that the GEPP is not a methodological artefact as far as scoring methods are concerned. Yet, researchers should be mindful of the potential effect of scoring methods on countries' rank order regarding gender differences.

Keywords: Five Factor Personality Model, Measurement Invariance, Test Scores, Gender Differences, Gender Equality

Acknowledgements

This master's thesis would not have been possible without the support of others. I could not have undertaken this endeavour without the helpful feedback and guidance from my supervisor, Petri Kajonius. I would also like to express my deepest gratitude to Martin Bäckström and Anton Andersson for introducing me to measurement invariance testing. Special thanks to John Johnson for making a large, cross-national personality dataset publicly available without which this master's thesis could not have been written. Finally, I thank my family and friends for their emotional support throughout the entire process of thesis writing.

Introduction

Objective

As living conditions for men and women converge, differences between men and women in personality become larger (Balducci, 2023; Herlitz et al., 2024). This phenomenon is known as the Gender-Equality-Personality-Paradox (GEPP; Fors Conolly et al., 2020). Typically, the GEPP is assessed cross-sectionally by correlating country-level gender differences in personality with country-level gender equality measures (Balducci, 2023).

Understanding how gender[1] differences develop with increasing gender equality can be crucial, especially for policy makers: If higher levels of gender equality truly increase gender differences, then this might imply that means to improve gender equality are reinstating gendered social roles. To properly evaluate gender-equality-policies, it is important to understand the mechanisms driving the GEPP.

However, there is no consensus regarding the origins of the GEPP. Some researchers view this phenomenon as an artefact of the analytical approach chosen (e.g., depending on the choice of gender equality index [Marsh et al., 2021] or on the investigated countries or time points [Guo et al., 2024]). No prior research has considered the scoring method used for calculating personality trait scores as a potential source of the GEPP. Broadly, three types of scores can be distinguished (McNeish & Wolf, 2020): (a) Sum-Scores, (b) Factor-Scores, and (c) Factor-Scores derived from measurement invariance (MI) models. Prior research suggests that the size of gender differences varies with the scoring method (Del Giudice et al., 2012; Kaiser et al., 2020). This might have consequences for the GEPP.

The present study explored the impact of these three scoring methods on gender-difference- and GEPP-estimates. By that, this master thesis hopes to contribute to a better understanding of the GEPP, which could ultimately have consequences for future gender policies. Moreover, comparing different scoring methods can help clarify whether choice of scoring method is an arbitrary decision or whether it changes statistical inferences. This will

[1] Despite the plurality of gender identities, I focused on personality differences between only men and women for pragmatic reasons: The dataset on which this study builds coded gender dichotomously. Hence, by the term gender, I refer to male and female gender throughout this study.

not only benefit research in personality psychology, but in all psychological disciplines concerned with measuring and scoring latent constructs.

Personality

The Five Factor Model of Personality

The most common taxonomy of personality traits is the Five Factor Model (FFM; McCrae & Costa, 2008; Mõttus et al., 2020). According to the FFM, personality consists of five broad traits (or factors) also known as the Big Five (McCrae & Costa, 2008; see Table 1). This taxonomy originates from the lexical hypothesis (i.e., the assumption that personality traits are expressed in natural language; Goldberg, 1993; McCrae & Costa, 2008). Grouping personality-describing words via factor analysis into broader factors revealed the underlying personality structure: the FFM (Ashton & Lee, 2005; Goldberg, 1993).

Gender Differences in Personality

Women tend to score higher on all traits than men (Kajonius & Johnson, 2018). This trend is most pronounced for the traits Agreeableness and Neuroticism (Costa et al., 2001; Kajonius & Johnson, 2018; McCrae & Costa, 2008): Women tend to describe themselves, for

Table 1

The Traits of the Five-Factor-Model with Example Facets and Items

Trait	Example Facet	Example Item
Openness	Intellect	"Love to rad challenging material"
	Liberalism	"Tend to vote for liberal political candidates"
Conscientiousness	Self-discipline	"Am always prepared"
	Dutifulness	"Keep my promises"
Extraversion	Excitement-seeking	"Love excitement"
	Cheerfulness	"Radiate joy"
Agreeableness	Altruism	"Love to help others"
	Sympathy	"Sympathise with the homeless"
Neuroticism	Immoderation	"Go on binges"
	Vulnerability	"Panic easily"

Note. Example facets and items taken from the IPIP-NEO-120 (Johnson, 2014). In this instrument, each trait is measured by 24 items grouped into six facets.

example, as more trusting and cooperative and as more anxious and emotionally instable than men. The size of gender differences depends on the level of analysis: In single traits, gender differences are small (Hyde, 2014; Murphy et al., 2021); in overall personality (i.e., across all traits simultaneously), gender differences are large (Del Giudice, 2013; Mac Giolla & Kajonius, 2018).

However, men's and women's traits are often not sufficiently comparable to calculate their difference (Dong & Dumas, 2020). In that case, a direct comparison of trait scores would not provide valid inferences about gender differences in that trait. To account for (non-) comparability, some researchers suggest calculating gender differences from trait scores that were derived from MI-models (Del Giudice et al., 2012; Kaiser et al., 2020).

Personality Across Countries

To assess the GEPP by comparing gender differences across countries, a personality model is needed that describes personality validly in each country. The FFM is well supported across countries (Allik et al., 2013) although it captures personality better in Western than non-Western societies (Fedvadjiev et al., 2015). Especially Conscientiousness, Extraversion, and Agreeableness replicate well across countries while Openness is more difficult to confirm cross-culturally (Fedvadjiev et al., 2015).

Even though the structure of the FFM is adequate to describe personality globally, trait means are typically not comparable across countries (Dong & Dumas, 2020). In other words, despite a similar structure of personality, traits have different meanings in different countries. This hampers conclusions about cross-country variation in gender differences. Hence, a proper investigation of the GEPP calls for a clear understanding of comparability across countries and genders and its effect on trait estimates.

Score Comparability: Measurement Invariance

Comparability of personality across gender and across countries holds when a given trait is assessed similarly in different groups. This is known as MI (Dong & Dumas, 2020; Putnick & Bornstein, 2016). The necessity for MI testing arises from personality traits being latent constructs that cannot be measured directly, but manifest in measurable behaviours, patterns of thought, or feelings (Robitzsch & Lüdtke, 2023). However, these manifestations might relate differently to the latent construct in different groups (i.e., the construct's meaning might differ across groups; Putnick & Bornstein, 2016).

MI is most frequently examined within a multigroup confirmatory factor analysis (CFA) framework (Dong & Dumas, 2020). Based on a configural model (i.e., a personality structure with adequate model fit in all groups), a sequence of models is compared whose parameters are increasingly constrained to be equal across groups (Dong & Dumas, 2020; Putnick & Bornstein, 2016). The first model to be tested is metric MI by constraining the factor loadings of the personality model to be equal. If the model fit of the metric model as compared to the configural model does not decrease substantially, metric MI across groups can be assumed. This indicates that the strength of association between manifestations and latent construct is the same in all groups. Nevertheless, this level of MI is yet not enough to compare mean trait scores. In addition to the loadings, items' intercepts need to be equal across groups too (Dong & Dumas, 2020; Putnick & Bornstein, 2016). This level is called scalar MI and can be assumed when the scalar model does not fit the data substantially worse than the metric model. Equal item intercepts mean that two groups with the same true latent trait score of zero also have the same baseline manifestation score (Putnick & Bornstein, 2016; Steinmetz, 2013).

In GEPP-research, (a) scalar MI across gender, and (b) metric MI across countries is needed. Scalar MI indicates that the trait scores of men and women are comparable so that meaningful gender differences can be computed. To compare these differences across countries, metric MI across countries is sufficient indicating that a given trait means the same across groups. Traits' intercepts do not need to be equal across countries: A given difference between men and women is the same in each country even if men and women score systematically higher in one country than in the other.

If full MI does not hold at the required level, partial MI can be tested (Putnick & Bornstein, 2016). To that end, invariant parameters (i.e., loadings or intercepts) are identified and freed from equality constraints (Lai et al., 2022). This still ensures invariant (i.e., comparable) latent constructs (Jung & Yoon, 2016; Steinmetz, 2013), while also acknowledging that some parameters need to be estimated differently in each group.

Scoring Methods

Why to Score

All scoring methods are data reduction techniques: On the one hand, the number of questionnaire items (which measure trait manifestations) is reduced. On the other hand, factor indeterminacy is reduced to a point estimate (Lai & Tse, 2024). Factor indeterminacy means that there is an infinite number of solutions to score calculation that all match a specified model (Grice, 2001; Lechner et al., 2021): Even though there is a model with a definite structure, this model could have emerged from an infinite combination of individual scores. Different extraction methods exist to reduce these infinite solutions to a point estimate; these yield different latent trait scores (Lechner et al., 2021). Consequently, trait scores (no matter the extraction method) introduce uncertainty as they are approximations of the latent trait (Rigdon, 2019).

To investigate relationships between latent variables (as is the case when investigating the GEPP), it is not necessary to extract trait scores at all. Instead, the relationship of interest can be directly modelled from the items in a structural equation model (SEM; Grice, 2001; Lai & Tse, 2024). Nevertheless, a SEM approach might not always be feasible from an applied researcher's perspective. Due to size and complexity of SEMs, these models can fail to converge, especially for small sample sizes (Lai & Tse, 2024). Moreover, some research questions might specifically call for point estimates: for example, assessing participants' position along a measurement scale for diagnostic purposes. Another reason could be research conventions in a field: The GEPP, for example, is most frequently assessed using a correlational approach with gender differences being calculated with point estimates of personality traits (Balducci, 2023; Herlitz et al., 2024). If an applied researcher strives for replication or better comparability with prior studies, following established conventions by using point estimates as trait scores can be an appropriate choice.

How to Score

Broadly, two approaches to summarizing items in trait scores can be distinguished: observed scores (or Sum-Scores) and Factor-Scores (McNeish & Wolf, 2020). Sum-Scores are calculated by adding up all items into an overall trait score. The underlying assumption is that each item is equally informative of the latent trait. In contrast, Factor-Scores are estimated from

CFA models of personality. Since items can have different factor loadings in a CFA (i.e., they can differ in the strength of their relationship with the latent trait), Factor-Scores account for the fact that items are not necessarily equally informative of the latent trait (McNeish & Wolf, 2020). Factor-Scores are computed from an ordinary CFA fitted to all individuals in a sample. If the sample comprises several groups (e.g., individuals from different countries or genders), this procedure ignores whether MI holds across these groups or not. This framework can be extended to base Factor-Scores on MI testing: Instead of estimating scores from a global CFA model, MI testing can be conducted, and scores can be estimated from a personality model at the required level of MI. However, when using MI-Factor-Scores, researchers should be aware that MI is specified for the latent trait and does not necessarily apply to the MI-Factor-Scores as they are only estimates of latent traits (Lai & Tse, 2024). This means that MI-Factor-Scores are not necessarily less biased than Sum-Scores or Factor-Scores.

The question which of these scoring methods is most appropriate when latent traits are fully or partially invariant is an ongoing debate (e.g., Lai & Tse, 2024; McNeish, 2023; Widaman & Revelle, 2024). Some researchers advocate (MI-)Factor-Scores claiming them to be more accurate especially under partial MI (e.g., McNeish, 2023; McNeish & Wolf, 2020; Steinmetz, 2013). Others caution against the use of (MI-)Factor-Scores as they can be biased just as observed scores (e.g., Lai & Tse, 2024), favour Sum-Scores (Widaman & Revelle, 2024) or question the necessity of MI testing (and hence deriving scores from such models) altogether (e.g., Robitzsch & Lüdtke, 2023).

The present study does not add to this debate from a theoretical or technical point of view, but from the perspective of applied researchers. Viewed from this perspective, the debate about different scoring methods condenses to the question whether theoretical differences between these scoring methods have implications for estimating gender differences and comparing them across countries. Sum-Scores and Factor-Scores, for example, result in highly similar trait scores in simulated data despite differences in underlying theoretical assumptions (McNeish, 2023). Hence, Sum-Scores and Factor-Scores seem interchangeable.

The use of MI-Factor-Scores seems scarce given that MI is seldomly tested in psychological research (Maassen et al., 2023). However, comparisons of Sum-Scores with MI-Factor-Scores suggest that the choice of scoring method impacts results (Del Giudice et al., 2012; Eigenhuis et al., 2015; Kaiser et al., 2020). Estimates of differences between men and women in the 16-personality-factor (16-PF) questionnaire were larger if calculated from MI-Factor-Scores than from Sum-Scores (Del Giudice et al., 2012; Kaiser et al., 2020). Likewise, the choice of scoring method impacted cross-country comparisons of personality (Eigenhuis et al., 2015). The authors of these studies consider MI-Factor-Scores to reflect true differences more accurately than Sum-Scores. To my knowledge, the impact of scoring methods has not been replicated for gender differences in the Big Five, nor has it been investigated in relation to the GEPP.

Gender Equality

Assessing the GEPP requires measuring gender equality across countries. Gender equality is a fundamental human right (UN General Assembly, 1948). While different operationalisations of gender equality exist (for an overview, see Else-Quest & Hamilton, 2018), gender equality can be defined as full parity between men and women in access to resources and opportunities (Hausman et al., 2012). Different levels of gender equality reflect partially fulfilled parity, i.e., the degree of disparity (Hausmann et al., 2012; World Economic Forum [WEF], 2024). According to the Global Gender Gap Index (GGGI), gender equality has increased globally at a slow pace from 2006 to 2024 (WEF, 2024).

The Gender-Equality-Personality-Paradox

Evidence for a general Gender-Equality-Paradox exists for several psychological constructs like mathematic attitudes and anxiety, episodic memory, mental health, and personality (Balducci, 2023; Herlitz et al., 2024). Of these constructs, personality seems the one with the clearest support for a Gender-Equality-Paradox (Herlitz et al., 2024).

An informal literature review of studies investigating the GEPP within a FFM framework yielded eight original studies. These vary in their methodology regarding (a) the choice of gender equality indices and FFM-questionnaires, (b) the calculation of gender differences, and (c) the analytical design (e.g., correlation, linear regression, path analysis). The majority of these eight studies support the GEPP (e.g., Costa et al., 2001; Kaiser, 2019; Mac Giolla & Kajonius, 2018; Schmitt, 2019). However, effect sizes of the GEPP as well as its interpretation vary between studies. Studies investigating gender differences in overall personality tend to yield large[2] associations between gender differences and gender equality (Costa et al., 2001; Kaiser, 2019; Mac Giolla & Kajonius, 2018; Schmitt et al., 2008). On the contrary, studies investigating traits separately tend to find smaller correlations for most traits (Ilmarinen & Lönnqvist, 2024; Lippa, 2010; Murphy et al., 2021; Schmitt, 2019). These studies differed from each other regarding the exact traits for which they observed a statistically significant GEPP: most frequently for Extraversion and Conscientiousness (Illmarinen & Lönnqvist, 2024; Murphy et al., 2021; Schmitt, 2019), followed by Agreeableness (Lippa, 2010; Schmitt, 2019) and Neuroticism (Schmitt, 2019). When correlations between gender equality and personality are regularised or controlled for the Human Development Index, the GEPP seems to vanish (Kaiser, 2019; Schmitt et al., 2008).

Explaining the Gender-Equality-Personality-Paradox from Theory

Most theory-based explanations of the GEPP come from evolutionary and social psychology (Balducci, 2023; Schmitt et al., 2008). According to evolutionary theories, gender differences evolved through sexual selection and parental invest (Hyde, 2014; Schmitt, 2015). The cultural variation of these differences are evolved adaptations to environmental conditions like ecological stress (Schmitt, 2015; Schmitt et al., 2008). Since ecological stress tends to co-vary with gender equality (Kaiser, 2019), gender differences vary with gender equality too.

From a social-psychological perspective, gender differences arise from gender stereotypes ascribing different social roles to men and women (Eagly & Wood, 2012). Accordingly, decreasing gender stereotypes should result in smaller gender differences. However, these stereotypes are not dissolved by increasing equality in participation opportunities (Breda et al., 2020), hence the GEPP.

Both explanatory approaches are not mutually exclusive (Balducci, 2023; Eagly & Wood, 2012). Considered separately, however, they draw different conclusions from the GEPP evidence: From an evolutionary perspective, the GEPP is simply an evolved adaptation of evolved traits to environmental conditions. From a social psychological perspective, the GEPP

[2] Correlation size interpreted as small for $r < .24$, medium for $r < .41$, or large for $r \geq .41$ (Lovakov &Agadllina, 2021).

indicates that current means to measure and foster gender equality fail to address gender stereotypes. This would call for adjusted gender policies.

Given these different interpretations, a thorough understanding of the origins of the GEPP seems important for guiding researchers and policy makers in interpreting the GEPP. However, the evidence for both theoretical explanations is mixed (Hyde, 2014; Balducci, 2023). So is, too, the evidence for the GEPP regarding size and specific traits (see above). Due to the methodological heterogeneity of GEPP-studies, searching for explanations of the GEPP within studies' methodology seems worthwhile. Should the GEPP turn out to be a methodological artefact, this would obviate the search for theory-based explanations and call for a refined methodology before any inferences could be drawn from the association between gender equality and gender differences in personality.

Explaining the Gender-Equality-Personality-Paradox from Methodology

Methodological explanations view the GEPP as an artefact arising from study characteristics or analytical procedures. Among the most frequently discussed biases are the choice of gender equality measures (Else-Quest & Hamilton, 2018; Guo et al., 2024; Marsh et al., 2021) and country coverage (Balducci, 2023; Richardson et al., 2020).

Regarding gender equality measures, size and significance of Gender-Equality-Paradoxes depend on the exact aspects of gender equality being measured: Using other gender-equality-indices than the GGGI, the Gender-Equality-Paradox in STEM[3] outcomes vanishes (Guo et al., 2024; Richardson et al., 2020). Moreover, the specific calculation of female to male ratios within the same gender-equality-index can also impact the relation between gender equality and STEM attitudes (Marsh et al., 2021). However, it is unclear how these results translate to research on gender differences in FFM-traits. Studies investigating more than one gender-equality-index found the GEPP irrespective of indices chosen (Costa et al., 2001; Lippa, 2010; Schmitt et al., 2008; Schmitt, 2019). Hence, the GEPP seems less susceptible to differences in gender equality assessment than the Gender-Equality-Paradox in STEM outcomes.

[3] STEM = Science, Technology, Engineering, Mathematics.

Regarding country coverage, western, educated, industrialised, rich, and democratic (WEIRD) countries dominate most samples (Balducci, 2023). This also holds for the eight studies that investigated the GEPP within an FFM framework: Especially South American and African countries are underrepresented (Costa et al., 2001; Ilmarinen & Lönnqvist, 2024; Kaiser, 2019; Lippa, 2010; Mac Giolla & Kajonius, 2018; Murphy et al., 2021; Schmitt et al., 2008). This aligns with evidence that gender differences in several psychological constructs are more strongly related to countries' affiliation with Western culture than with gender equality (Berggren & Bergh, 2023). In Western countries, gender differences tend to be larger than in non-Western countries, which seemingly supports the GEPP as Western countries tend to be more gender-egalitarian (Berggren & Bergh, 2023). When limiting samples to culturally more similar countries, the GEPP is reversed: Gender differences are smaller in more gender-egalitarian countries (Berggren & Bergh, 2023).

In addition to the WEIRD-bias in samples, most studies on Gender-Equality-Paradoxes used inventories developed in WEIRD countries (Berggren & Bergh, 2023). This applies to research on the GEPP too. Inventories developed in WEIRD countries do not necessarily work equally well in non-WEIRD countries (i.e., they might measure constructs non-invariantly across WEIRD and non-WEIRD countries). In the case of personality, measurements tend to be non-invariant across countries (Dong & Dumas, 2020).

If a measurement is non-invariant, some researchers point out that Sum-Scores and Factor-Scores might be biased (e.g., Del Giudice et al., 2012; Eigenhuis et al., 2015; Kaiser et al., 2020). Instead, they recommend using MI-Factor-Scores. Consequently, if a measurement is non-invariant, whether and to what extent researchers find a GEPP might depend on their choice of scoring method. Whether this is the case, cannot be judged based on prior GEPP-research. Only three GEPP studies reported MI-testing (Kaiser, 2019; Mac Giolla & Kajonius, 2018; Murphy et al., 2021). None of these compared scoring methods. Besides, only two of all GEPP studies explicitly stated how they calculated trait scores: Costa et al. (2001) used Factor-Scores and Kaiser (2019) used MI-Factor-Scores of personality traits. Both studies reported a large significant GEPP in overall personality. However, the effect of different scoring methods cannot be reliably judged from just two studies that assessed FFM-traits with different questionnaires and gender equality with different indices.

Research Gap and Present Study

Previous research generally affirmed the existence of a GEPP (at least for certain traits and gender equality measures; e.g., Mac Giolla & Kajonius, 2018; Murphy et al., 2021). However, the origins of the GEPP are yet unclear. In GEPP research, personality is compared across gender and across countries. Hence, measurements need to be invariant across gender and across countries to allow for valid comparisons. If invariance does not hold, some researchers recommend using MI-Factor-Scores instead of Sum-Scores or Factor-Scores. However, so far, no study has investigated whether the choice of scoring method affects the size of the GEPP. Moreover, the effect of these scoring methods on gender differences have only been investigated for personality models other than the FFM.

The present master's thesis was concerned with this gap in research. A publicly available FFM dataset with responses from men and women world-wide was reanalysed and three scoring methods for calculating trait scores were compared: Sum-Scores, Factor-Scores, and MI-Factor-Scores. Specifically, two research questions (RQs) were addressed:

RQ1: Do gender differences in personality vary with the scoring method?

RQ2: Do estimates of the Gender-Equality-Personality-Paradox vary with the scoring method?

Both, RQ1 and RQ2 were explorative. RQ1 was a preparatory step for RQ2 since the GEPP is concerned with gender differences across countries. Gender differences and the GEPP were assessed in single traits and in overall personality. While comparisons of different scoring methods do not allow for inferences about their appropriateness, they can illuminate whether choice of scoring method is arbitrary or whether it impacts conclusions drawn from research. A better understanding of these implications could inform the current debate about the appropriateness of different scoring methods. This could ultimately improve research methodology which would benefit future GEPP research: A refined methodology regarding scoring methods would allow for re-examining the existence of the GEPP which could justify further search of its origins or its potential use in political decision making.

Methods

Sample

The current sample is a subsample of the publicly available personality data collected by Johnson (2014) from 2001 to 2011 (available at https://osf.io/wxvth/). The age range of participants was limited to 19-69 years as personality is thought to be stable within this range (Briley & Tucker-Drop, 2014). All countries with less than 1000 respondents were excluded to ensure large enough sample sizes per country and gender for factor estimates to stabilise (see Hirschfeld et al., 2014). To account for overrepresentation of some countries (e.g, n_{USA} = 320128 as compared to n_{Norway} = 1059), countries with more than 1000 respondents were limited to 1000 observations by randomly choosing participants. Data from Hong-Kong was excluded as there were no gender equality scores obtainable for it. This led to a total sample of 21 countries with 1000 observations each. Missing values (< 1% per item) were imputed at item-level with the item mean.

Measures

Personality

FFM traits were measured with the IPIP-NEO-120 (Johnson, 2014), a free English-language online questionnaire (available at https://osf.io/tbmh5/). Each trait is measured by 24 items grouped into 6 facets. Participants rated their endorsement of each item on a five-point Likert-type scale (from 1 = "very inaccurate" to 5 = "very accurate"). Higher values indicate higher endorsement. The IPIP-NEO-120 is a valid and reliable personality questionnaire (e.g., Johnson, 2014; Kajonius & Johnson, 2019; Sleep et al., 2021). In the present sample, reliability was acceptable to good[4] (Table 2). Taking the IPIP-NEO-120 was voluntary, fully anonymous, and not related to any risks for the participants. Participants were compensated by receiving their personality results after completion of the questionnaire.

[4] Interpretation guidelines taken from Kalkbrenner (2023).

Gender Equality

Gender equality was measured by the GGGI (WEF, 2024) assessing equality in the domains of economic participation and opportunity, educational attainment, health and survival, and political empowerment. It ranges from 0 (no gender parity) to 1 (gender parity). The index is trunked at 1, so that disparity in favour of women is also indicated by 1. Despite being criticised for its lack of psychometric properties (Else-Quest & Hamilton, 2018) and for oversimplifying a complex social phenomenon (Liebowitz & Zwingel, 2014), the GGGI is frequently used in research (Balducci, 2023). The assessment of the GGGI typically involves counting, for example the number of parliamentary seats occupied by women as compared to men (WEF, 2024). Such ratings cannot be subject to country-specific interpretation so that MI-testing is not needed.

The GGGI commenced in 2006. To obtain one overall gender equality score for the duration of personality-data collection, I followed the analytical approach of Mac Giolla and Kajonius (2018) and averaged each country's GGGI scores from the years 2006 to 2011.

Country and Gender

Participants' country and gender were assessed with the IPIP-NEO-120 online questionnaire. In the present sample, gender was coded dichotomously as male/female. To assess country affiliation, participants were asked to indicate the country to which they felt most affiliated "by virtue of citizenship, length of residence, or acculturation" (Instructions for Completing the IPIP-NEO Short Form, n.d.).

Table 2

Reliability Coefficients of the IPIP-NEO-120 Scale for Each Trait

Reliability Coefficient	O	C	E	A	N
Cronbach's α	0.80	0.90	0.88	0.85	0.89
Coefficient *H*	0.72	0.85	0.87	0.77	0.88

Note. Cronbach's α is an appropriate reliability coefficient for Sum-Scores, coefficient *H* is the appropriate coefficient for (Measurement Invariance) Factor-Scores (McNeish, 2023). O = Openness. C = Conscientiousness. E = Extraversion. A = Agreeableness. N = Neuroticism.

Analytical Approach

General Approach

Prior to analyses, three types of scores were estimated: Sum-Scores, Factor-Scores and MI-Factor-Scores. Model fit was reported for (MI-)Factor-Scores only, since calculating Sum-Scores did not require model fitting. Goodness of fit was judged according to commonly used cut-off criteria: A comparative fit index (CFI) larger than 0.90 and a root mean square error of approximation (RMSEA) smaller than 0.08 were considered as indicating acceptable model fit (Hu & Bentler, 1999; Schermelleh-Engel et al., 2003).

To answer RQ1, gender differences in single traits and in overall personality were calculated from men's and women's trait scores in each country and based on each scoring method. The estimates of gender differences were then compared across scoring methods. Countries' absolute gender-difference-estimates were then used to assess the GEPP (RQ2) by correlating them with gender equality for each trait and for overall personality. GEPP-estimates were compared across scoring methods. All analyses were conducted in R statistical software (v4.3.1; R Core Team, 2023).

Scoring Methods

Sum-Scores. Sum-Scores were calculated by adding up participants' responses on each item belonging to a personality trait. For further analyses, Sum-Scores were z-standardised to a mean of 0 and standard deviation of 1.

Factor Scores. Factor-Scores were estimated from global CFA models. Robust maximum likelihood (MLR) was used as the estimator. Each trait was modelled separately. This was done to account for limited computational power, but also for factor indeterminacy: For univariate models, different extraction methods yield highly similar results (Lechner et al., 2013) so that factor indeterminacy is of less concern in univariate models.

Each trait was modelled from six item-parcels. Each parcel consisted of four items representing a personality facet. Participants' responses to these items were summed up to form the parcel's score. Parcels are appropriate to use when investigating a construct's relation to other constructs (Little et al., 2013), as is the case in this thesis (linking personality traits to gender equality). Item-parcels increase reliability (Meade & Kroustalis, 2005; Rioux et al., 2020) and reduce residual variance (Little et al., 2013). If parcels are well specified (e.g., by theory-driven parcelling as in the present case), parcels are better representations of a construct than single items (Lee & Whittaker, 2021). Models were fitted with the R package lavaan (Rosseel, 2012). Factor-Scores were estimated with the lavaan function lavPredict using the Empirical Bayes Modal approach as extraction method.

Measurement-Invariance Factor Scores. MI-Factor-Scores were estimated from full or partial scalar MI models. MI was assessed in a multi-group CFA framework modelling each trait as specified above. As these trait models included parcels not items, MI also applies to the parcel-level (i.e., facet-level) not to the item-level (Little et al., 2013)

Personality needed to be measurement invariant at scalar level across gender and at metric level across countries (for the rational of the required MI levels see chapter Score Comparability: Measurement Invariance above). To test the required MI levels across gender and across countries, personality-data was grouped into gender-by-country groups (i.e., in a female and a male group per country: Australia-women, Australia-men, Canada-women, Canada-men, etc.) resulting in 42 groups. Across these gender-by-country groups, configural and metric MI were assessed consecutively for each trait model separately. The metric MI model across gender-by-country groups served as a baseline model in testing scalar MI across gender in each country separately (resulting in 21 scalar invariance models). The MI testing approach is graphically exemplified in Appendix A. MI models were fitted with the R package lavaan (Rosseel, 2012). Metric and scalar MI were assumed when their model fit decreased only by ΔCFI ≤ 0.01 and ΔRMSEA ≤ 0.015 as compared to the preceding model (Putnick & Bornstein, 2016). Following the recommendations by Rutkowski and Svetina (2014) for large numbers of groups ($N = 20$), cut-off criteria were relaxed to ΔCFI ≤ 0.02 and ΔRMSEA ≤ 0.03 for testing metric MI.

In case the required level of MI could not be established, partial MI models were fitted using the R package SemTools (Jorgensen et al., 2022). To that end, modification indices were inspected to identify non-invariant items: The item with the largest χ^2-modification index was considered non-invariant and freed from equality constraints (see Lai et al., 2022). At metric level, the loading parameter was freed from equality constraints; at scalar level, the intercept was freed. After one parameter was freed, model fit was assessed again. This process was repeated until a partial metric/scalar model showed only acceptable decrease in model fit or parameters for three item-parcels were freed. As of now, there is no consensus regarding the number of parameters that can be freed without violating the assumption that a model is still mostly invariant (Dong & Dumas, 2020). Given that each trait model consisted of six item-parcels, freeing more than half of the latent trait indicators seemed to suggest non-invariance of this trait rather than partial invariance.

In case, a given trait did not reach at least partial scalar MI in a certain country, that trait in that country was excluded from further analysis for all scoring methods. MI-Factor-Scores were estimated with the lavaan function lavPredict using the Empirical Bayes Modal approach as extraction method.

Assessment of Gender Differences

Gender differences were computed as Cohen's *d* for single traits and as the Mahalanobis' Distance (*D*) for overall personality. Positive values indicate that women scored higher than men. Cohen's *d* was calculated with the R package effsize (v0.8.0; Torchiano, 2016) and *D* with an R-script developed by Del Giudice (2019). To assess whether *D* captured differences across all traits rather than being driven by a single trait, the heterogeneity coefficient H_2 and the equal proportion of variances (EPV_2) were inspected. H_2 ranges from 0 (homogeneity, all variables contribute equally) to 1 (heterogeneity, size of *D* depends on one variable alone; Del Giudice, 2017). The EPV_2 gives the proportion of variables producing the same amount of *D* when all other variables would not contribute to *D* (Del Giudice, 2017). For at least 5 variables, Del Giudice (2017) suggests $EPV_2 \leq 0.20$ as a cut-off criterion for too high levels of heterogeneity.

Both, Cohen's *d* and *D* can be interpreted in terms of standardised standard deviations: For example, if Cohen's $d = 0.5$, women's mean trait score would be half a standard deviation larger than men's mean trait score. The same applies to interpreting *D*. A Cohen's *d* or *D* of 0.15, 0.36, and 0.65 was interpreted as small, medium, and large, respectively (Lovakov & Agadullina, 2021).

Assessing the Gender-Equality-Personality-Paradox

Following prior research approaches (e.g., Lippa, 2010; Kaiser, 2019; Mac Giolla & Kajonius, 2018), the GEPP was judged from the correlation between country-level gender equality and country-level absolute gender differences (20 countries included). A positive correlation indicates a GEPP. The GEPP was calculated for all five single traits and for overall personality for each scoring method. To account for multiple testing, I adjusted the significance level for the correlation coefficients using Bonferroni's correction.

Prior to calculating the correlation, the assumptions of the Pearson product-moment correlation were assessed (continuous data, linear relation between variables, normality; Schober et al., 2018): Continuity of GGGI-scores and gender-difference-estimates could be assumed as well as linearity between these variables based on prior research that assessed personality with the same questionnaire as in the present study (e.g., Kaiser, 2019; Mac Giolla & Kajonius, 2018). Normality of GGGI-scores and gender-difference-estimates per trait and in overall personality were judged from Shapiro-Wilk tests and from density- and quantile-quantile-plots (QQ-plots). This double approach was chosen because small sample sizes (here: n = 20) tend to pass the Shapiro-Wilk test as normally distributed despite being actual non-normally distributed (Le Boedec, 2016). Correlation coefficients of .12, .24, and .41 were considered small, medium, and large, respectively (Lovakov & Agadullina, 2021).

Results

Sample Characteristics

The present dataset contains 21,000 responses from 21 countries. Overall, more women than men answered the questionnaire ($N_{WomenOverall}$ = 11307) with the average age of M_{men} = 28.0 years (SD_{men} = 9.85) and M_{women} = 28.3 years (SD_{women} = 10.1) across countries. Most data was obtained from WEIRD and Asian countries. Only one country from Africa (South Africa) and one from South America (Mexico) had large enough samples to be included. For a list of included countries see Appendix B.

Model Fit and Measurement Invariance

The global trait models (from which Factor-Scores were estimated) showed mostly acceptable to good model fit ranging from $CFI_{Openness}$ = 0.78 to $CFI_{Neuroticism}$ = 0.95 and $RMSEA_{Openness}$ = 0.15 to $RMSEA_{Neuroticism}$ = 0.10 (for all indices see Appendix C). All five traits showed full metric invariance across gender-by-country (Table D1). Tests of scalar MI across gender in each country are summarised in Table D2: In most countries, two to four traits reached at least partial scalar MI across gender. In China, India, Mexico, and the UK, all traits reached at least partial MI across gender. Full scalar MI across gender was most frequently reached for Extraversion and Neuroticism. Openness displayed full scalar MI across gender only in South

Korea. After exclusion of traits from single country samples that did not reach at least partial scalar MI across gender, final samples sizes were $n_{Openness}$ = $n_{Conscientiousness}$ = 13,000, $n_{Extraversion}$ = $n_{Agrerableness}$ = 14,000, and $n_{Neuroticism}$ = 16,000 containing data from 20 countries.

Gender Equality

The average gender equality between the years 2006 and 2011 was comparatively high in the investigated countries (M_{GGGI} = 0.72, SD_{GGGI} = 0.06; for a full list see Appendix B). The least gender-egalitarian country (India) reached an average parity between men and women of roughly 61%. The most gender-egalitarian countries (Finland and Norway) reached an average parity of 82%.

RQ1: Do Gender Differences in Personality Vary with the Scoring Method?

Similarities Between Scoring Methods Regarding Gender Differences

Gender differences across countries are given in Table 3 and gender differences per country in Appendix E. For all countries, coefficient H_2 was smaller than 1 (Sum-Scores: M_{H2} = 0.65; Factor-Scores: M_{H2} = 0.68; MI-Factor-Scores: M_{H2} = 0.72) and EPV_2 was larger than 0.20.

All three scoring methods led to the same interpretation of the size of gender differences in Openness, Conscientiousness, and in Agreeableness (Table 4). Across countries, gender differences were larger in overall personality (medium to large *D*) than in single traits regardless of scoring method. Of all countries, Australia, France, and the USA have the largest gender differences across scoring methods (large *D* ranging between 0.83 and 1.22), while the smallest gender differences are found in South Korea (small *D* ranging between 0.15 and 0.21). Thus, in line with prior research, gender differences tend to be negligible to medium in single traits and large in overall personality in the present sample regardless of scoring method.

Table 3

Descriptive Statistics of Absolute Gender Differences in Single Traits and Overall Personality Across Countries

Scoring Method	*M*	*SD*	*Md*	*IQR*	Min	Max	Skew	Kurtosis
Cohen's *d*: Openness [a]								
Sum-Scores	0.19	0.05	0.19	0.072	0.1	0.27	-0.12	-1.35
Factor Scores	0.17	0.06	0.17	0.093	0.07	0.26	-0.05	-1.41
MI-Factor Scores	0.15	0.12	0.13	0.11	0.01	0.44	1.07	0.64
Cohen's *d*: Conscientiousness [a]								
Sum-Scores	0.09	0.08	0.07	0.15	0	0.25	0.5	-1.27
Factor Scores	0.09	0.08	0.07	0.11	0.01	0.23	0.5	-1.31
MI-Factor Scores	0.11	0.08	0.07	0.10	0.02	0.27	0.87	-0.72
Cohen's *d*: Extraversion [b]								
Sum-Scores	0.12	0.09	0.12	0.14	0	0.28	0.2	-1.44
Factor Scores	0.15	0.11	0.16	0.19	0	0.31	0.1	-1.69
MI-Factor Scores	0.17	0.12	0.19	0.21	0	0.35	0.03	-1.69
Cohen's *d*: Agreeableness [b]								
Sum-Scores	0.45	0.14	0.46	0.22	0.25	0.65	-0.07	-1.6
Factor Scores	0.47	0.18	0.49	0.29	0.14	0.56	-0.31	-1.32
MI-Factor Scores	0.63	0.21	0.7	0.27	0.19	0.92	-0.56	-0.86
Cohen's *d*: Neuroticism [c]								
Sum-Scores	0.34	0.09	0.36	0.15	0.13	0.44	-0.8	-0.24
Factor Scores	0.41	0.1	0.45	0.12	0.16	0.35	-1.04	0.12
MI-Factor Scores	0.48	0.11	0.52	0.11	0.26	0.67	-0.57	-0.8
Mahalanobis' Distance: Overall Personality [d]								
Sum-Scores	0.63	0.2	0.63	0.19	0.17	0.91	-0.72	-0.16
Factor Scores	0.66	0.23	0.7	0.23	0.15	0.94	-0.77	-0.34
MI-Factor Scores	0.84	0.28	0.86	0.29	0.21	1.22	-0.62	-0.36

Note. $n_{Openness}$ = 13. $n_{Conscientiousness}$ = 13. $n_{Extraversiion}$ = 14. $n_{Agreeableness}$ = 14. $n_{Neuroticism}$ = 16. $n_{OverallPersonality}$ = 20.

Differences Between Scoring Methods Regarding Gender Differences

The three scoring methods led to diverging interpretations of the size of gender differences for Extraversion, Neuroticism, and overall personality (Table 4). For these traits, Sum-Scores led to smaller gender differences than Factor-Scores or MI-Factor-Scores. The largest discrepancy between scoring methods is observed in overall personality: The difference between MI-Factor-Scores and Sum-Scores is $\Delta D = 0.21$. While large enough to change the interpretation of gender differences in overall personality from medium to large, it is a rather small difference between the scoring methods.

When looking at the exact numeric size of gender differences and not only at their interpretation as small to large, scoring methods differ to a small degree for all traits. MI-Factor-Scores lead to the largest estimates of gender differences in overall personality and in all traits except Openness (Table 3). However, this trend does not hold in each country (see also Appendix E). To illustrate these country-specific differences between scoring methods, consider gender differences in Openness in China and Finland (Table 5). In Finland, gender-difference are smallest if derived from MI-Factor-Scores (in line with the generally observed trend). In China, gender differences in Openness are largest based on MI-Factor-Scores (reversed trend). Were both countries compared based on either Factor-Scores or Sum-Scores, researchers would conclude that Openness-gender-differences are larger in Finland than in

Table 4

Interpreted Effect Sizes of Gender Differences for each Scoring Method

Scoring Method	Size of gender differences					
	O	C	E	A	N	OP
Sum-Scores	small	negligible	negligible	medium	small	medium
Factor-Scores	small	negligible	small	medium	medium	large
MI-Factor-Scores	small	negligible	small	medium	medium	large

Note. Effect sizes of 0.15, 0.36, and 0.65 were interpreted as small, medium, and large, respectively (Lovakov & Agadullina, 2021). For the exact size of each gender difference see Table 3. O = Openness. C = Conscientiousness. E = Extraversion. A = Agreeableness. N = Neuroticism. OP = Overall Personality. MI = Measurement Invariance.

Table 5

Gender Differences [95%-CI] in Openness in China and Finland

Scoring method	China	Finland
Sum-Scores	0.16 [0.034, 0.29]	0.23 [0.11, 0.36]
Factor-Scores	0.14 [0.011, 0.26]	0.19 [0.062, 0.311]
MI-Factor Scores	0.44 [0.32, 0.57]	0.086 [-0.039, 0.21]

Note. CI = Confidence interval. MI = Measurement invariance.

China. If both countries were compared based on MI-Factor-Scores, then gender differences in Openness appear smaller in Finland than in China. Hence, the choice of scoring method can affect the rank-order of countries regarding the size of gender differences.

The direction of gender differences (whether men or women scored higher) seems mostly unaffected by scoring methods. Only for Openness in Canada, Germany, Ireland, and Sweden, scoring methods suggest diverging directions far enough apart to be considered relevant (i.e., confidence intervals do not overlap; Appendix E): In Openness in these countries, men's scores were higher than women's when estimated with MI-Factor-Scores rather than Factor-Scores or Sum-Scores.

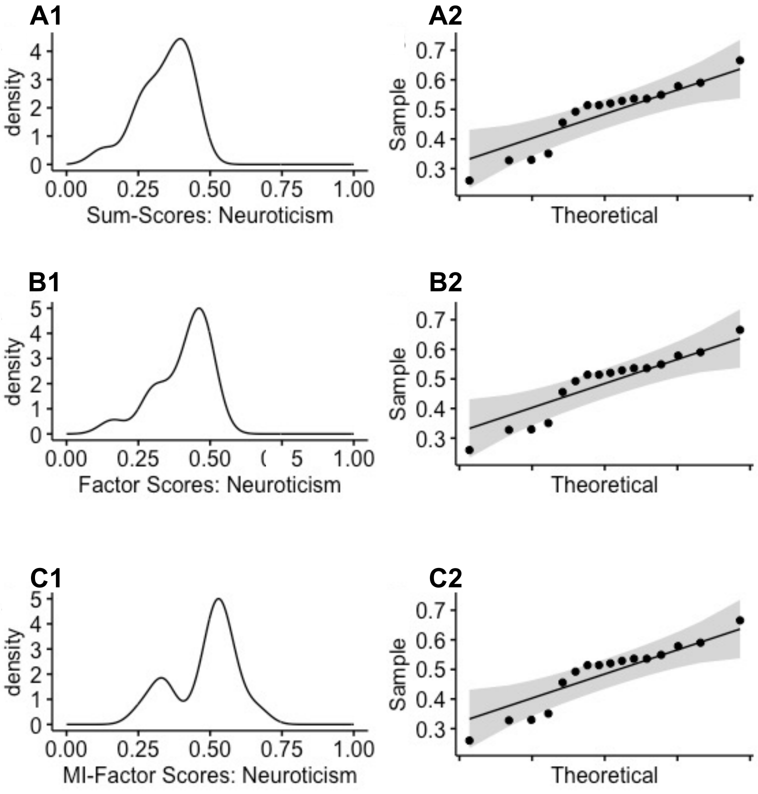
RQ2: Does the Gender-Equality-Personality-Paradox Vary with the Scoring Method?

Examining Assumptions of Pearson Correlation

Normally distributed data was given for gender equality (Table F1; Figures F1-F6). However, the absolute gender differences were not normally distributed across countries: While Shapiro-Wilk tests (Table F1) indicated normality for all traits and scoring methods, density- and QQ-plots of gender differences per trait and scoring method suggest non-normality for most traits and for overall personality (Figures F7-F21). Interestingly, scoring methods seem to impact the distribution shape of gender differences across countries: Sum-Score based gender differences approximate a normal distribution the best, while MI-Factor-Score based differences diverge from this shape the strongest of the three scoring methods (e.g., Figure 1). Hence, Spearman-Rank correlations were calculated instead of Pearson product-moment correlations.

Figure 1

Density- and QQ-plots for Gender Differences in Neuroticism Across Scoring Methods



Note. *N* = 16 countries. Gender Differences were measured as Cohen's *d*. Panels A1, B1, and C1 show the density distribution of gender differences across countries and scoring methods. Panel A2, B2, and C2 show the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample across scoring methods. MI = Measurement invariance.

Similarities Between Scoring Methods Regarding the Gender-Equality-Personality-Paradox

The correlations of gender equality with gender differences in single traits and in overall personality are given in Table 6. Regardless of scoring method, most correlations indicate an insignificant, negligible to moderate GEPP (i.e., positive correlation): Higher levels of gender equality co-occur with larger gender differences in personality traits and in overall personality. The only significant and large result was observed for the correlation between gender equality and Sum-Score-Extraversion. Sum-Scores and Factor-Scores yielded highly similar results for the correlations between gender equality and Openness, Agreeableness, and overall personality. Sum-Scores and MI-Factor-Scores resulted in similar GEPP-estimates for Neuroticism.

For the correlation between gender equality and Extraversion, all three scoring methods agreed regarding direction and size (Figure 2). This agreement of the scoring methods could indicate that the GEPP might be substantial for this trait. However, visual inspection of Figure 2 suggests that gender differences in Extraversion might fall into two groups: below and above medium ranked gender equality (roughly $0.68 < \text{GGGI}_{\text{medium}} < 0.70$). The group below $\text{GGGI}_{\text{medium}}$ displays smaller gender differences than the group above $\text{GGGI}_{\text{medium}}$, thus, confirming the GEPP. Nevertheless, it seems that at least within the group below $\text{GGGI}_{\text{medium}}$ a negative trend is present: Smaller gender differences are associated with higher ranked gender equality (until $\text{GGGI}_{\text{medium}}$ is reached).

Differences Between Scoring Methods Regarding the Gender-Equality-Personality-Paradox

The size of GEPP-estimates varies across scoring methods (see Table 6). Based on Sum-Scores and Factor-Scores, most correlations are medium to large. Based on MI-Factor-Scores, half of the correlations are negligible to small, the other half medium to large. The discrepancy between scoring methods is largest for Openness and overall personality. In overall personality, the GEPP based on MI-Factor-Scores is negligibly small while it is of medium size under Sum-Scores and Factor-Scores. For Openness, the discrepancy is even more extreme: Sum-Scores and Factor-Scores suggest a medium to large GEPP, while MI-Factor-Scores yielded a medium sized negative correlation (Figure 2): Higher levels of gender equality co-occur with smaller gender differences (reversed GEPP). For the scatter plots of the remaining traits consult Appendix G.

Table 6

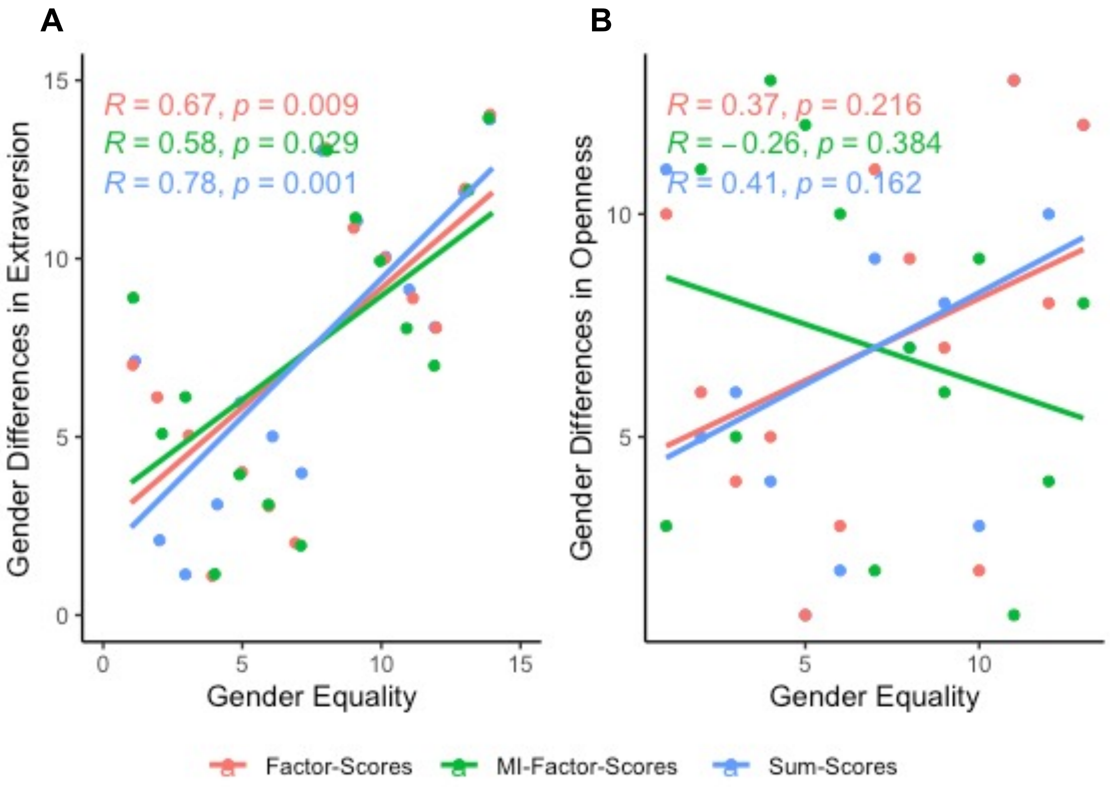
Spearman Correlation between Gender Equality and Gender Differences in Single Traits and Overall Personality

Scoring Method	ρ	95%CI	*p*-value
Openness			
Sum-Scores	.41	[-.18, .78]	0.16
Factor Scores	.37	[-.23, .76]	0.22
MI-Factor Scores	-.26	[-.71, .34]	0.38
Conscientiousness			
Sum-Scores	.16	[-.42, .66]	0.59
Factor Scores	-.033	[-.57, .53]	0.91
MI-Factor Scores	.093	[-.48, .61]	0.76
Extraversion			
Sum-Scores	.78	[.42, .93]	0.001 *
Factor Scores	.67	[.22, .89]	0.009
MI-Factor Scores	.58	[.075, .85]	0.029
Agreeableness			
Sum-Scores	.45	[-.10, .79]	0.10
Factor Scores	.44	[-.12, .79]	0.11
MI-Factor Scores	.22	[-.35, .67]	0.45
Neuroticism			
Sum-Scores	.21	[-.32, .64]	0.44
Factor Scores	.076	[-.44, .55]	0.78
MI-Factor Scores	.24	[-.29, .65]	0.38
Overall Personality			
Sum-Scores	.27	[-.20, .64]	0.25
Factor Scores	.26	[-.21, .63]	0.27
MI-Factor Scores	.096	[-.36, .52]	0.67

Note. Gender differences in single traits were estimated as the absolute Cohen's *d* between men and women. Gender differences in overall personality were estimated as the absolute Mahalanobis' Distance between men and women. Gender Equality was measured as the average GGGI between 2006 and 2011 per country. Using Bonferroni's correction, the significance level ($\alpha = 0.05$) was adjusted to $\alpha_{\text{adjusted}} = 0.003$. $n_{Openness} = 13$. $n_{Conscientiousness} = 13$. $n_{Extraversion} = 14$. $n_{Agreeableness} = 14$. $n_{Neuroticism} = 16$. $n_{OverallPersonality} = 20$.
* $p < 0.003$

Figure 2

Scatterplots: Spearman Correlations between Gender Equality and the absolute Gender Differences in Extraversion and Openness



Note. Gender differences were estimated as the absolute Cohen's *d* between men and women. Gender Equality was measured as the average Global Gender Gap Index between 2006 and 2011 per country. Panel A shows the correlation between gender equality and gender differences in Extraversion. Panel B shows the correlation between gender equality and gender differences in Openness. MI = Measurement invariance.

Discussion

The Present Results

The aim of the present master thesis was to explore the effect of different trait scoring methods on the estimation of gender differences and the GEPP (i.e., the phenomenon of larger gender differences in more gender-egalitarian countries) in the Big Five traits and in overall personality. Data from 20 countries was analysed. Across scoring methods, the GEPP seems substantial only for Extraversion. For Openness, an insignificant GEPP was found under Sum-Scores and Factor-Scores, while a reversed trend (smaller gender differences under higher gender equality) was found for MI-Factor-Scores. To my knowledge, this is the first study to investigate the effect of MI-Factor-Scores in the context of the GEPP.

Does the Gender-Equality-Personality-Paradox Exist?

Across all scoring methods, the correlations of gender equality with personality traits are mostly insignificant and negligible to medium in size. Since all scoring methods support this notion, the GEPP does not seem a methodological artefact arising from scoring methods. However, in the present study, the GEPP seems specific to Extraversion rather than being a universal phenomenon This is further supported by the insignificant, negligible to moderate correlations between gender equality and overall personality: Considering gender differences in all traits simultaneously, the GEPP does not seem relevant.

These results partly align with prior GEPP research. For example, one study found the largest gender-equality-trait-correlation for Extraversion as well (Illmarinen & Lönnquvist, 2024). Another study, in contrast, reported the smallest correlation for Extraversion (Lippa, 2010). This discrepancy might be due to the applied gender equality measure: In the present study I used the averaged GGGI like Illmarinen and Lönnqvist (2024), while Lippa (2010) examined the UN gender related development index and the UN gender empowerment index.

Moreover, studies investigating the GEPP in overall personality found larger correlations than the present study (Kaiser, 2019; MacGiolla & Kajonius, 2018). Both these studies investigated gender differences at facet level and not trait level as in the present study. Since gender equality correlations with facets seem larger than with traits, this might also support the notion that the GEPP applies to specific cases rather than being a universal phenomenon. For a clearer understanding of the influence of scoring methods, the three scoring methods should also be investigated at facet level and under different gender equality measures.

Do Scoring Methods Matter?

Differences between scoring methods were subtle in the present thesis. Generally, gender differences in all traits (except Openness) were larger when calculated from MI-Factor-Scores than from Sum-Scores or Factor-Scores. For some single countries (e.g., China and Finland), the rank order based on gender differences changed depending on the scoring method. Regarding the GEPP, MI-Factor-Scores led to smaller estimates than the other scoring methods (i.e., the larger MI-Factor-Score based gender differences co-varied to a lesser extent with gender equality). The largest discrepancy between scoring methods was found for the correlation between gender equality and Openness: Based on MI-Factor-Scores, the correlation was negative while it was positive based on the other scoring methods. While this seems an extreme difference, it might be due to normal fluctuations: The correlation coefficients' CIs for all scoring methods overlap.

Generally and across countries, however, differences between scoring methods were small and did not change inferences at large: Regardless of scoring methods, gender differences across countries seemed small in most traits and large in overall personality. The GEPP was mostly insignificant. Thus, the choice of scoring method did not change inferences about gender differences or the GEPP.

These results only partly align with prior research regarding gender differences: Just as in the present case, gender differences in the traits of the 16-PF-questionnaire were larger for MI-Factor-Scores than for Sum-Scores (Del Giudice et al., 2012; Kaiser et al., 2020). However, del Giudice et al. (2012) reported changing inferences: Based on MI-Factor-Scores, gender difference in several traits became medium to large as compared to Sum-Scores. This difference in the results between the present study and a prior study might be due to the underlying personality models. The 16-PF captures personality as 16 distinct factors. Most of these factors correspond to the facets of the FFM rather than its five traits examined in this study. Gender differences on the facets of these five traits are larger than gender differences on traits (MacGiolla & Kajonius, 2018). Thus, it seems plausible that gender differences on the facets of the Big Five might be even larger when calculated from MI-Factor-Scores.

In sum whether scoring methods impact inferences depends on the exact research aim: While inferences about the size of gender differences in traits across countries seem mostly unaffected by scoring methods, facet-level investigations, or ranking countries based on gender differences seem more susceptible to the choice of scoring method.

Limitations

The present study analysed data from 20 countries. This hampers generalisability of results, especially to African and South American countries which were underrepresented in the present sample. The conclusions about the GEPP of the present sample apply mostly to WEIRD and Asian countries. Results might therefore be confounded by socio-cultural variables reflecting cultural differences between WEIRD and Asian countries rather than gender equality. Besides, a globally representative sample of countries would also benefit sample size when comparing scoring methods: Since MI-Factor-Scores require at least partial MI which is not obtainable for all traits in all countries, larger numbers of countries could compensate for exclusions based on MI testing.

The necessity of excluding traits in some countries also hampers comparability of the Mahalanobis' Distance across countries: Only for four countries, gender differences in overall personality were calculated from all five traits. The remaining countries varied regarding how many and which traits were used in the calculation of the Mahalanobis' Distance. Especially the extremely small gender differences in overall personality in Singapore and South Korea might be because Agreeableness and Neuroticism had to be excluded in these countries. However, it were these two traits that exhibited the largest gender differences across other countries. Had they been comparable across gender in Singapore and South Korea, gender differences in overall personality would likely have been larger.

It should also be noted that results are limited to the scoring methods compared in this study: Different approaches exist for estimating Factor-Scores and MI-Factor-Scores depending on model specifications. For example, personality trait scores could be estimated from a hierarchical, multidimensional model, from a Bayesian approach to modelling, or from different ways of assessing MI (e.g., via item response theory or alignment method). I do not expect that minor changes to this study's specifications of personality models would alter results at large. For example, in prior research, Factor-Scores derived from a multidimensional model led to trait estimates highly similar to those based on Sum-Scores (McNeish, 2023); the same trend was observed in the present study. However, a completely different approach to modelling personality might lead to different inferences regarding the effects of scoring method. It is therefore important to keep in mind that the scoring methods compared in this thesis each represent only one way to estimate Sum-Scores, Factor-Scores, and MI-Factor-Scores. My choice of modelling approaches represents a compromise between the computational power available to me and comparability to prior GEPP studies. Nevertheless, comparing the implications of different approaches to estimating MI-Factor-Scores would also be insightful for future research but was beyond the scope of this master's thesis.

Finally, the present study was of explorative nature; its results do not allow inferences about the appropriateness of the compared scoring methods. Such inferences would require comparing scoring methods in simulated data for which the true trait scores are known. However, as prior research on potential differences in trait estimates based on different scoring methods is scarce, an explorative approach seemed necessary to first establish whether there are any differences between scoring methods before applying more complex research designs to test their appropriateness.

Implications for Future Research and Practical Use

As this study demonstrated, the GEPP is significant only for Extraversion and does not depend on the scoring method. Future studies should therefore aim for replicating this result and investigating potential causes of the covariation between gender differences in Extraversion and gender equality. Such studies could also strive for better country coverage or investigate the GEPP for more than two genders. It would also be insightful, to investigate whether the reversed GEPP found for Openness under MI-Factor-Scores replicates in different samples.

Whether applied researchers need to carefully consider their choice of scoring method depends on their research aim. If they are interested in general trends in gender differences or the GEPP, the choice of scoring method seems arbitrary. However, if researchers are interested in the exact size of the GEPP or in ranking countries based on gender differences, scoring methods can change inferences. Which method is the most appropriate to use in these cases, is still an ongoing debate. Nevertheless, researchers should report their choice of scoring method to facilitate comparability and replicability. They could also conduct their analyses with different scoring methods to determine whether inferences differ or not. This would help interpret results more carefully even in the absence of a scoring method commonly accepted as appropriate.

In the long run however, further psychometric studies are needed to provide researchers with recommendations on the appropriateness of each scoring method. Such studies could also investigate whether the impact of scoring method varies with the applied questionnaire or investigated countries, for instance. Ultimately, this would improve the methodology of applied personality research which could then be a reliable basis for policy makers, for example regarding gender policies. With this thesis, I hope to contribute to this endeavour by providing an insight into how scoring methods impact inferences regarding gender differences and the GEPP.

Conclusion

The Gender-Equality-Personality-Paradox seems stable across scoring methods: The association between country-level gender differences in personality and country-level gender equality is insignificant, weak, and positive for overall personality and single traits except Extraversion. For this trait, the paradox seems large and substantial regardless of scoring method. The size of gender differences across countries is also unaffected by choice of scoring method: small for single traits, large for overall personality. However, in some countries, differences between scoring methods are more pronounced which can affect countries' rank-order based on gender differences. In summary, researchers should report their chosen scoring method to enhance comparability and replicability of their results. The Gender-Equality-Personality-Paradox does not seem a statistical artefact as far as scoring methods are concerned.

References

Allik, J., Realo, A., McCrae, R. R. (2013). Universality of the five-factor model of personality. In T. A. Widiger, & P. T. Jr. Costa (Eds.): Personality disorders and the five-factor model of personality (3rd ed., pp. 61-74). American Psychological Association. https://doi.org/10.1037/13939-005

Ashton, M. C., & Lee, K. (2005). A defence of the lexical approach to the study of personality structure. *European Journal of Personality, 19*(1), 5-24. https://doi.org/10.1002/per.541

Balducci, M. (2023). Linking gender differences with gender equality: A systematic-narrative literature review of basic skills and personality. *Frontiers in Psychology*, *14*, Article 1105234. https://doi.org./10.3389/fpsyg.2023.1105234

Berggren, M., & Bergh, R. (2023). Simpson's Gender-Equality Paradox. https://doi.org/10.31234/osf.io/mfhyw

Breda, T., Jouini, E., Napp, C., & Thebault, G. (2020). Gender stereotypes can explain the gender-equality paradox. *Proceedings of the National Academy of Sciences*, *117*(49), 31063-31069. https://doi.org/10.1073/pnas.2008704117

Briley, D. A., & Tucker-Drob, E. M. (2014). Genetic and environmental continuity in personality development: A meta-analysis. *Psychological Bulletin*, 140(5), 1303–1331. https://doi.org/10.1037/a0037091

Costa Jr, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: robust and surprising findings. *Journal of Personality and Social Psychology*, *81*(2), 322-331. https://doi.org/10.1037/10022-3514.XI.2.322

Del Giudice, M. (2017). Heterogeneity coefficients for Mahalanobis' D as a multivariate effect size. *Multivariate Behavioral Research*, *52*(2), 216-221. https://doi.org/ 1080/00273171.2016.1262237

Del Giudice, M. (2019). Measuring sex differences and similarities. In D. P. van der Laan, & W. I. Wong (Eds.), *Gender and sexuality development: Contemporary theory and research*. Springer. h, https://doi.org/10.1007/978-3-030-84273-4_1

Del Giudice, M., Booth, T., & Irwing, P. (2012). The distance between Mars and Venus: Measuring global sex differences in personality. *PloS ONE*, *7*(1), Article e29265. https://doi.org/10.1371/journal.pone.0029265

Dong, Y., & Dumas, D. (2020). Are personality measures valid for different populations? A systematic review of measurement invariance across cultures, gender, and age. *Personality and Individual Differences*, *160*, Article 109956. https://doi.org/10.1016/j.paid.2020.109956

Eagly, A. H., & Wood, W. (2012). Social role theory. In P. A. M. van Lange, & A. W. Kruglanski (Eds.), *Handbook of Theories of Social Psychology* (pp. 458-476). Sage.

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

Else-Quest, N. M., & Hamilton, V. (2018). Measurement and analysis of nation-level gender equity in the psychology of women. In C. B Travis, J. W. White, A. Rutherford, W. S. Williams, S. L. Cook, & K. F. Wyche (Eds.), *APA handbook of the psychology of women: Perspectives on women's private and public lives* (pp. 545–563). American Psychological Association. https://doi.org/10.1037/0000060-029

Fedvadjiev, V., & van de Vijver, F. J. R. (2015). Universality of the five-factor model of personality. In J. D. Wright (Ed.), I*nternational Encyclopedia of Social and Behavioral Sciences* (2nd ed., pp. 249-253). Reed Elsevier.

Fors Connolly, F., Goossen, M., & Hjerm, M. (2020). Does gender equality cause gender differences in values? Reassessing the gender-equality-personality paradox. *Sex Roles*, *83*(1), 101-113. https://doi.org/10.1007/s11199-019-01097-x

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, *48*(1), 26-34.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*(4), 322-331. https://doi.org/10.1037//1082-989X.6.4.430

Guo, J., Marsh, H. W., Parker, P. D., & Hu, X. (2024). Cross-Cultural Patterns of Gender Differences in STEM: Gender Stratification, Gender Equality and Gender-Equality Paradoxes. *Educational Psychology Review*, *36*(2), 37.

Hausmann, R., Tyson, L. D., Bekhouche, Y., & Zahidi, S. (2012). The global gender gap index 2012. *The Global Gender Gap Report*, *2012*, 3-27.

Herlitz, A., Hönig, I., Hedebrant, K., & Asperholm, M. (2024). A systematic review and new analyses of the gender-equality paradox. *Perspectives on Psychological Science*, 17456916231202685. https://doi.org/10.1177/17456916231202685

Hirschfeld, G., Von Brachel, R., & Thielsch, M. (2014). Selecting items for Big Five questionnaires: At what sample size do factor loadings stabilize? *Journal of Research in Personality*, *53*, 54–63. https://doi.org/10.1016/j.jrp.2014.08.003

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. https://doi.org/10.1080/10705 51990 9540118

Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology*, *65*(1), 373-398. https://doi.org/10.1146/annurev-psych-010213-115057

Ilmarinen, V. J. & Lönnqvist, J. E. (2024). Deconstructing the gender-equality paradox. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000508

Instructions for Completing the IPIP-NEO Short Form. (n.d.). Retrieved August 12, 2024, from https://drj.virtualave.net/IPIP/shortipipneo1.cgi

Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, *51*, 78-89. https://doi.org/10.1016/j.jrp.2014.05.003

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). semTools: Useful tools for structural equation modeling. R package version 0.5-6. Retrieved from https://CRAN.R-project.org/package=semTools

Jung, E., & Yoon, M. (2016). Comparisons of three empirical methods for partial factorial invariance: Forward, backward, and factor-ratio tests. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 567-584. https://doi.org/10.1080/10705511.2015.1138092

Kaiser, T. (2019). Nature and evoked culture: Sex differences in personality are uniquely correlated with ecological stress. *Personality and Individual Differences*, *148*, 67-72. https://doi.org/10.1016/j.paid.2019.05.011

Kaiser, T., Del Giudice, M., & Booth, T. (2020). Global sex differences in personality: Replication with an open online dataset. *Journal of Personality*, *88*(3), 415-429. https://doi.org/10.1111/jopy.12500

Kajonius, P. J., & Johnson, J. (2018). Sex differences in 30 facets of the five factor model of personality in the large public (N= 320,128). *Personality and Individual Differences*, *129*, 126-130. https://doi.org/10.1016/j.paid.2018.03.026

Kajonius, P. J., & Johnson, J. A. (2019). Assessing the structure of the Five Factor Model of Personality (IPIP-NEO-120) in the public domain. *Europe's Journal of Psychology*, *15*(2), 260-275. https://doi.org/10.5964/ejop.v15i2.1671

Kalkbrenner, M. T. (2023). Alpha, omega, and H internal consistency reliability estimates: Reviewing these options and when to use them. *Counseling Outcome Research and Evaluation*, *14*(1), 77-88. https://doi.org/10.1080/21501378.2021.1940118

Lai, M. H., Liu, Y., & Tse, W. W. Y. (2022). Adjusting for partial invariance in latent parameter estimation: Comparing forward specification search and approximate invariance methods. *Behavior Research Methods*, *54*(1), 414-434. https://doi.org/10.3758/s13428-021-01560-2

Lai, M. H. C., & Tse, W. W.-Y. (2024). Are factor scores measurement invariant? *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000658

Lechner, C. M., Bhaktha, N., Groskurth, K., & Bluemke, M. (2021). Why ability point estimates can be pointless: a primer on using skill measures from large-scale assessments in secondary analyses. *Measurement Instruments for the Social Sciences*, *3*, 1-16. https://doi.org/10.1186/s42409-020-00020-5

Lee, J., & Whittaker, T. A. (2021). The impact of item parceling on structural parameter invariance in multi-group structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(5), 684-698. https://doi.org/10.1080/10705511.2021.1890604

Liebowitz, D. J., & Zwingel, S. (2014). Gender equality oversimplified: Using CEDAW to counter the measurement obsession. *International Studies Review*, *16*(3), 362-389. https://doi.org/10.1111/misr.12139

Lippa, R. A. (2010). Sex differences in personality traits and gender-related occupational preferences across 53 nations: Testing evolutionary and social-environmental theories. *Archives of Sexual Behavior*, *39*, 619-636. https://doi.org/10.1007/s10508-008-9380-7

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, *18*(3), 285-300. https://doi.org/10.1037/a0033266

Le Boedec, K. (2016). Sensitivity and specificity of normality tests and consequences on reference interval accuracy at small sample size: a computer-simulation study. *Veterinary Clinical Pathology*, *45*(4), 648-656. https://doi.org/10.1111/vcp.12390

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, *51*(3), 485-504. https://doi.org/10.1002/ejsp.2752

Maassen, E., D'Urso, E. D., Van Assen, M. A., Nuijten, M. B., De Roover, K., & Wicherts, J. M. (2023). The dire disregard of measurement invariance testing in psychological science. *Psychological Methods*. Advance online publication. https://dx.doi.org/10.1037/met0000624

Mac Giolla, E., & Kajonius, P. J. (2018). Sex differences in personality are larger in gender equal countries: Replicating and extending a surprising finding. *International Journal of Psychology*, *54*(6), 705-711. https://doi.org/10.1002/ijop.12529

Marsh, H. W., Parker, P. D., Guo, J., Basarkod, G., Niepel, C., & Van Zanden, B. (2021). Illusory gender-equality paradox, math self-concept, and frame-of-reference effects: New integrative explanations for multiple paradoxes. *Journal of Personality and Social Psychology, 121*(1), 168–183. https://doi.org/10.1037/pspp0000306

McCrae, R. R., & Costa, P. T. (2008). Empirical and theoretical status of the five-factor model of personality traits. In G. J. Boyle, G. Matthews, & D. Saklofske (Eds.), *Sage Handbook of Personality Theory and Assessment: Vol. 1. Personality Theories and Models* (pp. 273-294). Sage.

McNeish, D. (2023). Psychometric properties of sum scores and factor scores differ even when their correlation is 0.98: A response to Widaman and Revelle. *Behavior Research Methods*, *55*(8), 4269-4290.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*, 2287-2305. https://doi.org/10.3758/s13428-020-01398-0

Meade, A. W., & Kroustalis, C. M. (2005). Problems of item parceling with CFA tests of measurement invariance. In *20th Annual Conference of the Society for Industrial and Organizational Psychology*, Los Angeles, CA (pp. 1-14).

Mõttus, R., Wood, D., Condon, D. M., Back, M. D., Baumert, A., Costantini, G., Epskamp, A., Greiff, S., Johnson, W., Lukaszewski, A., Murray, A., Revelle, W., Wright, A. G. C., Yarkoni, T., Ziegler, M., & Zimmermann, J. (2020). Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the Big Few traits. *European Journal of Personality*, *34*(6), 1175-1201. https://doi.org/10.1002/per.2311

Murphy, S. A., Fisher, P. A., & Robie, C. (2021). International comparison of gender differences in the five-factor model of personality: An investigation across 105 countries. *Journal of Research in Personality*, *90*, Article 104047. https://doi.org/10.1016/j.jrp.2020.104047

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71-90. https://doi.org/10.1016/j.dr.2016.06.004

R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria.

Richardson, S. S., Reiches, M. W., Bruch, J., Boulicault, M., Noll, N. E., & Shattuck-Heidorn, H. (2020). Is there a gender-equality paradox in science, technology, engineering, and math (STEM)? Commentary on the study by Stoet and Geary (2018). *Psychological Science*, *31*(3), 338-341. https://doi.org/10.1177/0956797619872762

Rigdon, E. E., Becker, J. M., & Sarstedt, M. (2019). Parceling cannot reduce factor indeterminacy in factor analysis: A research note. *Psychometrika*, *84*(3), 772-780. https://doi.org/10.1007/s11336-019-09677-2

Rioux, C., Stickley, Z. L., Odejimi, O. A., & Little, T. D. (2020). Item parcels as indicators: Why, when, and how to use them in small sample research. In R. van de Shoot, & M. Miočević (Eds.), *Small Sample Size Solutions* (pp. 203-214). Routledge.

Robitzsch, A., & Lüdtke, O. (2023). Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons.

Structural Equation Modeling: A Multidisciplinary Journal, *30*(6), 859-870. https://doi.org/10.1080/10705511.2023.2191292

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31-57. https://doi.org/10.1177/0013164413498257

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2)*,* 23–74.

Schmitt, D. P. (2015). The evolution of culturally-variable sex differences: Men and women are not always different, but when they are... It appears not to result from patriarchy or sex role socialization. In T. K. Shackelford & R. D. Hansen (Eds.), *The Evolution of Sexuality* (pp. 221–256). Springer. https://doi.org/10.1007/978-3-319-09384-0_11

Schmitt, D. P. (2019). Why sometimes a man is more like a woman: Insights into the "Gender Paradox" of psychological sex differences around the world. In A. Realo (Ed.), *In Praise of an Inquisitive Mind. A Festschrift in Honor of Jüri Allik on the Ocassion of his 70th Birthday.* (pp. 141-150).

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*(1), 168-182. https://doi.org/10.1037/0022-3514.94.1.168

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia*, *126*(5), 1763-1768.

Sleep, C. E., Lynam, D. R., & Miller, J. D. (2021). A comparison of the validity of very brief measures of the Big Five/Five-Factor Model of personality. *Assessment*, 28(3), 739-758. https://doi.org/10.1177/1073191120939160

Steinmetz, H. (2013). Analyzing observed composite differences across groups. *Methodology*, *9*(1), 1-12.

Torchiano, M. (2016). *Effsize - A package for efficient effect size computation*. Zenodo. https://doi.org/10.5281/ZENODO.1480624

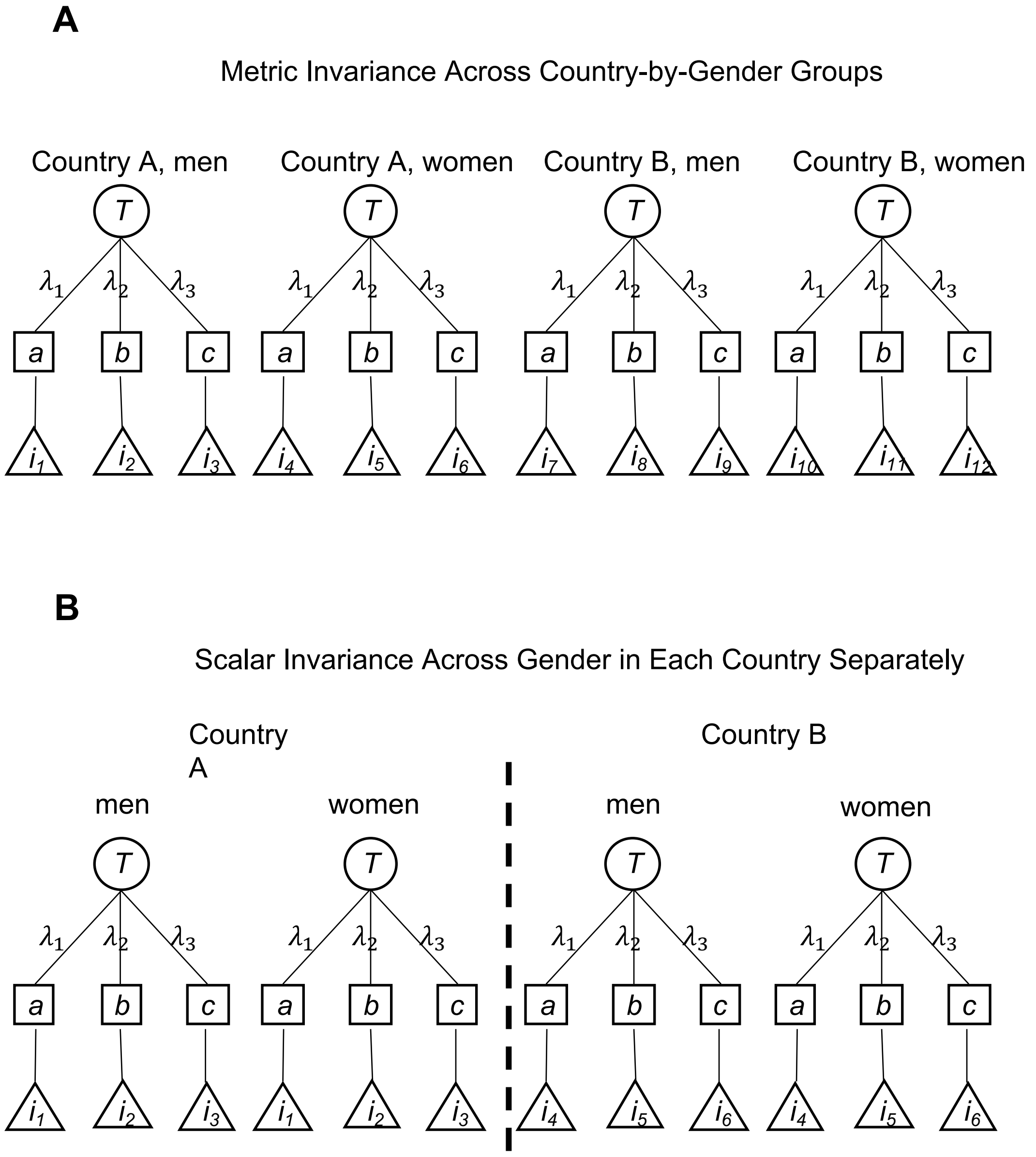
UN General Assembly. (1948). *Universal declaration of human rights* (217 [III] A)

Widaman, K. F., & Revelle, W. (2024). Thinking about sum scores yet again, maybe the last time, we don’t know, Oh no…: a comment on. *Educational and Psychological Measurement*, *84*(4), 637-659. https://doi.org/10.1177/00131644231205310

World Economic Forum. (2024). *Global Gender Gap 2024. Insight Report June 2024* https://www3.weforum.org/docs/WEF_GGGR_2024.pdf

Appendix A

Illustrative Example of Measurement-Invariance-Testing Procedure


A
Metric Invariance Across Country-by-Gender Groups
Country A, men
Country A, women
Country B, men
Country B, women
T
λ1 λ2 λ3
a b c
i1 i2 i3 i4 i5 i6 i7 i8 i9 i10 i11 i12
B
Scalar Invariance Across Gender in Each Country Separately
Country A
Country B
men
women
men
women
i1 i2 i3 i1 i2 i3 i4 i5 i6 i4 i5 i6

Note. This example is for illustrative purposes; trait models are simplified in this figure and the number of countries reduced from 21 to 2. In panel A, all factor loadings are the same across country-by-gender groups indicating metric invariance. Scalar invariance was tested across gender in each country separately (Panel B): Men and women in both countries have the same loadings estimated in the metric model across country-by-gender groups; the intercepts, however, are constrained to be equal only across gender for each country separately. T = any given trait. λ = factor loading for the items a, b, and c. i = intercept of items a, b, and c.

Appendix B

Sample Size and Gender Equality in each Country

Country	Sample Size Men	Sample Size Women	GGGI
Australia	371	629	0.72
Canada	393	607	0.72
China	406	594	0.68
Finland	474	526	0.82
France	563	437	0.70
Germany	498	502	0.75
India	641	359	0.61
Ireland	428	572	0.76
Malaysia	361	639	0.65
Mexico	562	438	0.65
Netherlands	517	483	0.74
New Zealand	424	576	0.78
Norway	556	444	0.82
Philippines	346	654	0.76
Romania	438	562	0.68
Singapore	455	545	0.67
South Africa	449	551	0.74
South Korea	477	523	0.62
Sweden	512	488	0.81
United Kingdom	441	559	0.74
United States of America	381	619	0.72

Note. GGGI = Global Gender Gap Index averaged across the years 2006 to 2011.

Appendix C

Model Fit Indices of Global Personality Models for each Trait

Trait	*df*	χ^2	*p*	*CFI*	*RMSEA* [95%-CI]
Openness [a]	9	2722.4	0	0.78	0.15 [0.15, 0.16]
Conscientiousness [a]	9	1315.7	0	0.94	0.11 [0.10, 0.11]
Extraversion [b]	9	2551.1	0	0.90	0.14 [0.14, 0.15]
Agreeableness [b]	9	2878.6	0	0.83	0.15 [0.15, 0.16]
Neuroticism [c]	9	1572.2	0	0.95	0.10 [0.10, 0.11]

Note. CFI = robust comparative fit index. *RMSEA* = robust root mean square error of approximation.

[a] $n = 13000$

[b] $n = 14000$

[c] $n = 16000$

Appendix D

Measurement Invariance Tests Across Gender-by-Country Groups and Across Gender

Table D1

Model Fit Indices for Configural and Metric Invariance Models Across Gender-by-Country

Model	df	χ^2	*p*	*CFI*	*RMSEA* [95%-CI]	Δ*CFI*	Δ*RMSEA*
				Openness			
configural	378	5235.1	0	0.76	0.16 [0.16, 0.16]	/	/
metric	583	5699.9	0	0.74	0.13 [0.13, 0.14]	0.02	-0.03
				Conscientiousness			
configural	378	2761.6	0	0.94	0.11 [0.11, 0.12]	/	/
metric	583	3454.8	0	0.93	0.098 [0.094, 0.10]	0.01	-0.012
				Extraversion			
configural	378	5430.5	0	0.87	0.16 [0.16, 0.17]	/	/
metric	583	6163.0	0	0.86	0.14 [0.13, 0.14]	0.01	-0.03
				Agreeableness			
configural	378	5456.6	0	0.78	0.16 [0.16, 0.17]	/	/
metric	583	6072.0	0	0.77	0.14 [0.13, 0.14]	0.01	-0.02
				Neuroticism			
configural	378	2473.1	0	0.95	0.10 [0.10, 0.11]	/	/
metric	583	3039.9	0	0.94	0.091 [0.087, 0.094]	0.01	-0.009

Note. Configural and metric invariance were tested across 42 gender-by-country groups (two genders by 21 countries). The differences between fit indices were calculated so that positive values indicate a decrease in model fit and negative values an increase in model fit. *CFI* = robust comparative fit index. *RMSEA* = robust root mean square error of approximation. Δ*CFI* = $CFI_{configural} - CFI_{metric}$; Δ*RMSEA* = $RMSEA_{metric} - RMSEA_{configural}$.

Table D2

Model Fit Indices for Scalar (Partial) Invariance Models Across Gender in Each Country

Model	*df*	χ^2	*p*	*CFI*	*RMSEA* [95%-CI]	Δ*CFI*	Δ*RMSEA*
Australia: Openness							
scalar	33	453.3	0	0.55	0.16 [0.15, 0.17]	0.19	0.03
Partial scalar (intercepts freed for facets O3, O2, and O1)	30	322.8	0	0.69	0.14 [0.13, 0.15]	0.05	0.01
Australia: Conscientiousness							
scalar	33	223.1	0	0.91	0.11 [0.092, 0.012]	0.02	0.012
Partial scalar (intercept of facet C6 freed)	32	203.1	0	0.92	0.10 [0.088, 0.12]	0.01	0.002
Australia: Extraversion							
Scalar	33	390.9	0	0.81	0.15 [0.14, 0.16]	0.06	0.01
Partial scalar (intercepts for facets E3, E6, E4 freed)	30	378.7	0	0.82	0.15 [0.14, 0.17]	0.04	0.01
Australia: Agreeableness							
Scalar	33	379.8	0	0.70	0.14 [0.13, 0.16]	0.07	0
Partial scalar (intercepts A5, A1 freed)	31	306.7	0	0.76	0.13 [0.12, 0.15]	0.01	-0.01
Australia: Neuroticism							
Scalar	33	170.1	0	0.94	0.091 [0.077, 0.11]	0	0
Canada: Openness							
Scalar	33	469.8	0	0.59	0.16 [0.15, 0.18]	0.15	0.03
Partial scalar (intercepts of facets O3, O2 freed)	31	307.8	0	0.74	0.13 [0.12, 0.15]	0	0
Canada: Conscientiousness							
Scalar	33	195.1	0	0.92	0.097 [0.083, 0.11]	0.01	-0.001
Canada: Extraversion							
Scalar	33	465.0	0	0.78	0.17 [0.15, 0.18]	0.08	0.02
Partial scalar (intercepts for facets E3, E6, E5 freed)	30	425.5	0	0.80	0.16 [0.15, 0.18]	0.06	0.02
Canada: Agreeableness							
Scalar	33	362.4	0	0.68	0.14 [0.13, 0.16]	0.09	0
Partial scalar (intercepts for facets A5, A1, A6 freed)	30	323.4	0	0.71	0.14 [0.13, 0.15]	0.06	0

Model	*df*	χ^2	*p*	*CFI*	*RMSEA* [95%-CI]	Δ*CFI*	Δ*RMSEA*
Canada: Neuroticism							
Scalar	33	179.6	0	0.93	0.093 [0.08, 0.11]	0.01	0.002
China: Openness							
Scalar	33	269.9	0	0.72	0.12 [0.11, 0.13]	0.02	-0.01
Partial scalar (intercept for facet O5 freed)	32	241.0	0	0.75	0.11 [0.10, 0.13]	-0.01	-0.02
China: Conscientiousness							
Scalar	33	115.3	0	0.95	0.068 [0.053, 0.084]	-0.02	-0.03
China: Extraversion							
Scalar	33	229.0	0	0.87	0.11 [0.094, 0.12]	-0.01	-0.03
China: Agreeableness							
Scalar	33	283.6	0	0.7	0.12 [0.11, 0.14]	0.07	-0.02
Partial scalar (intercepts for facets A6, A3, A1 freed)	30	236.3	0	0.76	0.12 [0.10, 0.13]	0.01	-0.02
China: Neuroticism							
Scalar	33	160.4	0	0.91	0.086 [0.072, 0.10]	0.03	-0.005
Partial Scalar (intercepts for facets N3, N4 freed)	31	140.4	0	0.93	0.082 [0.067, 0.097]	0.01	-0.009
Finland: Openness							
Scalar	33	372.7	0	0.67	0.14 [0.13, 0.16]	0.07	0.01
Partial scalar (intercept for facet O3 freed)	32	304.7	0	0.73	0.13 [0.12, 0.14]	0.01	0
Finland: Conscientiousness							
Scalar	33	215.0	0	0.90	0.10 [0.09, 0.12]	0.03	0.002
Partial scalar (intercepts of facets C6, C4, C1 freed)	30	202.8	0	0.90	0.11 [0.092, 0.12]	0.03	0.003
Finland: Extraversion							
Scalar	33	191.7	0	0.92	0.097 [0.083, 0.11]	-0.06	-0.043
Finland: Agreeableness							
Scalar	33	285.6	0	0.76	0.12 [0.083, 0.11]	0.01	-0.02
Finland: Neuroticism							
Scalar	33	236.9	0	0.9	0.11 [0.096, 0.12]	0.04	0.019

Model	*df*	χ^2	*p*	*CFI*	*RMSEA* [95%-CI]	Δ*CFI*	Δ*RMSEA*
Partial scalar (intercepts of facets N4, N2 freed)	31	175.8	0	0.93	0.095 [0.081, 0.11]	0.01	0.004
France: Openness							
Scalar	33	470.4	0	0.46	0.16 [0.15, 0.18]	0.28	0.03
Partial scalar (intercepts for facets O3, O2, O1)	30	281.2	0	0.69	0.13 [0.12, 0.14]	0.05	0
France: Conscientiousness							
Scalar	33	211.7	0	0.87	0.10 [0.089, 0.12]	0.06	0.002
Partial scalar (intercepts for facets C4, C3, C6 freed)	30	170.5	0	0.90	0.096 [0081, 0.11]	0.03	-0.002
France: Extraversion							
Scalar	33	305.6	0	0.86	0.13 [0.11, 0.14]	0	-0.01
France: Agreeableness							
Scalar	33	306.8	0	0.76	0.13 [0.11, 0.14]	0.01	-0.01
France: Neuroticism							
Scalar	33	243.5	0	0.89	0.11 [0.099, 0.13]	0.05	0.019
Partial scalar (intercepts for facets N4, N3 freed)	31	160.0	0	0.93	0.091 [0.076, 0.11]	0.01	0
Germany: Openness							
Scalar	33	401.5	0	0.60	0.15 [0.14., 0.16]	0.14	0.02
Partial scalar (intercepts for facets O3, O2 freed)	31	253.1	0	0.76	0.12 [0.11, 0.13]	-0.02	-0.01
Germany: Conscientiousness							
Scalar	33	208.4	0	0.90	0.10 [0.088, 0.12]	0.03	0.002
Partial scalar (intercepts for facets C6, C3, C1 freed)	30	173.6	0	0.91	0.097 [0.082, 0.11]	0.02	-0.001
Germany: Extraversion							
Scalar	33	255.8	0	0.89	0.12 [0.10, 0.13]	-0.03	-0.02
Germany: Agreeableness							
Scalar	33	386.6	0	0.66	0.15 [0.13, 0.16]	0.11	0.01
Partial scalar (intercepts for facets A1, A4, A5 freed)	30	336.4	0	0.71	0.14 [0.13, 0.16]	0.06	0
Germany: Neuroticism							
Scalar	33	201.3	0	0.92	0.10 [0.086, 0.11]	0.02	0.009

Model	*df*	χ^2	*p*	*CFI*	*RMSEA* [95%-CI]	Δ*CFI*	Δ*RMSEA*
Partial scalar (intercept for facet N3 freed)	32	179.5	0	0.93	0.095 [0.081, 0.11]	0.01	0.004
India: Openness							
Scalar	33	307.1	0	0.66	0.13 [0.12, 0.14]	0.08	0
Partial scalar (intercepts for facets O3, O2 freed)	31	245.9	0	0.73	0.12 [0.10, 0.13]	0.01	-0.01
India: Conscientiousness							
Scalar	33	196.8	0	0.91	0.098 [0.084, 0.11]	0.02	0
Partial scalar (intercept for facet C1 freed)	32	178.3	0	0.92	0.094 [0.079, 0.11]	0.01	-0.004
India: Extraversion							
Scalar	33	281.9	0	0.84	0.12 [0.11, 0.14]	0.02	-0.02
Partial scalar (intercepts for facet E3 freed)	32	264.6	0	0.85	0.12 [0.11, 0.13]	0.01	-0.02
India: Agreeableness							
Scalar	33	235.9	0	0.83	0.11 [0.096, 0.12]	-0.06	-0.03
India: Neuroticism							
Scalar	33	156.3	0	0.93	0.086 [0.072, 0.10]	0.01	-0.005
Ireland: Openness							
Scalar	33	479.8	0	0.61	0.16 [0.15, 0.18]	0.13	0.03
Partial scalar (intercepts for facets O3, O2 freed)	31	342.8	0	0.73	0.14 [0.13, 0.16]	0.01	0.01
Ireland: Conscientiousness							
Scalar	33	169.0	0	0.93	0.09 [0.075, 0.10]	0	-0.008
Ireland: Extraversion							
Scalar	33	453.5	0	0.78	0.16 [0.14, 0.17]	0.08	0.02
Partial scalar (intercepts for facets E3, E1, E5 freed)	30	424.4	0	0.80	0.16 [0.15, 0.18]	0.06	0.02
Ireland: Agreeableness							
Scalar	33	420.0	0	0.64	0.15 [0.14, 0.17]	0.13	0.01
Partial scalar (intercepts for facets A5, A1, A6 freed)	30	373.6	0	0.68	0.15 [0.14, 0.17]	0.09	0.01
Ireland: Neuroticism							
Scalar	33	175.0	0	0.94	0.092 [0.079, 0.11]	0	0.001

Model	*df*	χ^2	*p*	*CFI*	*RMSEA* [95%-CI]	Δ*CFI*	Δ*RMSEA*
Malaysia: Openness							
Scalar	33	290.4	0	0.66	0.12 [0.11, 0.14]	0.08	-0.01
Partial scalar (intercepts for facets O3, O2, O4 freed)	30	250.1	0	0.70	0.12 [0.11, 0.14]	0.04	-0.01
Malaysia: Conscientiousness							
Scalar	33	177.1	0	0.93	0.092 [0.078, 0.11]	0	-0.006
Malaysia: Extraversion							
Scalar	33	313.8	0	0.83	0.13 [0.11, 0.14]	0.03	-0.01
Partial scalar (intercept for facet E3 freed)	32	286.1	0	0.85	0.13 [0.11, 0.14]	0.01	-0.01
Malaysia: Agreeableness							
Scalar	33	327.4	0	0.73	0.13 [0.12, 0.15]	0.04	-0.01
Partial scalar (intercepts for facets A2, A5 freed)	31	299.4	0	0.76	0.13 [0.12, 0.15]	0.01	-0.01
Malaysia: Neuroticism							
Scalar	33	129.7	0	0.95	0.075 [0.06, 0.09]	-0.01	-0.016
Mexico: Openness							
Scalar	33	368.3	0	0.65	0.14 [0.13, 0.16]	0.09	0.01
Partial scalar (intercepts for facets O3, O2, O1 freed)	30	287.9	0	0.73	0.13 [0.12, 0.15]	0.01	0
Mexico: Conscientiousness							
Scalar	33	190.0	0	0.93	0.096 [0.082, 0.11]	0.02	-0.002
Partial scalar (intercepts for facets C2, C6 freed)	31	175.5	0	0.92	0.095 [0.080, 0.11]	0.01	-0.003
Mexico: Extraversion							
Scalar	33	219.1	0	0.89	0.11 [0.091, 0.12]	-0.03	-0.03
Mexico: Agreeableness							
Scalar	33	289.8	0	0.77	0.12 [0.11, 0.14]	0	-0.02
Mexico: Neuroticism							
Scalar	33	160.9	0	0.93	0.087 [0.073, 0.10]	0.01	-0.004
Netherlands [a]: Openness							
Scalar	33	494.8	0	0.58	0.17 [0.15, 0.18]	0.16	0.04

Model	*df*	χ^2	*p*	*CFI*	*RMSEA* [95%-CI]	Δ*CFI*	Δ*RMSEA*
Netherlands [a]: Conscientiousness							
Scalar	33	168.9	0	0.92	0.089 [0.074, 0.10]	0.01	-0.009
Netherlands [a]: Extraversion							
Scalar	33	258.9	0	0.89	0.12 [0.10, 0.13]	-0.03	-0.02
Netherlands [a]: Agreeableness							
Scalar	33	309.9	0	0.72	0.13 [0.12, 0.14]	0.05	-0.01
Netherlands [a]: Neuroticism							
Scalar	33	192.8	0	0.93	0.097 [0.084, 0.11]	0.01	0.006
New Zealand [a]: Openness							
Scalar	33	329.1	0	0.66	0.13 [0.12, 0.15]	0.08	0
New Zealand [a]: Conscientiousness							
Scalar	33	238.3	0	0.90	0.11 [0.097, 0.071]	0.03	0.02
New Zealand [a]: Extraversion							
Scalar	33	401.9	0	0.82	0.15 [0.14, 0.16]	0.04	0.01
New Zealand [a]: Agreeableness							
Scalar	33	366.2	0	0.71	0.14 [0.13, 0.16]	0.06	0
New Zealand [a]: Neuroticism							
Scalar	33	209.6	0	0.92	0.10 [0.089, 0.12]	0.02	0.009
Norway: Openness							
Scalar	33	389.8	0	0.65	0.15 [0.13, 0.16]	0.09	0.02
Partial scalar (intercepts for facet O3 freed)	32	276.2	0	0.76	0.12 [0.11, 0.14]	-0.02	-0.01
Norway: Conscientiousness							
Scalar	33	234.4	0	0.89	0.11 [0.095, 0.12]	0.04	0.012
Partial scalar (intercepts for facets C6, C4, C1 freed)	30	193.9	0	0.91	0.10 [0.088, 0.12]	0.02	0.002
Norway: Extraversion							
Scalar	33	294.0	0	0.88	0.13 [0.11, 0.14]	-0.02	-0.01
Norway: Agreeableness							

Model	*df*	χ^2	*p*	*CFI*	*RMSEA* [95%-CI]	Δ*CFI*	Δ*RMSEA*
Scalar	33	341.3	0	0.74	0.14 [0.12, 0.15]	0.03	0
Partial scalar (intercepts for facet A5 freed)	32	306.4	0	0.77	0.13 [0.11, 0.14]	0	-0.01
Norway: Neuroticism							
Scalar	33	265.6	0	0.88	0.12 [0.10, 0.13]	0.06	0.029
Partial scalar (intercepts for facets N4, N3, N5 freed)	30	213.4	0	0.91	0.11 [0.095, 0.13]	0.03	0.019
Philippines: Openness							
Scalar	33	334.4	0	0.58	0.14 [0.12, 0.15]	0.15	0.01
Partial scalar (intercepts for facets O3, O2, O4 freed)	30	290.1	0	0.64	0.13 [0.12, 0.15]	0.1	0
Philippines: Conscientiousness							
Scalar	33	224.1	0	0.91	0.11 [0.092, 0.12]	0.02	0.012
Partial scalar (intercepts for facets C3, C4 freed)	31	203.4	0	0.92	0.10 [0.09, 0.12]	0.01	0.011
Philippines: Extraversion							
Scalar	33	322.7	0	0.82	0.13 [0.12, 0.15]	0.04	-0.01
Partial scalar (intercepts for facets E3, E5, E6 freed)	30	305.6	0	0.83	0.14 [0.12, 0.15]	0.03	0
Philippines: Agreeableness							
Scalar	33	291.9	0	0.76	0.12 [0.11, 0.14]	0.01	-0.02
Philippines: Neuroticism							
Scalar	33	170.3	0	0.92	0.089 [0.075, 0.10]	0.02	-0.002
Partial scalar (intercept for facet N3 freed)	33	137.3	0	0.94	0.079 [0.064, 0.095]	0	-0.012
Romania: Openness							
Scalar	33	461.9	0	0.57	0.16 [0.15, 0.18]	0.17	0.03
Partial scalar (intercepts for facets O3, O2, O1 freed)	30	363.8	0	0.67	0.15 [0.13, 0.16]	0.07	0.02
Romania: Conscientiousness							
Scalar	33	326.1	0	0.88	0.13 [0.12, 0.15]	0.05	0.032
Partial scalar (intercepts for facets C3, C4, C6 freed)	30	285.7	0	0.89	0.13 [0.11, 0.14]	0.04	0.032
Romania: Extraversion							
Scalar	33	349.9	0	0.83	0.14 [0.12, 0.15]	0.03	0

Model	*df*	χ^2	*p*	*CFI*	*RMSEA* [95%-CI]	Δ*CFI*	Δ*RMSEA*
Partial scalar (intercepts for facets E3, E2 freed)	31	314.9	0	0.85	0.13 [0.12, 0.15]	0.01	-0.01
Romania: Agreeableness							
Scalar	33	333.8	0	0.77	0.13 [0.12, 0.15]	0	-0.01
Romania: Neuroticism							
Scalar	33	183.7	0	0.95	0.094 [0.08, 0.11]	0.01	0.003
Singapore [a]: Openness							
Scalar	33	430.4	0	0.61	0.16 [0.14, 0.17]	0.13	0.03
Singapore: [a] Conscientiousness							
Scalar	33	168.8	0	0.92	0.089 [0.074, 0.10]	0.01	-0.009
Singapore [a]: Extraversion							
Score	33	266.7	0	0.87	0.12 [0.11, 0.13]	-0.01	-0.02
Singapore [a]: Agreeableness							
Scalar	33	413.8	0	0.65	0.15 [0.14, 0.17]	0.12	0.01
Singapore [a]: Neuroticism							
Scalar	33	252.0	0	0.88	0.11 [0.10, 0.13]	0.06	0.019
South Africa: Openness							
Scalar	33	444.9	0	0.69	0.16 [0.14, 0.17]	0.14	0.03
Partial scalar (intercepts for facets O3, O2 freed)	31	416.7	0	0.73	0.14 [0.12, 0.15]	0.01	0.01
South Africa: Conscientiousness							
Scalar	33	211.2	0	0.90	0.10 [0.089, 0.12]	0.03	0.002
Partial scalar (intercepts for facets C4, C2, C3 freed)	30	187.8	0	0.91	0.10 [0.087, 0.12]	0.02	0.002
South Africa: Extraversion							
Scalar	33	268.1	0	0.82	0.14 [0.13, 0.16]	0.04	0
Partial scalar (intercepts for facets E3, E5, E6 freed)	30	352.8	0	0.82	0.15 [0.13, 0.16]	0.04	0.01
South Africa: Agreeableness							
Scalar	33	393.6	0	0.73	0.15 [0.13, 0.16]	0.04	0.01
Partial scalar (intercept for facet A1 freed)	32	346.0	0	0.77	0.14 [0.13, 0.15]	0	0

Model	*df*	χ^2	*p*	*CFI*	*RMSEA* [95%-CI]	Δ*CFI*	Δ*RMSEA*
South Africa: Neuroticism							
Scalar	33	168.3	0	0.94	0.09 [0.076, 0.10]	0	-0.001
South Korea: Openness							
Scalar	33	279.2	0	0.74	0.12 [0.11, 0.14]	0	-0.01
South Korea: Conscientiousness							
Scalar	33	136.2	0	0.93	0.077 [0.063, 0.093]	0	-0.021
South Korea: Extraversion							
Scalar	33	274.4	0	0.85	0.12 [0.11, 0.13]	0.01	-0.02
South Korea: Agreeableness							
Scalar	33	384.2	0	0.64	0.15 [0.13, 0.16]	0.13	0.01
Partial scalar (intercepts for facets A4, A2, A5 freed)	30	343.7	0	0.68	0.14 [0.13, 0.16]	0.09	0
South Korea: Neuroticism							
Scalar	33	168.1	0	0.91	0.089 [0.075, 0.10]	0.03	-0.002
Partial scalar (intercepts for facets N4, N5, N3 freed)	30	147.1	0	0.92	0.087 [0.072, 0.10]	0.02	-0.004
Sweden: Openness							
Scalar	33	438.4	0	0.61	0.15 [0.14, 0.17]	0.13	0.02
Partial scalar (intercepts for facets O3, O2 freed)	31	267.1	0	0.77	0.12 [0.11, 0.14]	-0.03	-0.01
Sweden: Conscientiousness							
Scalar	33	234.7	0	0.87	0.11 [0.095, 0.12]	0.06	0.012
Partial scalar (intercepts for facets C6, C3, C4 freed)	30	181.1	0	0.91	0.099 [0.084, 0.11]	0.02	0.001
Sweden: Extraversion							
Scalar	33	268.4	0	0.90	0.12 [0.11, 0.13]	-0.04	-0.02
Sweden: Agreeableness							
Scalar	33	358.7	0	0.75	0.14 [0.13, 0.15]	0.02	0
Partial scalar (intercept for facet A6 freed)	32	352.0	0	0.76	0.14 [0.13, 0.16]	0.01	0
Sweden: Neuroticism							
Scalar	33	220.1	0	0.94	0.091 [0.087, 0.094]	0.04	0.019

Model	*df*	χ^2	*p*	*CFI*	*RMSEA* [95%-CI]	Δ*CFI*	Δ*RMSEA*
Partial scalar (intercepts for facets N4, N3, N6 freed)	30	176.5	0	0.92	0.098 [0.083, 0.11]	0.02	0.007
UK: Openness							
Scalar	33	363.8	0	0.67	0.14 [0.14, 0.16]	0.07	0.01
Partial scalar (intercept for facet O3 freed)	32	272.6	0	0.76	0.12 [0.11, 0.14]	-0.02	-0.01
UK: Conscientiousness							
Scalar	33	201.3	0	0.91	0.10 [0.086, 0.11]	0.02	0.002
Partial scalar (intercept for facet C4 freed)	32	178.9	0	0.92	0.095 [0.08, 0.11]	0.01	-0.003
UK: Extraversion							
Scalar	33	381.4	0	0.83	0.14 [0.13, 0.16]	0.03	0
Partial scalar (intercepts for facets E4, E3, E5)	30	355.8	0	0.85	0.15 [0.13, 0.16]	0.01	0.01
UK: Agreeableness							
Scalar	33	362.1	0	0.74	0.14 [0.13, 0.16]	0.03	0
Partial scalar (intercept for facet A5 freed)	32	327.5	0	0.76	0.14 [0.12, 0.15]	0.01	0
UK: Neuroticism							
Scalar	33	223.6	0	0.91	0.11 [0.094, 0.12]	0.03	0.019
Partial scalar (intercepts for facets N4, N3 freed)	31	183.2	0	0.93	0.099 [0.085, 0.11]	0.01	0.008
USA: Openness							
Scalar	33	374.9	0	0.64	0.14 [0.13, 0.16]	0.10	0.01
Partial scalar (intercepts for facets O3, O2 freed)	31	245.7	0	0.78	0.12 [0.10, 0.13]	-0.04	-0.01
USA: Conscientiousness							
Scalar	33	200.6	0	0.92	0.10 [0.086, 0.11]	0.01	0.002
USA: Extraversion							
Scalar	33	417.3	0	0.81	0.15 [0.14, 0.17]	0.05	0.01
Partial scalar (intercepts for E5, E3, E6 freed)	30	351.5	0	0.84	0.15 [0.13, 0.16]	0.02	0.01
USA: Agreeableness							
Scalar	33	311.4	0	0.77	0.13 [0.12, 0.14]	0	-0.01
USA: Neuroticism							

Model	*df*	χ^2	*p*	*CFI*	*RMSEA* [95%-CI]	Δ*CFI*	Δ*RMSEA*
Scalar	33	193.9	0	0.93	0.098 [0.085, 0.11]	0.01	0.007

Note. *N* = 1000 in each country. Scalar models were compared to the metric models summarised in table D1. Regarding partial models, facets are reported in the sequence in which their intercepts were freed. The differences between fit indices were calculated so that positive values indicate a decrease in model fit and negative values an increase in model fit. *CFI* = robust comparative fit index. *RMSEA* = robust root mean square error of approximation. $\Delta CFI = CFI_{configural} - CFI_{metric}$; $\Delta RMSEA = RMSEA_{metric} - RMSEA_{configural}$.. Facets: O1 = Imagination. O2 = Artistic Interests. O3 = Emotionality. O4 = Adventurousness. O5 = Intellect. C1 = Self-efficacy. C2 = Orderliness. C3 = Dutifulness. C4 = Achievement-striving. C6 = Cautiousness. E3 = Assertiveness. E4 = Activity Level. E5 = Excitement-seeking. E6 = Cheerfulness. A1 = Trust. A2 = Morality. A3 = Altruism. A4 = Cooperation. A5 = Modesty. A6 = Sympathy. N2 = Anger. N3 = Depression. N4 = Self-consciousness. N5 = Immoderation. N6 = Vulnerability.

[a] Modification indices could not be calculated in these countries. Consequently, no partial MI models were fitted.

Appendix E

Gender Differences per Country and Trait

Scoring Method	Cohen's *d* [95%CI]					*D* [95%CI]	H_2	EPV_2
	O	C	E	A	N			
				Australia				
Sum-Scores	/	-0.015 [-0.14, 0.11]	/	0.65 [0.52, 0.78]	0.42 [0.29, 0.55]	0.91 [0.76, 1.04]	0.65	0.57
Factor-Scores	/	0.018 [-0.11, 0.15]	/	0.70 [0.57, 0.83]	0.46 [0.33, 0.59]	0.94 [0.78, 1.07]	0.64	0.57
MI-Factor-Scores	/	0.051 [-0.078, 0.18]	/	0.92 [0.79, 1.06]	0.51 [0.38, 0.64]	1.17 [1.01, 1.30]	0.70	0.53
				Canada				
Sum-Scores	0.11 [-0.022, 0.23]	0.17 [0.039, 0.29]	/	/	0.40 [0.27, 0.52]	0.59 [0.45, 0.71]	0.64	0.57
Factor-Scores	0.12 [-0.011, 0.24]	0.14 [0.016, 0.27]	/	/	0.47 [0.34, 0.29]	0.65 [0.51, 0.78]	0.72	0.52
MI-Factor-Scores	-0.20 [-0.33, -0.073]	0.17 [0.044, 0.30]	/	/	0.54 [0.41, 0.67]	0.70 [0.56, 0.82]	0.76	0.49
				China				
Sum-Scores	0.16 [0.034, 0.29]	-0.063 [-0.19, 0.063]	-0.058 [-0.18, 0.068]	0.25 [0.13, 0.38]	0.13 [0.001, 0.25]	0.39 [0.24, 0.50]	0.63	0.50
Factor-Scores	0.14 [0.011, 0.26]	-0.066 [-0.19, 0.061]	-0.037 [-0.16, 0.090]	0.14 [0.012, 0.27]	0.16 [0.034, 0.29]	0.31 [0.17, 0.41]	0.53	0.58
MI-Factor-Scores	0.44 [0.32, 0.57]	-0.080 [-0.21, 0.046]	-0.047 [-0.17, 0.079]	0.78 [0.65, 0.91]	0.26 [0.13, 0.39]	1.06 [0.89, 1.19]	0.75	0.40
				Finland				
Sum-Scores	0.23 [0.11, 0.36]	/	0.24 [0.12, 0.37]	0.27 [0.14, 0.39]	0.42 [0.29, 0.54]	0.77 [0.64, 0.89]	0.58	0.57
Factor-Scores	0.19 [0.062, 0.311]	/	0.27 [0.15, 0.39]	0.29 [0.16, 0.41]	0.40 [0.27, 0.53]	0.72 [0.58, 0.84]	0.60	0.55

Scoring Method	Cohen's *d* [95%CI]					*D* [95%CI]	*H₂*	*EPV₂*
	O	C	E	A	N			
MI-Factor-Scores	0.086 [-0.039, 0.21]	/	0.30 [0.18, 0.43]	0.37 [0.24, 0.49]	0.56 [0.33, 0.58]	0.84 [0.69, 0.95]	0.57	0.40
			France					
Sum-Scores	/	/	0.24 [0.12, 0.37]	0.52 [0.39, 0.65]	0.38 [0.26, 0.51]	0.87 [0.74, 0.99]	0.24	0.84
Factor-Scores	/	/	0.28 [0.15, 0.40]	0.55 [0.42, 0.67]	0.47 [0.35, 0.59]	0.90 [0.77, 1.03]	0.28	0.81
MI-Factor-Scores	/	/	0.32 [0.19, 0.44]	0.71 [0.58, 0.84]	0.67 [0.54,0.79]	1.22 [1.07, 1.34]	0.35	0.77
			Germany					
Sum-Scores	0.20 [0.073, 0.32]	/	0.15 [0.028, 0.28]	/	0.37 [0.25, 0.50]	0.58 [0.44, 0.70]	0.58	0.61
Factor-Scores	0.17 [0.050, 0.30]	/	0.19 [0.068, 0.32]	/	0.47 [0.34, 0.59]	0.73 [0.60, 0.85]	0.65	0.57
MI-Factor-Scores	-0.12 [-0.24, 0.008]	/	0.21 [0.086, 0.33]	/	0.58 [0.45, 0.71]	0.85 [0.71, 0.97]	0.76	0.49
			India					
Sum-Scores	0.24 [0.11, 0.37]	0.004 [-0.13, 0.13]	0.11 [-0.016, 0.24]	0.35 [0.22, 0.48]	0.27 [0.14, 0.40]	0.61 [0.46, 0.72]	0.58	0.54
Factor-Scores	0.22 [0.090, 0.35]	-0.014 [-0.14, 0.11]	0.15 [0.025, 0.28]	0.32 [0.19, 0.45]	0.31 [0.18, 0.44]	0.59 [0.45, 0.71]	0.55	0.56
MI-Factor-Scores	0.043 [-0.086, 0.17]	0.052 [-0.077, 0.18]	0.21 [0.085, 0.34]	0.39 [0.26, 0.52]	0.35 [0.22, 0.48]	0.68 [0.54, 0.80]	0.64	0.49
			Ireland					
Sum-Scores	0.14 [0.013, 0.26]	-0.039 [-0.16, 0.087]	/	/	0.44 [0.31, 0.56]	0.52 [0.38, 0.64]	0.84	0.44
Factor-Scores	0.10 [-0.022, 0.23]	-0.017 [-0.14, 0.11]	/	/	0.48 [0.35, 0.61]	0.56 [0.44, 0.68]	0.68	0.92

Scoring Method	Cohen's *d* [95%CI]					*D* [95%CI]	H_2	EPV_2
	O	C	E	A	N			
MI-Factor-Scores	-0.16 [-0.29, -0.036]	-0.017 [-0.14, 0.11]	/	/	0.54 [0.41, 0.66]	0.61 [0.48, 0.74]	0.92	0.39
Malaysia								
Sum-Scores	/	0.068 [-0.061, 0.20]	0.001 [-0.13, 0.13]	0.27 [0.14, 0.40]	0.27 [0.14, 0.40]	0.52 [0.37, 0.63]	0.68	0.49
Factor-Scores	/	0.035 [-0.094, 0.16]	0.050 [-0.079, 0.18]	0.22 [0.089, 0.35]	0.30 [0.17, 0.43]	0.47 [0.34, 0.59]	0.75	0.44
MI-Factor-Scores	/	0.044 [-0.085, 0.17]	0.099 [-0.030, 0.23]	0.19 [0.057, 0.32]	0.33 [0.20, 0.46]	0.52 [0.38, 0.64]	0.73	0.45
Mexico								
Sum-Scores	0.17 [0.045, 0.030]	-0.16 [-0.29, -0.040]	-0.016 [-0.14, 0.11]	0.36 [0.23, 0.48]	0.44 [0.31, 0.57]	0.71 [0.57, 0.83]	0.78	0.38
Factor-Scores	0.13 [0.0008, 0.25]	-0.13 [-0.25, -0.001]	0.003 [-0.12, 0.13]	0.41 [0.28, 0.54]	0.51 [0.39, 0.64]	0.77 [0.62, 0.89]	0.81	0.35
MI-Factor-Scores	-0.10 [-0.23, 0.022]	-0.12 [-0.24, 0.008]	0.004 [-0.12, 0.13]	0.54 [0.41, 0.66]	0.59 [0.46, 0.72]	0.97 [0.82, 1.09]	0.76	0.39
Netherlands								
Sum-Scores	/	0.22 [0.093, 0.34]	0.20 [0.072, 0.32]	/	0.35 [0.22, 0.47]	0.70 [0.57, 0.82]	0.41	0.73
Factor-Scores	/	0.21 [0.085, 0.33]	0.24 [0.12, 0.37]	/	0.45 [0.33, 0.58]	0.83 [0.70, 0.96]	0.54	0.64
MI-Factor-Scores	/	0.25 [0.13, 0.38]	0.28 [0.15, 0.40]	/	0.51 [0.39, 0.64]	0.95 [0.82, 1.07]	0.52	0.66
Norway								
Sum-Scores	0.26 [0.14, 0.39]	/	0.28 [0.15, 0.40]	0.54 [0.41, 0.66]	/	0.61 [0.47, 0.73]	0.65	0.57
Factor-Scores	0.24 [0.12, 0.37]	/	0.31 [0.18, 0.44]	0.58 [0.46, 0.71]	/	0.61 [0.47, 0.72]	0.83	0.44

Scoring Method	Cohen's *d* [95%CI]					*D* [95%CI]	H_2	EPV_2
	O	C	E	A	N			
MI-Factor-Scores	0.14 [0.016, 0.27]	/	0.35 [0.22, 0.47]	0.70 [0.57, 0.83]	/	0.74 [0.60, 0.85]	0.80	0.46
Philippines								
Sum-Scores	/	0.007 [-0.12, 0.14]	/	0.44 [0.31, 0.57]	0.33 [0.20, 0.46]	0.74 [0.60, 0.86]	0.56	0.63
Factor-Scores	/	0.007 [-0.12, 0.14]	/	0.39 [0.26, 0.52]	0.41 [0.28, 0.54]	0.72 [0.58, 0.84]	0.55	0.63
MI-Factor-Scores	/	-0.067 [-0.20, 0.063]	/	0.50 [0.37, 0.63]	0.55 [0.42, 0.68]	0.93 [0.78, 1.05]	0.55	0.64
Romania								
Sum-Scores	/	/	-0.041 [-0.17, 0.083]	0.39 [0.26, 0.52]	0.42 [0.30, 0.55]	0.74 [0.61, 0.87]	0.55	0.63
Factor-Scores	/	/	-0.028 [-0.15, 0.097]	0.46 [0.34, 0.59]	0.49 [0.36, 0.62]	0.81 [0.67, 0.93]	0.58	0.61
MI-Factor-Scores	/	/	0.041 [-0.084, 0.17]	0.59 [0.46, 0.71]	0.53 [0.40, 0.66]	0.96 [0.83, 1.10]	0.54	0.64
Singapore								
Sum-Scores	/	-0.25 [-0.37, -0.12]	-0.089 [-0.21, 0.036]	/	/	0.25 [0.11, 0.36]	0.88	0.56
Factor-Scores	/	-0.23 [-0.35, -0.10]	-0.046 [-0.17, 0.078]	/	/	0.23 [0.10, 0.34]	0.94	0.53
MI-Factor-Scores	/	-0.27 [-0.40, -0.15]	-0.052 [-0.18, 0.073]	/	/	0.27 [0.14, 0.39]	0.94	0.53
South Africa								
Sum-Scores	0.22 [0.096, 0.35]	/	/	0.47 [0.35, 0.60]	0.24 [0.11, 0.36]	0.64 [0.50, 0.75]	0.54	0.64
Factor-Scores	0.23 [0.10, 0.35]	/	/	0.52 [0.39, 0.64]	0.29 [0.16, 0.42]	0.67 [0.54, 0.79]	0.55	0.63

Scoring Method	Cohen's *d* [95%CI]					*D* [95%CI]	H_2	EPV_2
	O	C	E	A	N			
MI-Factor-Scores	0.031 [-0.094, 0.16]	/	/	0.71 [0.58, 0.84]	0.33 [0.20, 0.46]	0.87 [0.72, 0.98]	0.78	0.48
			South Korea					
Sum-Scores	0.16 [0.040, 0.29]	0.004 [-0.12, 0.13]	-0.003 [-0.13, 0.12]	/	/	0.17 [0.052, 0.27]	1.00	0.34
Factor-Scores	0.14 [0.017, 0.27]	0.026 [-0.10, 0.15]	0.061 [-0.064, 0.18]	/	/	0.15 [0.029, 0.23]	0.86	0.43
MI-Factor-Scores	0.21 [0.088, 0.34]	0.049 [-0.075, 0.17]	0.065 [-0.059, 0.19]	/	/	0.21 [0.077, 0.32]	0.96	0.36
			Sweden					
Sum-Scores	0.27 [0.14, 0.39]	/	0.12 [-0.006, 0.24]	0.61 [0.48, 0.73]	/	0.63 [0.49, 0.74]	0.87	0.42
Factor-Scores	0.26 [0.14, 0.39]	/	0.16 [0.036, 0.28]	0.65 [0.52, 0.78]	/	0.65 [0.52, 0.77]	0.97	0.35
MI-Factor-Scores	-0.011 [-0.14, 0.11]	/	0.18 [0.051, 0.30]	0.77 [0.64, 0.90]	/	0.82 [0.69, 0.94]	0.98	0.35
			UK					
Sum-Scores	0.19 [0.063, 0.31]	0.092 [-0.033, 0.22]	0.16 [0.033, 0.28]	0.63 [0.50, 0.75]	0.27 [0.14, 0.40]	0.87 [0.73, 0.98]	0.74	0.41
Factor-Scores	0.20 [0.078, 0.33]	0.11 [-0.012, 0.24]	0.20 [0.078, 0.33]	0.68 [0.55, 0.81]	0.37 [0.24, 0.49]	0.92 [0.77, 1.04]	0.70	0.44
MI-Factor-Scores	0.13 [0.006, 0.26]	0.072 [-0.053, 0.20]	0.24 [0.12, 0.37]	0.84 [0.71, 0.97]	0.49 [0.37, 0.62]	1.15 [1.00, 1.27]	0.71	0.43
			USA					
Sum-Scores	0.10 [-0.26, 0.23]	0.11 [-0.019, 0.24]	/	0.59 [0.46, 0.72]	0.33 [0.20, 0.46]	0.83 [0.68, 0.95]	0.72	0.46
Factor-Scores	0.067 [-0.061, 0.20]	0.13 [0.004, 0.26]	/	0.64 [0.51, 0.77]	0.44 [0.31, 0.57]	0.93 [0.78, 1.05]	0.68	0.49

Scoring Method	Cohen's *d* [95%CI]					*D* [95%CI]	H_2	EPV_2
	O	C	E	A	N			
MI-Factor-Scores	-0.27 [-0.40, -0.14]	0.15 [0.024, 0.28]	/	0.82 [0.69, 0.95]	0.52 [0.39, 0.65]	1.22 [1.06, 1.35]	0.65	0.52

Note. N = 1000 for all countries. Positive values indicate that women scored higher than men on a given trait and vice versa. The confidence intervals for D were estimated via bootstrapping with 500 repitions. O = Openness. C = Conscientiousness. E = Extraversion. A = Agreeableness. N = Neuroticism. MI = Measurement Invariance. D = Mahalanobis' Distance. H_2 = Heterogeneity coefficient. EPV_2 = Equal Proportions of Variance coefficient.

Appendix F

Testing Normal Distribution of Gender Equality and Gender Differences Across Countries

Table F1

Shapiro-Wilk Tests of Normality for the Distributions of Absolute Gender Differences and Gender Equality in Single Traits and Overall Personality Across Countries

Trait	Sum-Scores		Factor-Scores		MI-FS		Gender Equality	
	W	*p*	*W*	*p*	*W*	*p*	*W*	*p*
Openness [a]	0.91	0.61	0.97	0.86	0.90	0.13	0.94	0.43
Conscientiousness [a]	0.90	0.13	0.86	0.061	0.85	0.028	0.92	0.23
Extraversion [b]	0.94	0.42	0.91	0.14	0.92	0.19	0.93	0.28
Agreeableness [b]	0.93	0.27	0.95	0.54	0.95	0.49	0.96	0.67
Neuroticism [c]	0.91	0.10	0.87	0.024	0.91	0.10	0.97	0.76
Overall Personality [d]	0.94	0.20	0.91	0.075	0.94	0.24	0.97	0.66

Note. Gender differences in single traits were estimated as the absolute Cohen's *d* between men and women. Gender differences in overall personality were estimated as the absolute Mahalanobis' Distance between men and women. Using Bonferroni's correction, the significance level ($\alpha = 0.05$) was adjusted to $\alpha_{\text{adjusted}} = 0.002$. MI-FS = Measurement-Invariance-Factor-Scores.

[a] $n = 13$

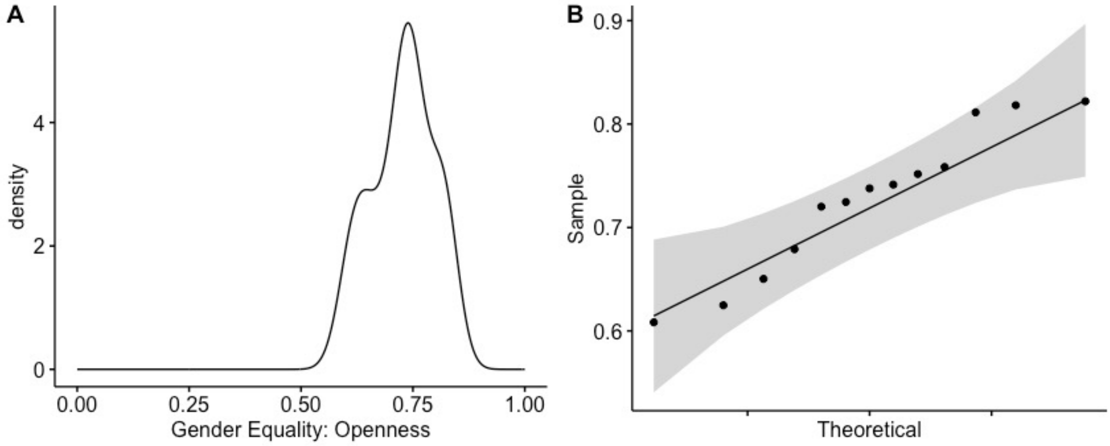
[b] $n = 14$

[c] $n = 16$

[d] $n = 20$

Figure F1

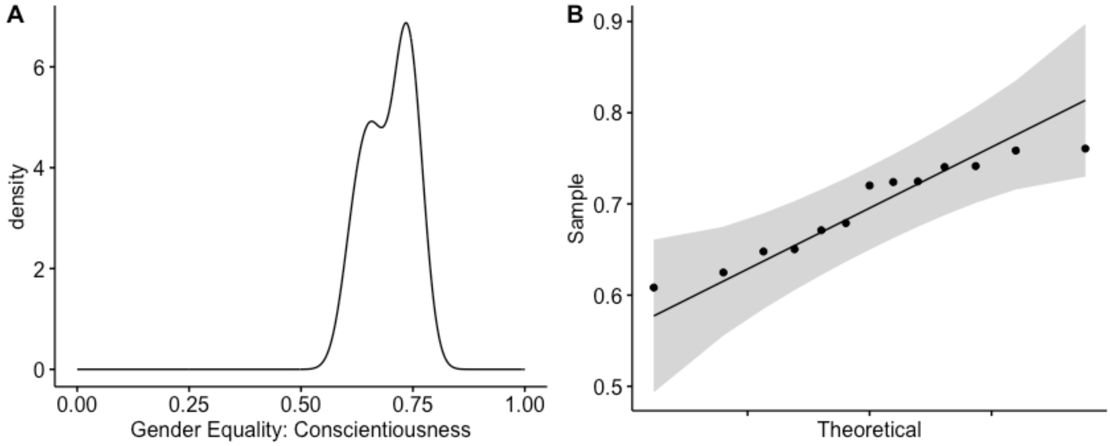
Density- and QQ-Plot of Gender Equality for the Openness-Sample



Note. N = 13 countries. Gender Equality was measured with the Gender Equality Gap Index (GGGI). Panel A shows the density distribution of the GGGI across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample.

Figure F2

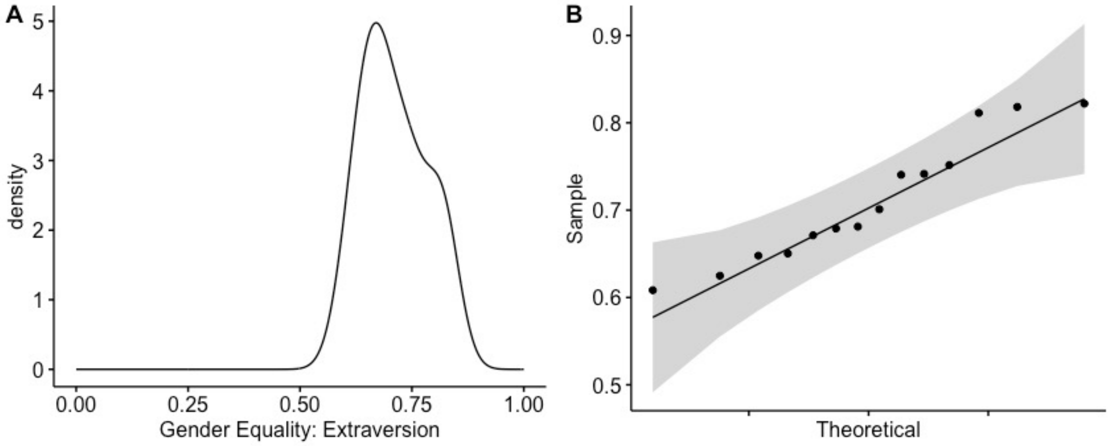
Density- and QQ-Plot of Gender Equality for the Conscientiousness-Sample



Note. N = 13 countries. Gender Equality was measured with the Gender Equality Gap Index (GGGI). Panel A shows the density distribution of the GGGI across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample.

Figure F3

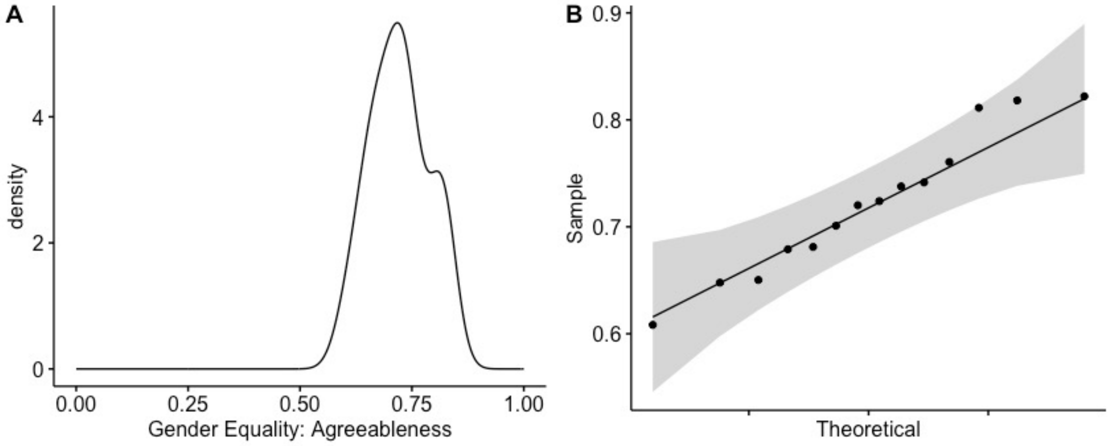
Density- and QQ-Plot of Gender Equality for the Extraversion-Sample



Note. N = 14 countries. Gender Equality was measured with the Gender Equality Gap Index (GGGI). Panel A shows the density distribution of the GGGI across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample.

Figure F4

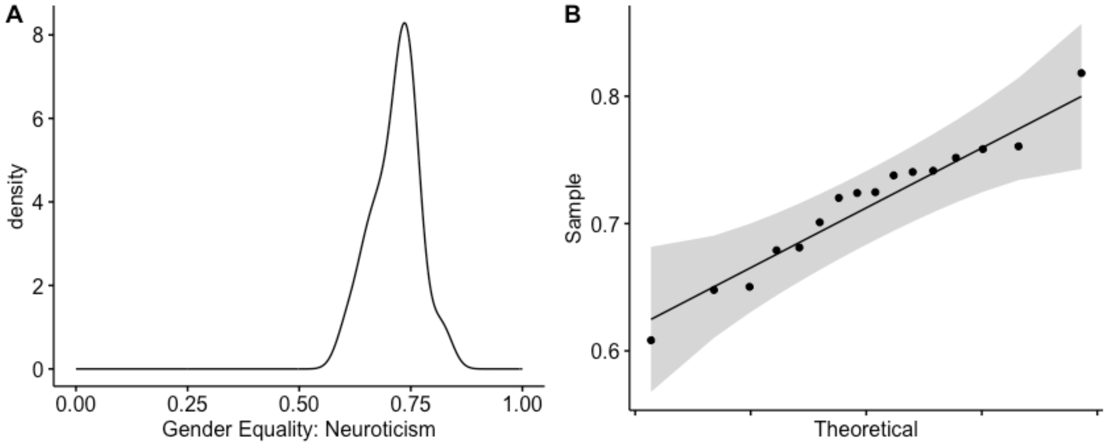
Density- and QQ-Plot of Gender Equality for the Agreeableness-Sample



Note. N = 14 countries. Gender Equality was measured with the Gender Equality Gap Index (GGGI). Panel A shows the density distribution of the GGGI across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample.

Figure F5

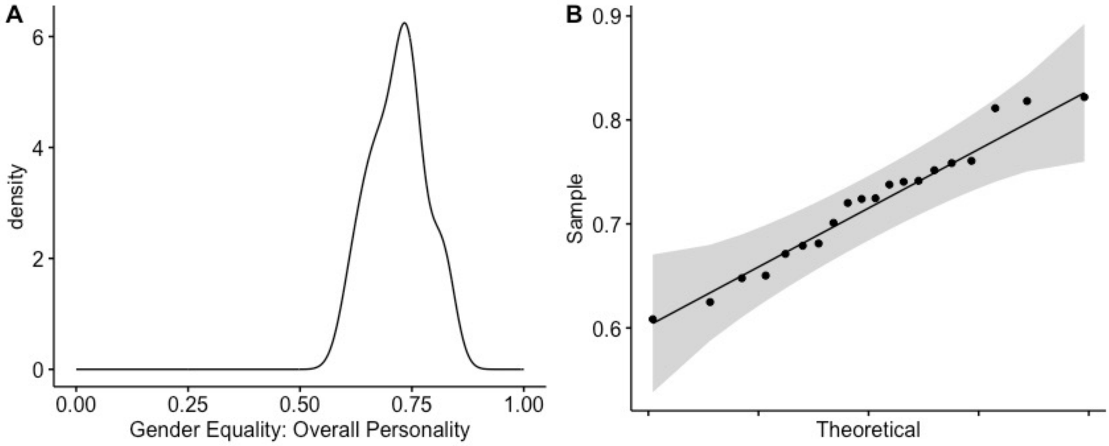
Density- and QQ-Plot of Gender Equality for the Neuroticism-Sample



Note. $N = 16$ countries. Gender Equality was measured with the Gender Equality Gap Index (GGGI). Panel A shows the density distribution of the GGGI across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample.

Figure F6

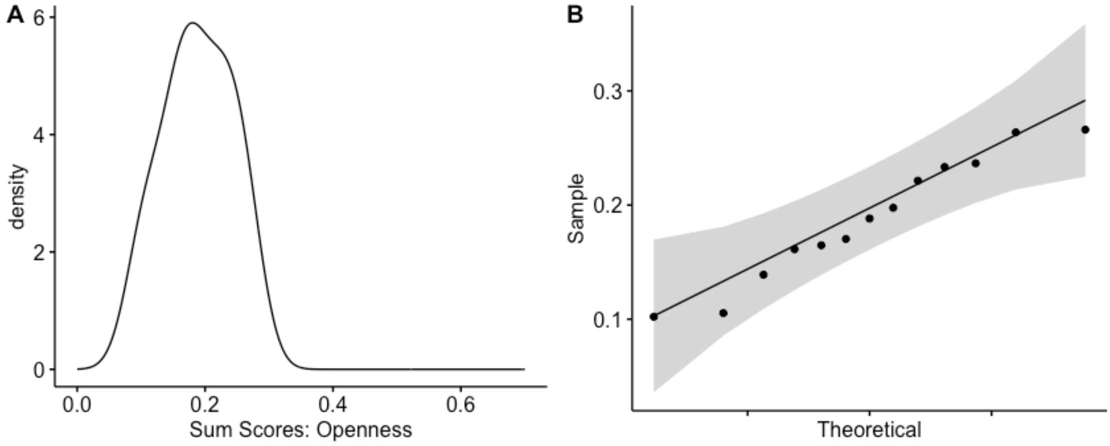
Density- and QQ-Plot of Gender Equality for the Overall-Personality-Sample



Note. $N = 20$ countries. Gender Equality was measured with the Gender Equality Gap Index (GGGI). Panel A shows the density distribution of the GGGI across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample.

Figure F7

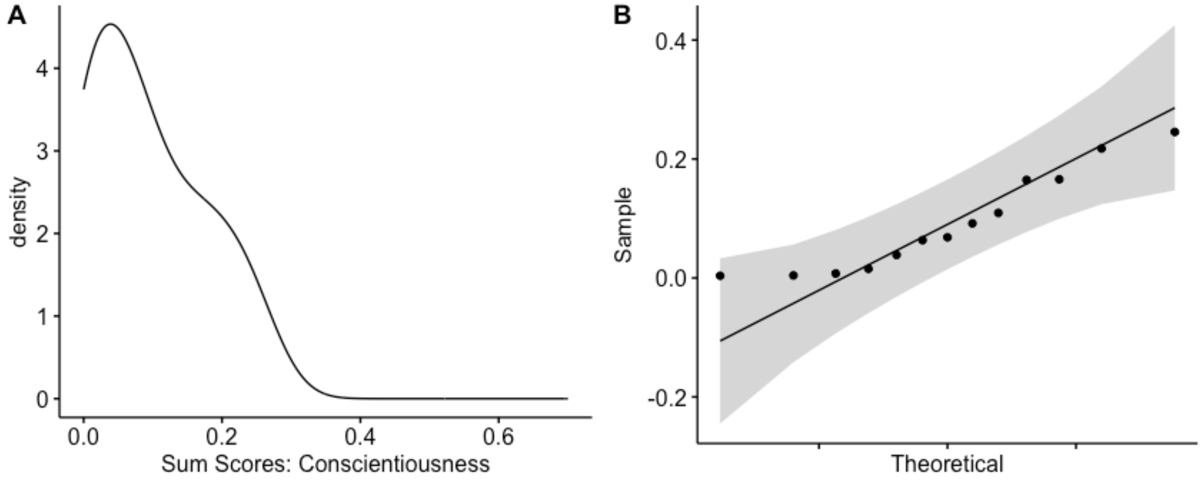
Density- and QQ-Plot of Gender Differences for Sum-Score Openness



Note. *N* = 13 countries. Gender Differences were measured as Cohen’s *d*. Panel A shows the density distribution of gender differences across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample.

Figure F8

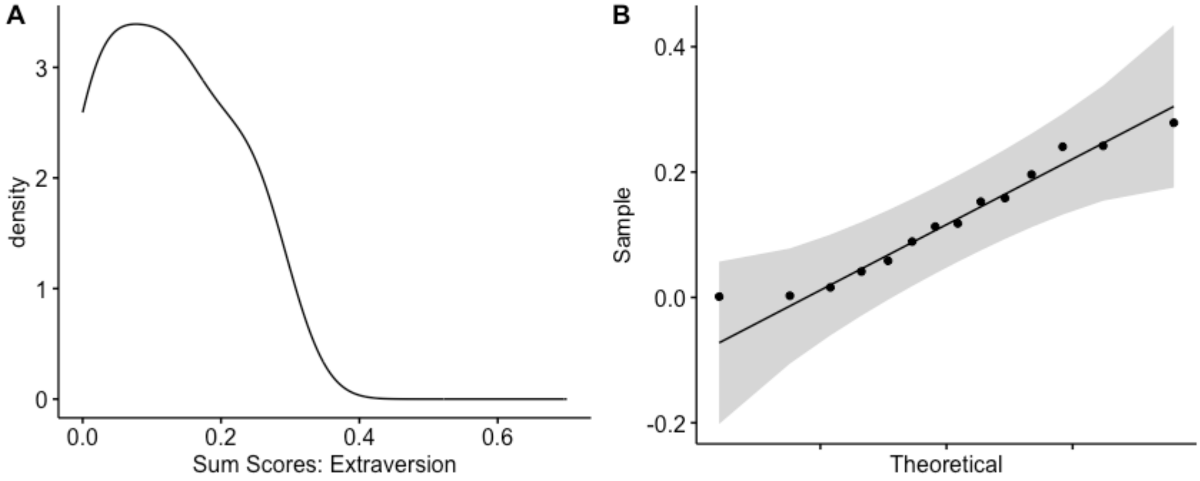
Density- and QQ-Plot of Gender Differences for Sum-Score Conscientiousness



Note. *N* = 13 countries. Gender Differences were measured as Cohen’s *d*. Panel A shows the density distribution of gender differences across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample.

Figure F9

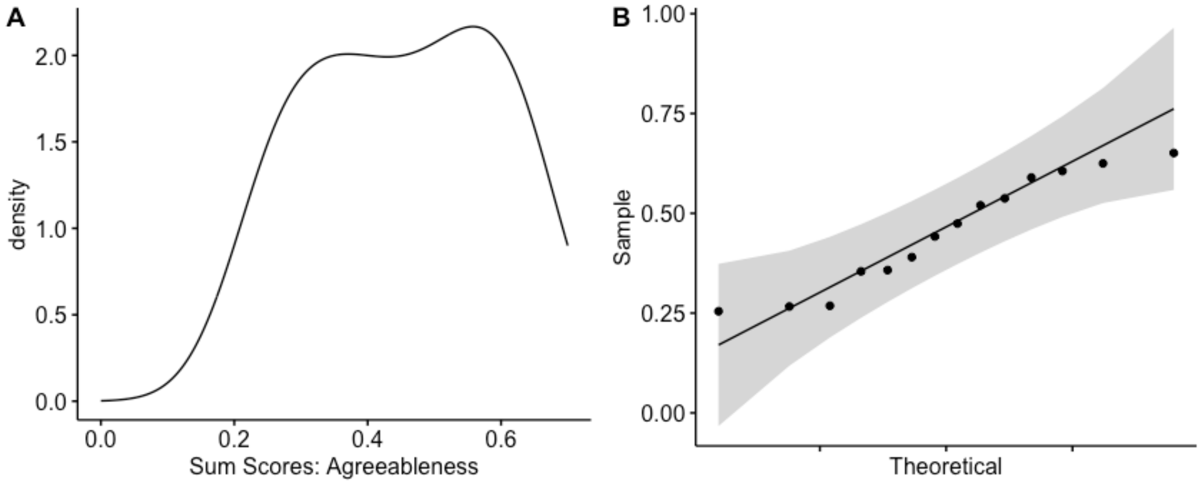
Density- and QQ-Plot of Gender Differences for Sum-Score Extraversion



Note. N = 14 countries. Gender Differences were measured as Cohen's *d*. Panel A shows the density distribution of gender differences across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample.

Figure F10

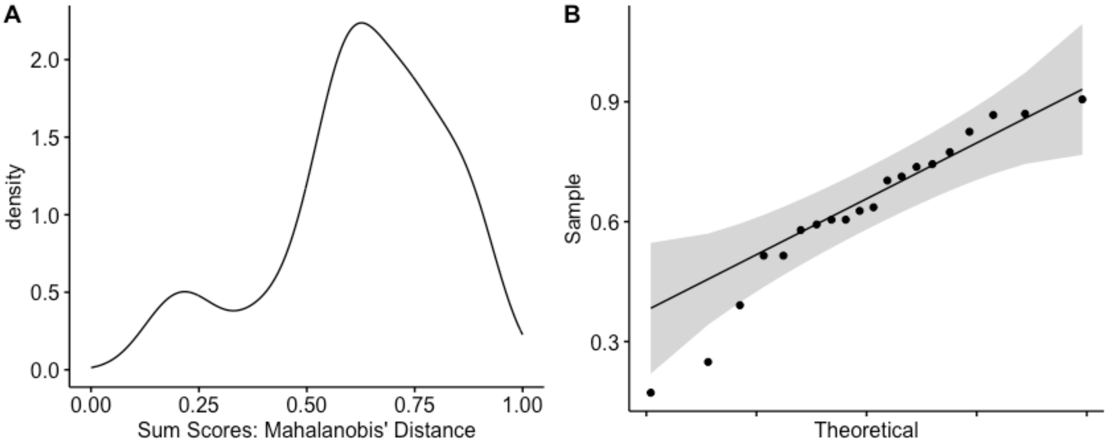
Density- and QQ-Plot of Gender Differences Sum-Score Agreeableness



Note. N = 14 countries. Gender Differences were measured as Cohen's *d*. Panel A shows the density distribution of gender differences across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample.

Figure F11

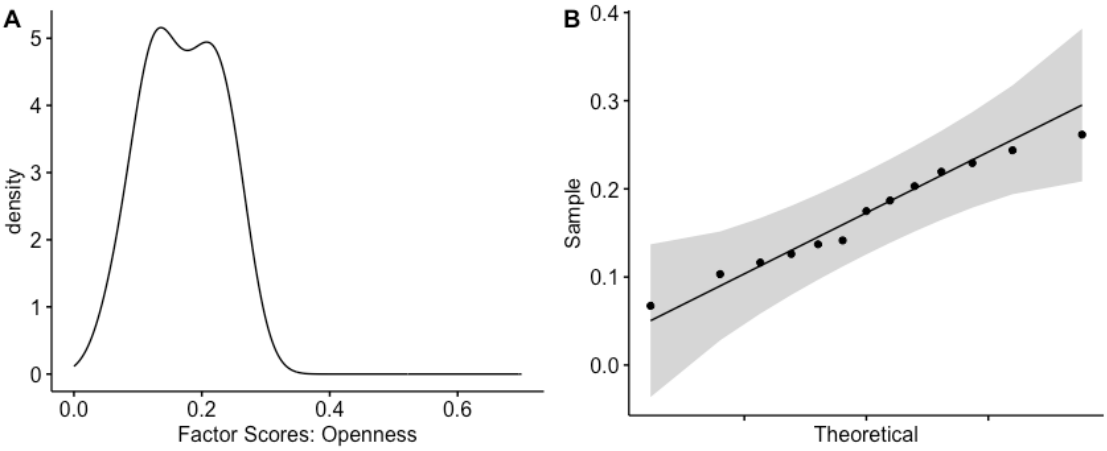
Density- and QQ-Plot of Gender Differences for Sum-Score Overall Personality



Note. *N* = 20 countries. Gender Differences in overall personality were measured as the Mahalanobis' Distance. Panel A shows the density distribution of gender differences across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample.

Figure F12

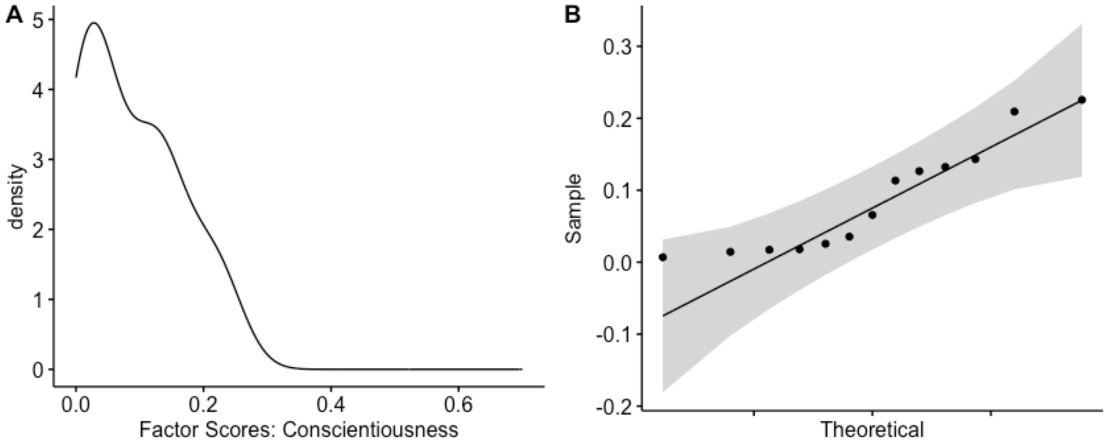
Density- and QQ-Plot of Gender Differences for Factor-Score Openness



Note. *N* = 13 countries. Gender Differences were measured as Cohen's *d*. Panel A shows the density distribution of gender differences across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample.

Figure F13

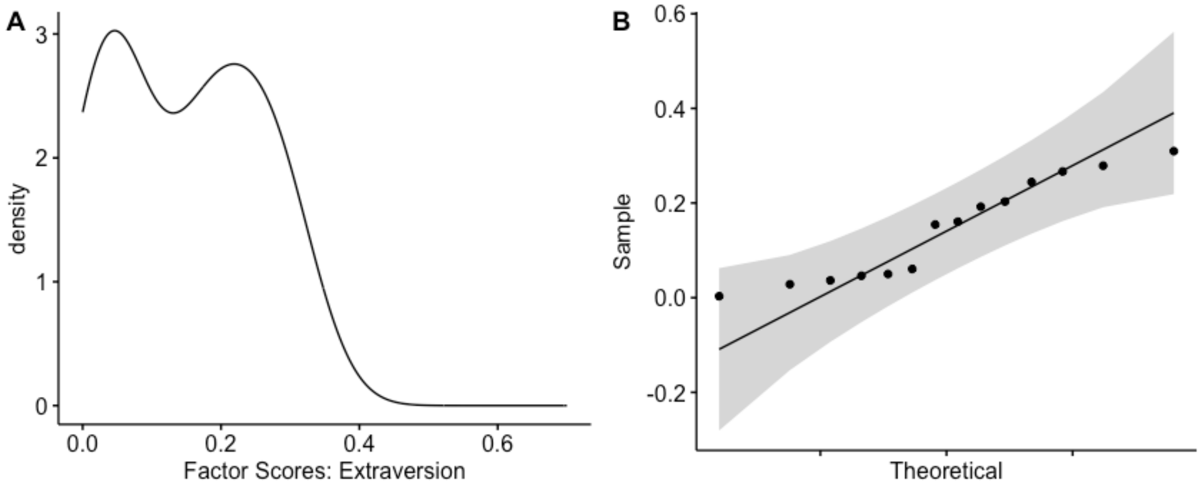
Density- and QQ-Plot of Gender Differences for Factor-Score Conscientiousness



Note. N = 13 countries. Gender Differences were measured as Cohen's *d*. Panel A shows the density distribution of gender differences across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample.

Figure F14

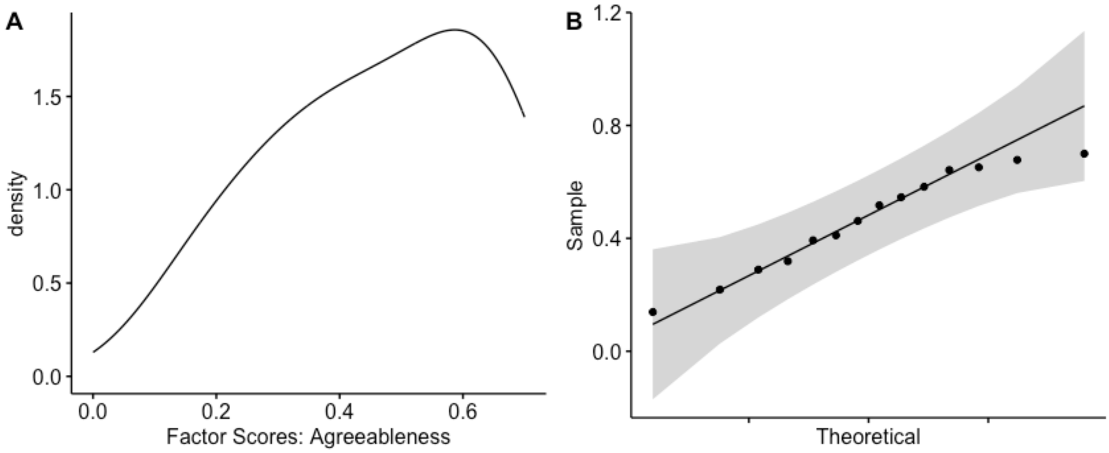
Density- and QQ-Plot of Gender Differences for Factor-Score Extraversion



Note. N = 14 countries. Gender Differences were measured as Cohen's *d*. Panel A shows the density distribution of gender differences across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample.

Figure F15

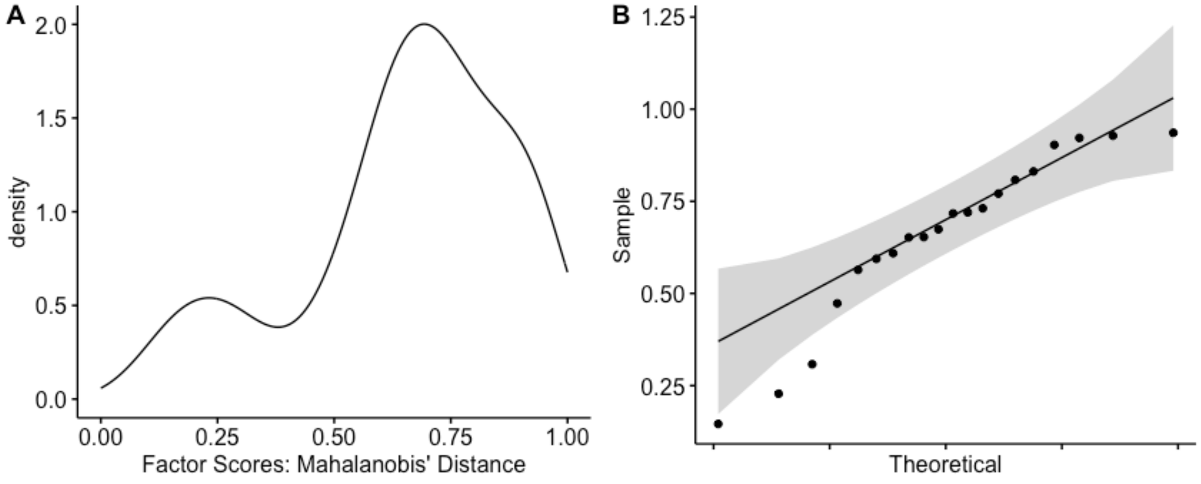
Density- and QQ-Plot of Gender Difference for Factor-Score Agreeableness



Note. *N* = 14 countries. Gender Differences were measured as Cohen's *d*. Panel A shows the density distribution of gender differences across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample.

Figure F16

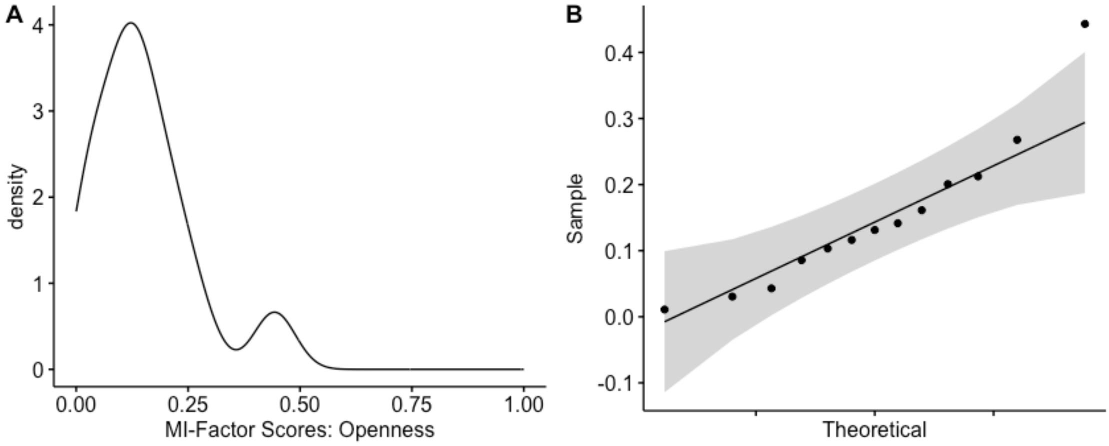
Density- and QQ-Plot of Gender Differences for Factor-Score Overall Personality



Note. *N* = 20 countries. Gender Differences in overall personality were measured as the Mahalanobis' Distance. Panel A shows the density distribution of gender differences across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample.

Figure F17

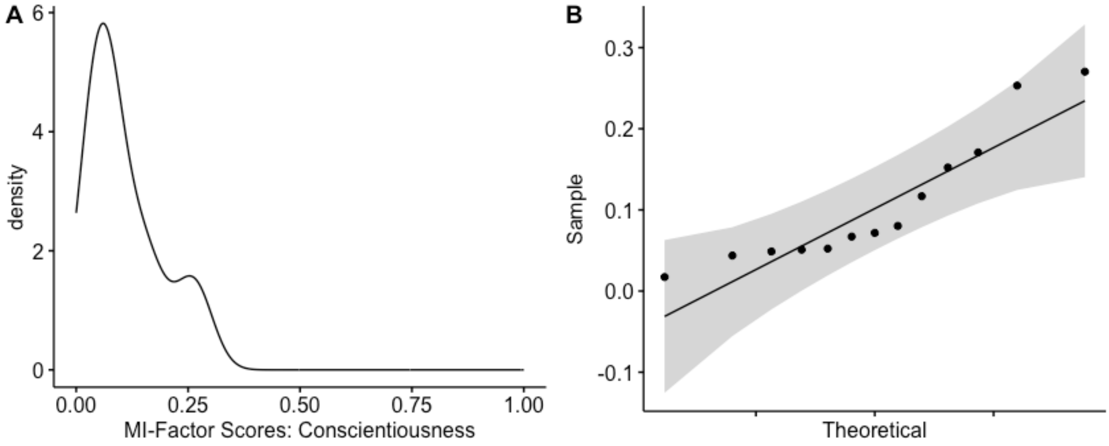
Density- and QQ-Plot of Gender Differences for MI-Factor-Score Openness



Note. N = 13 countries. Gender Differences were measured as Cohen's d. Panel A shows the density distribution of gender differences across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample. MI = Measurement invariance.

Figure F18

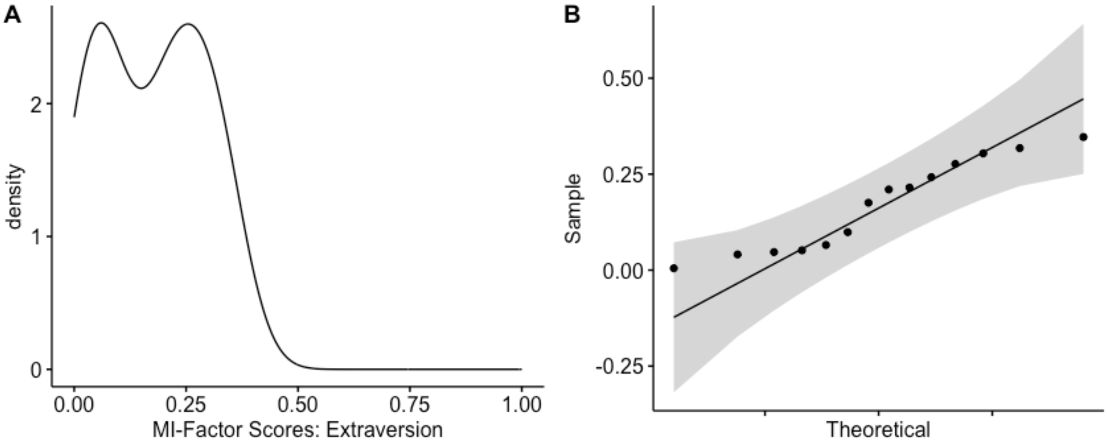
Density- and QQ-Plot of Gender Differences for MI-Factor-Score Conscientiousness



Note. N = 13 countries. Gender Differences were measured as Cohen's d. Panel A shows the density distribution of gender differences across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample. MI = Measurement invariance.

Figure F19

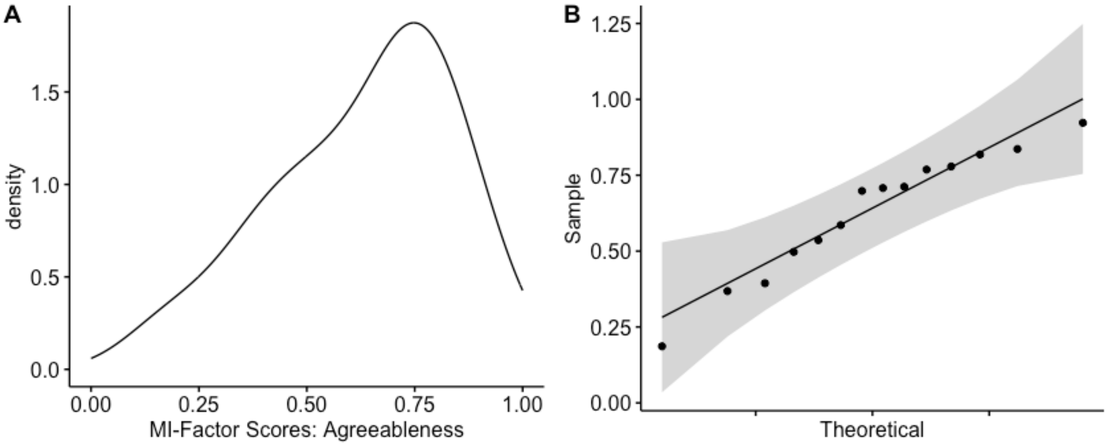
Density- and QQ-Plot of Gender Differences for MI-Factor-Score Extraversion



Note. *N* = 14 countries. Gender Differences were measured as Cohen's *d*. Panel A shows the density distribution of gender differences across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample. MI = Measurement invariance.

Figure F20

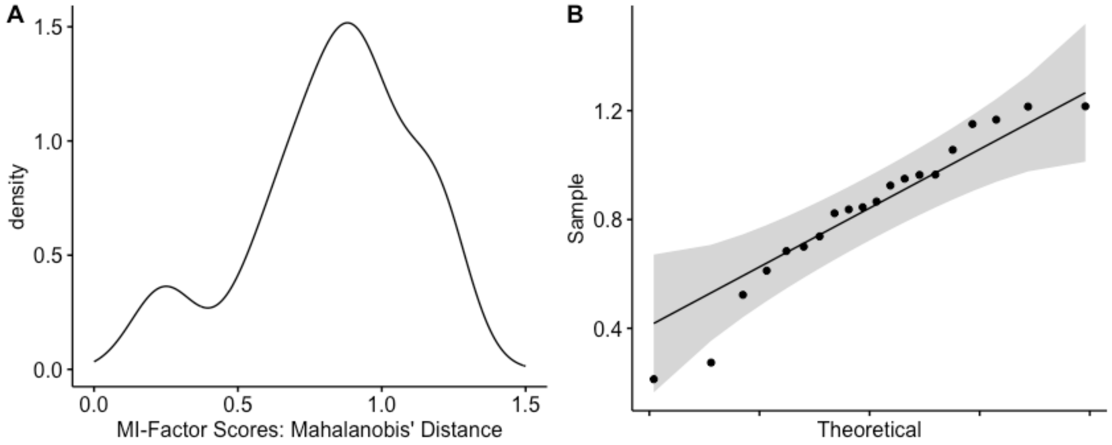
Density- and QQ-Plot of Gender Differences for MI-Factor-Score Agreeableness



Note. *N* = 14 countries. Gender Differences were measured as Cohen's *d*. Panel A shows the density distribution of gender differences across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample. MI = Measurement invariance.

Figure F21

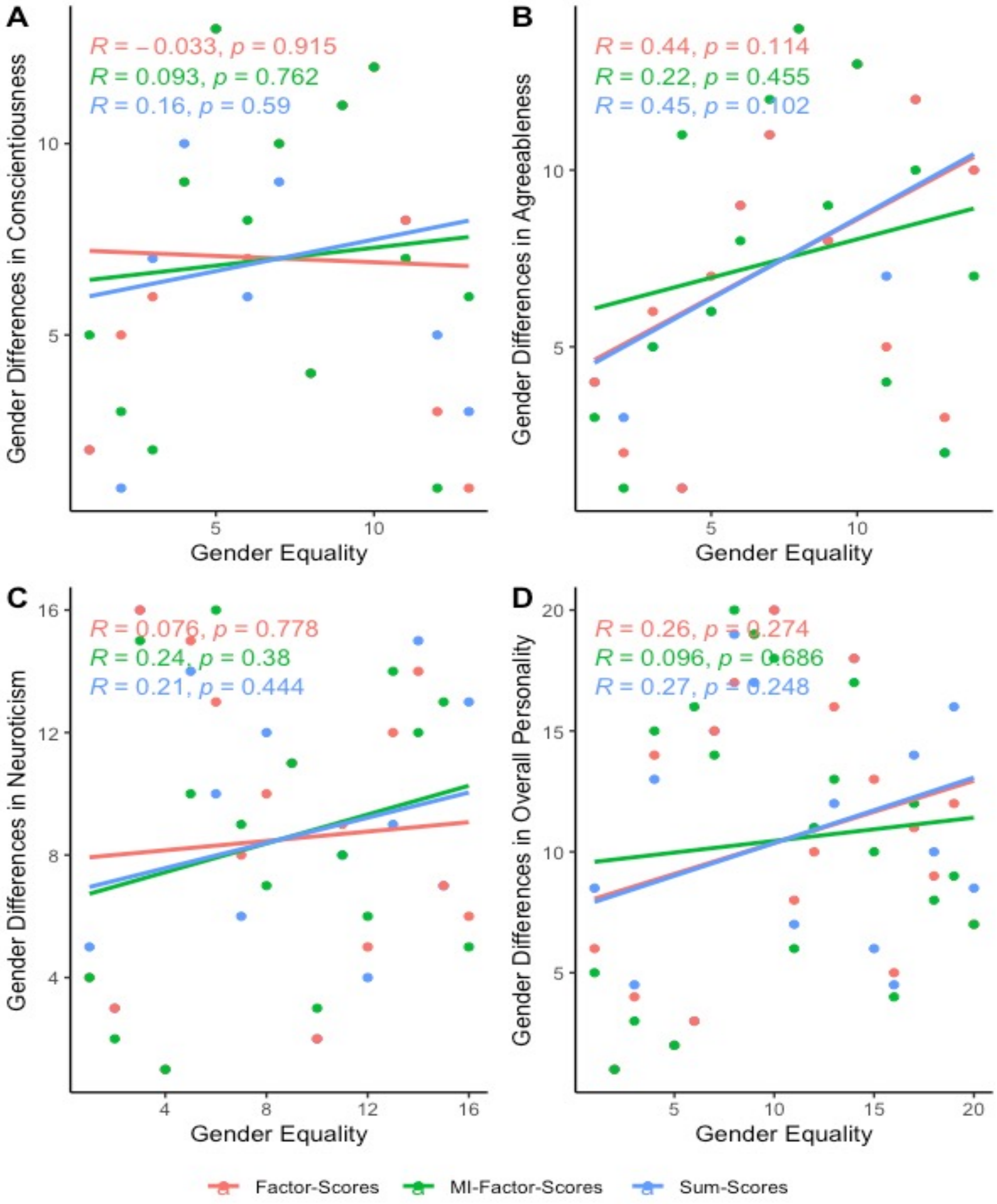
Density- and QQ-Plot of Gender Differences for MI-Factor-Score Overall Personality



Note. N = 20 countries. Gender Differences in overall personality were measured as the Mahalanobis' Distance. Panel A shows the density distribution of gender differences across countries. Panel B shows the theoretical quantiles of a normal distribution plotted against the quantiles in the actual sample. MI = Measurement invariance.

Appendix G

Scatterplots: Spearman Correlations between Gender Equality and Gender Differences in Conscientiousness, Agreeableness, Neuroticism and Overall Personality



Note. Gender differences were estimated as the absolute Cohen's d between men and women. Gender Equality was measured as the average Global Gender Gap Index between 2006 and 2011 per country. Panels A, B, C, and D show the correlation between gender equality with Conscientiousness, Agreeableness, Neuroticism, and overall personality, respectively. MI = Measurement invariance.