LUNDS UNIVERSITET

Department of Psychology

Accounting for Partial Measurement Invariance in Score Computation: Does it Matter for Diagnostic Personality Assessment in Europe?

Marie Elisabeth von Westerman

Master's Thesis

2024

Supervisor: Petri Kajonius

Abstract

A lack of full scalar measurement invariance (MI) across groups of test takers is widely assumed to render sum scores incomparable between groups. Computing factor scores from partial-scalar-MI models has been proposed as a remedy. To what extent this scoring method affects personality assessment in diagnostic contexts has yet not been investigated. The present master thesis explored this question re-analysing an example dataset with the responses of men and women from four northern European countries (n = 7504) to the Big-Five questionnaire IPIP-NEO-120. First, each of the Big-Five scales was tested for full and partial scalar MI across nationality-by-gender groups with multigroup confirmatory factor analysis. Second, the study compared Big-Five scores computed from partial-scalar-MI models with sum scores. The analyses revealed that for each Big-Five personality trait (a) full scalar MI had to be rejected, and (b) the two scoring methods yielded highly similar trait scores on average. However, for up to 6% of the sample, trait scores computed from partial-scalar-MI models did not fall within the 95%-confidence interval of respective sum scores. Depending on the reason for the personality assessment, these differences might be relevant in high-stakes situations. Practitioners should be aware of the potential effect of accounting for partial scalar MI in score computation. Future research could (a) try to replicate the results in more diverse and representative samples, and (b) examine which scoring method yields in fact the least biased trait scores.

Keywords: Personality Measures, Five Factor Personality Model, Measurement Invariance, Test Scores

Acknowledgements

This thesis would not have been possible without the help of others. I am extremely grateful to Petri Kajonius for his supervision throughout the thesis process. I would like to express my deepest appreciation to Martin Bäckström and Anton Andersson for introducing me to measurement invariance testing. Thanks should also go to John Johnson for his IPIP repository: Without his publicly available cross-national personality dataset, this thesis would not have been possible. Last but not least, I thank my family and friends for their emotional support.

Introduction

Personality predicts several life outcomes like subjective well-being, personality disorders or financial security (for an overview: Soto, 2019). The popularity of personality tests openly available in online or print media is therefore not surprising. Personality assessment is also common in work-related contexts, for example in personnel selection (Armoneit et al., 2020). If a measurement reflects not only personality but also irrelevant characteristics of an individual (e.g., nationality or gender), it might lead to wrong conclusions about the individual's personality. Whereas this would be annoying when personality is measured out of curiosity, in work-related contexts it could have severe consequences for an individual (e.g., being hired or not).

If the responses to a measurement instrument reflect only the construct to be measured and not test takers' belonging to a subpopulation, the measurement is said to be invariant (Meredith, 1993). It is possible that a measurement is only partially invariant, meaning that, across groups of test takers, only parts of the instrument are equally indicative of the construct to be measured (Byrne et al., 1989). Partial measurement invariance (MI) is widely assumed to lead to biased sum scores for test takers from at least some groups (e.g., Minkov et al., 2024). Hence, comparing these test takers regarding their scale scores would be unfair. For such cases, computing alternative scale scores from a partial-MI model has been suggested to yield comparable and fair scores (McNeish, 2023).

The theoretical appropriateness of such alternative scores is a matter of ongoing debate (e.g., Lai & Tse, 2024). In comparison, the practical relevance of accounting for partial MI in score computation has received less attention so far (cf. Eigenhuis et al., 2015). In particular, the effect on diagnostic evaluations of individual test results has been neglected (cf. Terwee et al., 2021).

The present master thesis set out to explore the practical relevance of accounting for partial MI in score computation for diagnostic personality assessment in Europe. If sum scores and trait scores computed from partial-MI models led to similar diagnostic evaluations of personality, applied test users (e.g., recruiters) would not have to concern themselves with computing trait scores from partial-MI models. Instead, they could stick to the easier to calculate and communicate sum scores. However, they should be aware of the potential bias in sum scores if full MI does not hold.

Diagnostic Personality Assessment Across Subpopulations

Personality refers to individuals' behavioural and affective dispositions that are relatively stable over time and consistent over situations (Mottus et al., 2017). These dispositions are open to descriptions at different levels of detail. So far, the measurement of personality has predominately focused on a few broad dimensions of personality (Mottus et al., 2020): The five-factor model of personality with five dimensions (also referred to as the Big Five; see Table 1 for descriptions) and the HEXACO model with six dimensions are prominent examples hereof.

In research, the interest in more detailed descriptions of personality has recently increased (e.g., Achaa-Amankwaa et al., 2020; for an overview: Mottus et al., 2020). Undoubtedly, reducing the information contained in questionnaire items to only five or six personality trait scores comes with a loss of information. In diagnostic contexts, however, it is necessary to preserve as much information as possible while also reducing complexity. Test results should be easy to communicate to and understand for lay persons not trained in psychometrics. Hence, personality descriptions on a few broad dimensions (e.g., the Big Five) are a good compromise between the former and scientific rigour.

As opposed to personality measurement in research, diagnostic personality assessment

Table 1

The Five Factor Model of Personality

Factor	Example facets	Example items
Openness to experience	O1: Imagination O2: Artistic interests	"Love to get lost in thought." "Do not like poetry." [a]
Conscientiousness	C1: Self-efficacy C2: Orderliness	"Excel in what I do." "Like to tidy up"
Extraversion	E1: Friendliness E2: Gregariousness	"Feel comfortable around people." "Prefer to be alone." [a]
Agreeableness	A1: Trust A2: Morality	"Distrust people." [a] "Use others for my own ends" [a]
Neuroticism	N1: Anxiety N2: Anger	"Am afraid of many things." "Am not easily annoyed." [a]

Note. Displayed are the Big Five as measured by the IPIP-NEO-120 (Johnson, 2014). Each of the five factors comprises six facets á four items. The first two facets with an accompanying example item are stated for each Big-Five trait.
[a] Reversed keyed item.

does neither aim for deeper understanding of personality as a construct, nor for mean-level comparisons between groups. Instead, in diagnostic contexts, an individual's personality is of interest. A meaningful interpretation of individual test results requires relating them to a norm sample. Only then is it possible to make judgments such as "Person A is more agreeable than the average". It is common practice to choose a norm sample that matches important characteristics (e.g., country of origin and gender) of the tested individual (American Psychological Association [APA], 2020). For example, the test results of a Swedish woman would be compared against a sample of Swedish women. Consequently, the judgment "Person A is more agreeable than the average Swedish woman" would be more correct than just saying "Person A is more agreeable than the average".

However, not always do test users have access to suitable norm samples. Sometimes, depending on the occasion for which personality is measured, it does not make sense to use norm samples matching the characteristics of a test taker. This is the case when individuals from different subpopulations are going to be compared. Such a scenario can happen, for example, in personnel selection. Take for instance employers, such as institutions of the European Union, selecting future employees from a multi-national applicant pool. So, comparing the applicants with each other is of interest. In such cases, research associations involved in psychometric testing and diagnostics recommend checking for MI (APA, 2020; Society for Industrial Organizational Psychology [SIOP], 2018).

Measurement Invariance in Diagnostic Contexts

The concept of MI builds on the notion that latent variables (i.e., hypothetical and not directly observable constructs; e.g., the personality trait extraversion) can be measured via manifest variables (i.e., directly observable variables; e.g., questionnaire items describing concrete actions or emotions related to extraversion). Broadly speaking, a measurement is invariant if test takers' scores on the manifest variables (e.g., responses to questionnaire items) depend only on the latent variable to be measured (Meredith, 1993). Take, for example, test takers from two different countries but equally extraverted. If the measurement of the personality trait extraversion is invariant, these test takers should exhibit the same responses to the extraversion items, irrespective of test takers home country. If the measurement is not invariant, the responses to the items would depend additionally on test takers' home country.

If a measurement is invariant across groups of test takers, it is still possible that the considered groups differ on average in their standing on the latent variable measured (i.e., group

A can be more extraverted on average than group B although the measurement is invariant across both groups). Besides, it should be noted that MI is a property of a particular measurement. Hence, it would be wrong to attribute MI to a psychometric test or a construct. Instead, MI needs to be evaluated for each measurement anew.

The latter point poses a challenge for measurements in diagnostic contexts. The sample sizes in such contexts were often too small to carry out the necessary statistical tests (SIOP, 2018). Hence, it was recommended to consult the results of prior MI testing for the inventory and particular subgroups intended to measure (APA, 2020). Such results could provide an idea whether the measurement is likely to be invariant across the subpopulations or not.

Configural, Metric, and Scalar Measurement Invariance

Measurements can be invariant on different levels of strictness. Typically, three to four increasingly strict levels of MI have been tested (Putnick & Bornstein, 2016): configural, metric, scalar, and, sometimes, strict MI. These levels correspond to measurement models with an increasing number of model parameters constrained to be equal across groups of test takers (Meredith, 1993; see Table 2). Starting with the model with the least equality constraints (viz., the configural-MI model), the different models are being compared in order of increasing constraints, with regard to model fit (Putnick & Bornstein, 2016; see Table 2). A level of MI is assumed to hold if the fit of the corresponding model is not substantially worse than the fit of the model corresponding to the preceding MI level (i.e., the MI level with equality constraints on one group of model parameter less).

Comparability of the latent variable across groups of test takers is generally assumed to require scalar MI (Lai & Tse, 2024). If scalar MI holds, the measurement exhibits, across groups, the same factor structure (i.e., which items measure which latent variable), factor loadings (indicating the strength of the relationship between items and latent variable), and intercepts of manifest variables (i.e., the baseline responses which test takers give to an item irrespective of their standing on the latent variable). The equality of intercepts is key for comparability because it allows the latent variable in the measurement model to be on the same metric across groups (Lai & Tse, 2024).

It is widely assumed that the comparability of the latent variable across groups in the measurement model (i.e., scalar MI) is necessary for comparing sum scores of individuals across groups (Minkov et al., 2024; cf. Lai & Tse, 2024). Consequently, diagnostic personality assessment across subpopulations would require that personality is measured on a scalar level.

Table 2

Four Levels of Measurement Invariance

Level of invariance	Equality constraints across groups of test takers on …	Meaning: Groups do not differ in …
Configural	---	… the factor structure.
Metric	… all factor loadings λ_i	… how strongly items are associated with the latent variable.
Scalar	… all factor loadings λ_i … all item intercepts α_i	… how strongly items are associated with the latent variable. … the difficulty of items.
Strict	… all factor loadings λ_i … all item intercepts α_i … all residual variances ε_i	… how strongly items are associated with the latent variable. … the difficulty of items. … the item-specific measurement error.

Note. Displayed are the equality constraints across groups each level of measurement invariance poses on the one-dimensional measurement model of the form $y_i = \alpha_i + \lambda_i * \eta + \varepsilon_i$. In the case of partial metric, scalar, or strict measurement invariance, these equality constraints refer not to all but only some of the items *i* of a scale. y_i = Response to item *i*. η = Latent variable to be measured. α_i = Difficulty of item *i* (the higher α_i, the more do people agree to the item irrespective of their standing on the latent variable). λ_i = Factor loading of item *i* (indicates the strength of the relationship between item *i* and latent variable η). ε_i = Item-specific measurement error.

However, in personality research, scalar MI has often not been supported across gender or across countries (for an overview: Dong & Dumas, 2020). Hence, comparing individuals of different genders and from different countries without bias is potentially hindered.

Full and Partial Measurement Invariance

Traditionally, equality constraints are introduced for all items in the measurement model (also referred to as full metric, scalar, or strict MI). If full metric, scalar, or strict MI is rejected, there is the possibility to test the fit of partial MI models (Byrne et al., 1989). In a partial MI model, equality constraints are generally retained but relaxed for single items.

Even if a measurement is only partially scalar, the items still constrained to be equal across groups of test takers allow the latent variable in the measurement model to be comparable across groups (Lai & Tse, 2024). In such a case, it is believed that, by accounting for partial scalar MI in score computation, one could obtain alternative scale scores that are, as opposed to sum scores, fair and comparable across groups (McNeish, 2023; cf. Lai & Tse, 2024).

However, partial MI testing is seen controversially since it involves a lot of arbitrary decisions on the side of the modeller and there are yet no best-practice procedures for it (Robitzsch & Lütke, 2023; Welzel et al., 2021). Most importantly, there is no consensus regarding (a) for how many items equality constraints can be relaxed but partial MI still be assumed, and (b) how to determine the items for which to lift constraints (Han et al., 2019).

Accounting for Partial-Scalar Measurement Invariance in Score Computation

There are several ways to compute individual scores on a psychometric scale. Broadly, it can be differentiated between sum scores (also referred to as observed scores [McNeish & Wolf, 2020] or non-refined scores [DiStefano et al., 2009]) and factor scores (also referred to as latent scores [McNeish & Wolf, 2020] or refined scores [DiStefano et al., 2009]).

Basically, sum scores are computed by adding up the responses to all items of a scale. Factor scores are computed from measurement models obtained, for example, from confirmatory factor analysis (CFA). A major difference between sum scores and factor scores is that the former (i.e., sum scores) are influenced by all items equally whereas the latter (i.e., factor scores) considers that items can differ in how much information on the latent variable they carry (McNeish & Wolf, 2020). The proposed alternative scale scores that account for partial-scalar MI (partial-MI scores) belong to the category of factor scores because they are computed from a partial-scalar-MI measurement model.

Theoretical Appropriateness

Intuitively, partial-MI scores could seem most appropriate in the case of partial scalar MI as they are derived from partial-scalar-MI models that achieve comparability in the latent structure between groups. However, they are subject to several theoretical debates.

First, the appropriateness of factor scores in general as compared to sum scores is a matter of ongoing debate (e.g., McNeish [2023] and McNeish & Wolf [2020] vs. Widamann & Revelle [2023] and Widamann & Revelle [2024]): From an applied diagnostics perspective, the most important argument in favour of sum scores is that they are easy to compute and to communicate. So, lay people without psychometric training too can understand their calculation. Using factor scores in diagnostic contexts would either mean fitting a measurement model to each sample anew or using a measurement model that has been cross-validated in large representative samples. The former might be hindered by the typical small samples in diagnostic situations. The latter would be rather time-consuming.

Second, the lack of consensus regarding partial-MI testing affects partial-MI scores too. Partial-MI scores are computed from partial-scalar-MI models. Hence, a lack of clarity how to build partial-MI models entails also a lack of clarity how to compute partial-MI scores. Third, MI testing evaluates the invariance of latent constructs, not of scores (Lai & Tse, 2024). Factor scores, and thus partial-MI scores too, are point estimates of an infinite number of possible latent trait scores. Thus, being estimates, factor scores cannot be equated with the true underlying latent trait. Even if a latent construct is only partially scalar invariant, this does neither necessarily imply that sum scores are biased nor that partial-MI scores are unbiased (Lai & Tse, 2024).

Practical Relevance

The question of practical relevance has seldom been addressed in MI testing (Maassen et al., 2023). This is especially true if the measurement results were used for diagnostic purposes rather than for group-level comparisons (Lai et al., 2017). Prior studies found that the usage of partial-MI scores has only a small impact on classification decisions when identifying clinical risk groups (Lai et al., 2019; Terwee et al., 2021). If MI was lacking and scores were not computed from partial-MI models, classification consistency was found to be reduced (Gonzalez et al., 2021). None of these studies investigated the practical relevance for the scoring of personality traits.

On the group-level, using partial-strict-MI scores was found to alter differences between Dutch and US-Americans on the personality questionnaire MPQ-BF-NL on four out of eleven scales (Eigenhuis et al., 2015). Similarly, scores obtained from partial-MI models led to larger group-level differences between men and women in the 16PF questionnaire than sum scores (Del Giudice et al., 2012; Kaiser et al., 2020). Hence, it stands to reason that, in these samples, the scoring method made a difference for individual trait scores.

Research Gap and Present Master Thesis

The practical relevance of accounting for partial-scalar MI in diagnostic personality assessment has yet not been examined. However, the results of such an endeavour could inform test users whether to worry about the theoretical discussions on partial MI in this diagnostic context. Furthermore, if one follows the traditional but challenged view that partial-MI scores are less biased than sum scores, such a study could reveal how strongly sum scores are biased in this context when only partial scalar MI holds. If the results of diagnostic personality

assessments should not vary between sum scores and partial-MI scores, test users could stick with the easier to compute and communicate sum scores.

The present master thesis addressed the identified research gap by re-analysing personality data from men and women from different European countries. The study had two objectives: First, the data was tested for full and partial scalar MI across nationality-by-gender groups (e.g., Swedish women, Swedish men, etc.) of test takers (objective 1). Second, the study set out to compare partial-MI scores with sum scores (objective 2). The results from objective 1, strictly speaking the partial-scalar MI models fitted to the data, were necessary for objective 2, to compute partial-MI scores. The decision to focus on a European sample was motivated by the facts that (a) no prior cross-country study on personality had reported MI testing for several European countries, and (b) personality tests are popular in personnel selection in European countries (Goodman, 2024).

Methods

Dataset

An openly available dataset[1] (Johnson, 2014) was re-analysed. The data was collected between 2001 and 2011. In total, n = 619,150 people from all over the world filled out an online personality questionnaire. The questionnaire could be found via search engines and word-of-mouth. Participants were informed that their responses would be used for research only and that careless responding would invalidate study results. Filling out the questionnaire took participants on average 20-30 minutes. After completion, participants received short written feedback on their personality. The dataset contains (a) test takers' responses to a Big-Five questionnaire on item level, (b) age, (c) gender, and (d) national affiliation of participants, as well as (e) information on date of completion of the questionnaire.

Only data from test takers meeting three inclusion criteria (see Table 3 for the criteria and their rationale) was considered in the analyses. In the resulting subsample, four nationalities were present (viz., Dutch, German, Irish, and Swedish). The eight nationality-by-gender groups of test takers to be considered in the analyses differed in size (n_{min} = 938, n_{max} = 1565). Since unequal group sizes potentially distort the results of the planned analyses (Kline, 2023), random

[1] https://osf.io/tbmh5/

Table 3

Inclusion Criteria

Criterium	Rationale
Test takers stated their national affiliation with a European country.	The raised research questions refer to nationals from European countries.
Test takers' age ranged from 19 to 69 years.	Personality is relatively stable within this age range (Bleidorn et al., 2022).
Per stated national affiliation, at least n = 900 men and n = 900 women had answered the questionnaire.	Factor loadings for Big-Five models stabilise for sample sizes of n > 1000 (Hirschfeld et al., 2014). However, this would lead to a sample containing only two nationalities. Hence, required sample size was reduced.

Note. The stated criteria were applied to the full dataset collected by John Johnson (2014). The resulting subsample was used in the present thesis.

samples were drawn from the groups to get eight groups with n = 938. The characteristics of the resulting sample can be taken from Appendix A.

Measure of Personality

The Big-Five personality traits were measured with the IPIP-NEO-120 (Johnson, 2014). The IPIP-NEO-120 is a non-proprietary adaptation of the NEO-PI-R (Costa & McCrae, 1995). With 120 items from the openly available international item pool (Goldberg et al., 2006), it measures 30 facets (four items each) and five broad personality traits (six facets each). The broad traits are openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism (see Table 1 for example facets and items). Items are responded to on a five-point Likert-scale ("very inaccurate" – "very accurate").

Prior research indicated good reliability of the IPIP-NEO-120 on trait level ($0.81 \leq \alpha \leq 0.90$) and acceptable reliability on facet level ($0.63 \leq \alpha \leq 0.88$; Johnson, 2014). The five-factor structure of the measure has been supported in a US sample (Kajonius & Johnson, 2019). Furthermore, the IPIP-NEO-120 exhibits an association with the NEO-PI-R as expected (Johnson, 2014).

Measure of National Affiliation and Gender

Prior to answering the IPIP-NEO-120, participants were asked to select the country they felt most affiliated with ("Please indicate the country to which you feel you belong the most, whether by virtue of citizenship, length of residence, or acculturation."). The so-measured national affiliation was used as the measure of nationality in the present study. Furthermore, participants stated their gender. Since the original data collection allowed only for two possibilities (viz., "male" and "female"), only these two gender identities can be considered in the present analysis.

Analytical Approach

Objective 1: Testing for (Partial) Scalar Measurement Invariance

Whether full or partial scalar MI held in the present sample was investigated via multigroup CFA for each of the five personality traits. Introduced by Jöreskog in 1971, multi-group CFA is a prominent approach to testing for MI (Putnick & Bornstein, 2016). This approach is also very common in personality research (Greiff & Scherer, 2018). I followed the conventional procedure to compare the model fit of a succession of measurement models with increasing equality constraints across groups of test takers (as described by Putnick & Bornstein, 2016; see Table 2 for the equality constraints demanded by each level of MI). The decision to reject/accept a (partial) metric- or scalar-MI model was based on conventional cut-offs for differences in Comparative Fit Index (CFI) and Standardized Root Mean-square Residual (SRMR; Chen, 2007): If model fit indices dropped by $\Delta CFI > .01$ and $\Delta SRMR > .03$ in comparison with configural-MI models, metric MI was rejected. Scalar MI was rejected if model fit indices dropped by $\Delta CFI > .01$ and $\Delta SRMR > .015$ in comparison with metric-MI models. All models were fitted with the R package lavaan (Version 0.6.17; Rosseel, 2012).

Chi-square and Root Mean Square Error of Approximation (RMSEA) were reported out of convention but not considered in the decision to reject/accept a model. Chi-square has increased type-I error rates (i.e., falsely rejecting a correctly specified model) in the face of large samples (Putnick & Bornstein, 2016). RMSEA, which builds on chi-square, is sensitive to the ratio of degrees of freedom and sample size (Kenny et al., 2015). Furthermore, RMSEA was shown to be overly strict in MI testing in the presence of ten groups and six indicators per latent variable (Rutkowski & Svetina, 2013; in the present analysis, there were six indicators per trait and eight groups).

If a full scalar-MI model was rejected, partial-scalar-MI models were built with the partialInvariance function (Pornprasertmanit, 2022) of the R package semTools (Version 0.5.6; Jorgensen et al., 2022). This open-source software was chosen to make the construction of partial-scalar-MI models as transparent and easily replicable as possible. Basically, after specifying the rejected full-scalar-MI model and the desired level of MI, the partialInvariance function gives out modification indices indicating by how much model fit would increase if a model parameter was freed.

Model parameters were freed sequentially, starting with the parameter with the highest modification index, and not all at once, since the latter procedure is associated with larger type-I error rates (Yoon & Kim, 2014). Model parameters were freed until the partial-scalar-MI model reached an acceptable model fit (i.e., met the specified cut-offs for change in CFI and SRMR in comparison with the respective metric-MI model). As long as two indicators were still constraint to be equal across groups, partial scalar MI was accepted (Byrne et al., 1989).

Objective 2: Comparing Partial-MI Scores with Sum Scores

First, sum scores and partial-MI scores were computed per Big-Five personality trait for each test taker, resulting in ten trait scores per test taker. Sum scores were computed adding up the responses to all items of the respective Big-Five scale. Partial-MI scores were extracted with Bartlett's method from the partial-scalar-MI models obtained within Objective 1. Bartlett's method uses a maximum-likelihood approach that is assumed to yield relatively unbiased estimates of the latent variable (DiStefano et al., 2009).

Second, for each personality trait, partial-MI scores were compared with sum scores. The perspective of diagnostic assessment guided the comparison. In reports of diagnostic assessments, *z*-standardised scale scores and frequentist confidence intervals (CIs) are typically interpreted (Schmidt-Atzert et al., 2021). *Z*-standardised scores allow for judging how high or low a test takers scores in comparison to a reference sample. Therefore, differences between z-standardised partial-MI scores and z-standardised sum scores were examined. Furthermore, I judged the deviation of partial-MI scores from sum scores as meaningful for diagnostic evaluations if the partial-MI score fell outside the 95%-CI of the sum score.

General Considerations

All analyses were carried out with the statistical programming language R (Version 4.3.2). Facet means instead of items were used as indicator variables in the measurement

models. Using item composites as indicators (also referred to as item parcelling) is discussed controversially in psychometrics, with lots of reasons for as well as against parcelling (for an overview: Little et al., 2013). Thanks to parcelling, it was possible to fit the measurement models with the maximum-likelihood (ML) estimator, which requires continuous indicator variables.

To account for a possible lack of multivariate normality, a robust variant of the ML estimator, based on the Satorra-Bentler scaled chi-square statistic, was used. Missing values (maximum per person: 10 of 120 items; maximum per item: 80 of 7504 test takers) were replaced by individual facet means. Though missing values do not hinder multigroup CFA, the calculation of sum scores would not have been possible with them.

Results

Objective 1: Testing for (Partial) Scalar Measurement Invariance

The results of MI testing are summarised in Table 4 (for a detailed description of model comparisons following the recommendations by Putnick and Bornstein [2016], please consult Tables B1 to B5 in Appendix B). As indicated by Table 4, scalar MI was only partially accepted for all the Big Five. This means that the same factor structure and factor loadings were

Table 4

Summary of MI Testing Results

Big-Five Scale	Level of MI		
	configural	metric	scalar
Openness	Accepted	Fully accepted	Partially accepted (4/6 intercepts freed)
Conscientiousness	Accepted	Fully accepted	Partially accepted (2/4 intercepts freed)
Extraversion	Accepted	Fully accepted	Partially accepted (3/6 intercepts freed)
Agreeableness	Accepted	Fully accepted	Partially accepted (4/6 intercepts freed)
Neuroticism	Accepted	Fully accepted	Partially accepted (3/6 intercepts freed)

Note. Which level of MI held was tested via multigroup confirmatory factor analysis. Partial invariance was established by sequentially freeing model parameters from constraints, starting with the highest modification index. MI = Measurement invariance.

supported across nationality-by-gender groups. Two to four out of six intercepts had to be freely estimated for each group to achieve partial scalar MI. Model fit of the configural-MI models was acceptable with CFI ≥ .90 and SRMR ≤ .06 for all personality traits but agreeableness (CFI = .79 and SRMR = .07; for exact model fitindices for all traits, please consult Tables B1 to B5). Since this mediocre model fit did not come as a surprise for a Big-Five scale (Hopwood & Donnellan, 2010), configural MI was accepted for all personality-trait scales, nevertheless.

Objective 2: Comparing Partial-MI Scores with Sum Scores

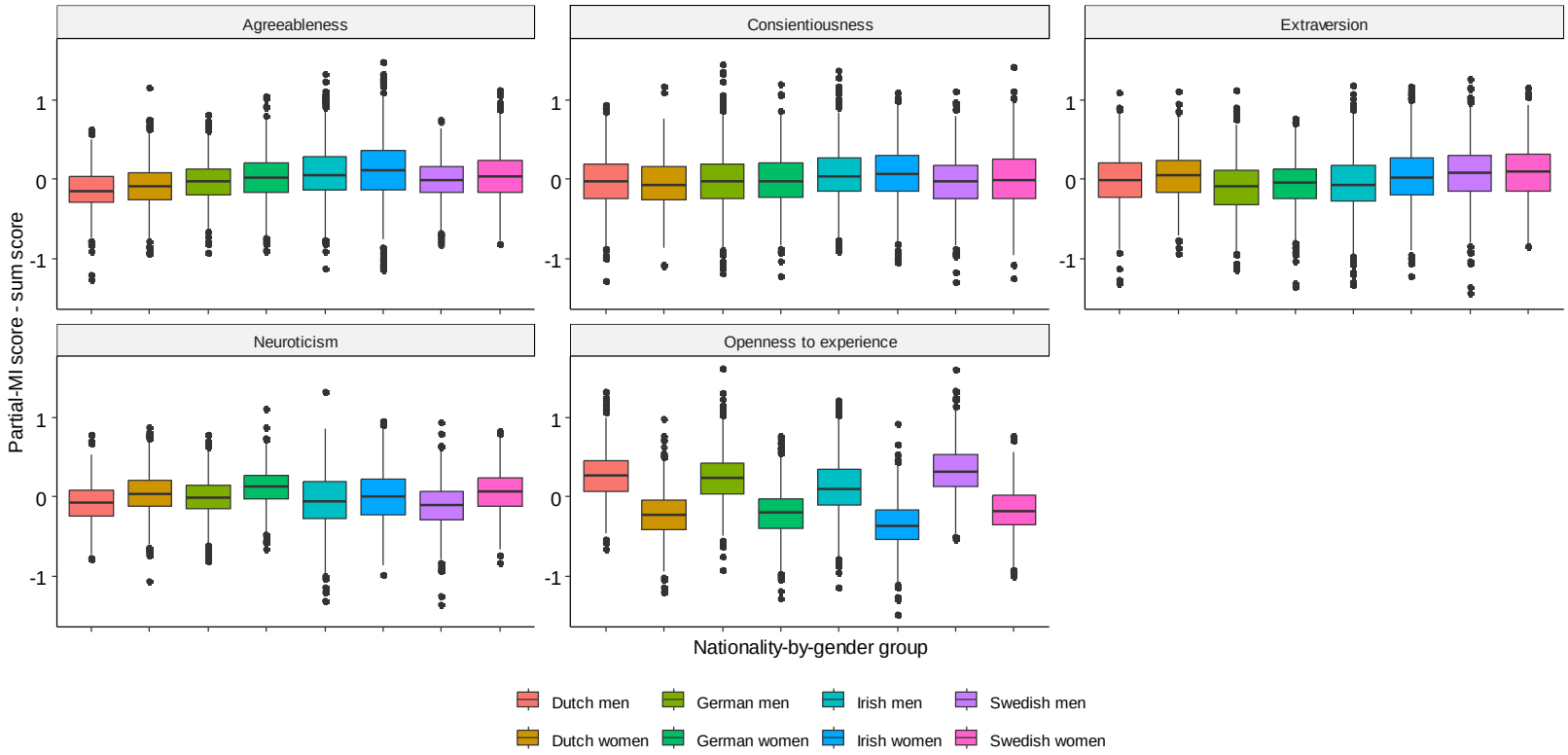
For the sample as a whole, the difference between partial-MI scores and sum scores was small. The mean difference between the two scores was close to zero regarding all the Big Five ($M_{\text{difference}} < 0.001$ for all Big-Five traits). However, for 20% of test takers, the scores differed by $\Delta = 0.4$ (in the case of neuroticism by $\Delta = 0.5$) or more, for 10% of test takers by $\Delta = 0.5$ or more (in the case of neuroticism by more than $\Delta = 0.6$). Such a difference might alter a diagnostic evaluation. Take for example an individual with a sum score of $z_{\text{sum}} = 0.0$ and a partial-MI score of $z_{\text{partial}} = 0.5$ on the neuroticism scale. The sum score would indicate that the individual is as neurotic as the average of the sample, the partial-MI score would indicate that the individual is more neurotic than 69% of test takers in the sample.

If one takes the nationality-by-gender groups into account, an interesting pattern emerges for the trait openness to experience (see Figure 1). Apparently, sum scores and partial-MI scores of openness to experience deviated systematically for some groups: Partial-MI scores tended to be higher than sum scores for men, and vice versa for women. 18 post-hoc t-tests revealed that these differences were significant on a total significance level of $\alpha = .05$ (the significance level was Bonferroni-corrected to $\alpha = .003$ for individual t-tests; for test statistics and exact p-values, please consult Appendix C). However, the average differences were not very large, with absolute mean differences between partial-MI scores and sum scores ranging between $0.1 \leq M_{\text{difference}} \leq 0.4$ for openness to experience (see Appendix C). This means that, on average, partial-MI scores and sum scores disagreed by 0.1 *SD* to 0.4 *SD* where to place test takers in relation to the whole sample.

Surprisingly, examining the proportion of test takers for whom the alternative partial-MI score falls outside the 95%-CI of the respective sum score revealed that not openness to experience, the trait with the largest mean differences per group, was the most noteworthy one in this regard (see Table 5). In fact, conscientiousness and extraversion were the traits for which

Figure 1

Boxplots – Differences Between Partial-MI Scores and Sum Scores per Big-Five Trait and Group



Note. Partial-MI scores and sum scores were *z*-standardised each. A positive difference indicates that the partial-MI score was higher than the respective sum score.

Table 5

Proportion of Test Takers with Partial-MI Scores Outside the 95%-CI of Sum Scores

(Sub)sample	Big-Five personality trait				
	A	C	E	N	O
Dutch men [b]	1.07%	**6.29%**	4.69%	1.60%	3.30%
Dutch women [b]	1.49%	**5.12%**	3.01%	2.77%	1.71%
German men [b]	0.53%	**7.36%**	**5.76%**	1.28%	3.20%
German women [b]	1.49%	**5.33%**	3.20%	1.49%	1.49%
Irish men [b]	4.69%	**7.14%**	**7.80%**	**7.57%**	1.71%
Irish women [b]	**6.50%**	**6.40%**	**7.78%**	**6.29%**	**6.72%**
Swedish men [b]	0.64%	4.80%	**7.89%**	3.94%	4.69%
Swedish women [b]	2.24%	**8.64%**	**5.76%**	1.92%	1.28%
Total [a]	2.33%	**6.38%**	**5.77%**	3.36%	3.01%

Note. Stated is the proportion of test takers whose partial-MI score was outside the 95%-CI of the sum score. **Numbers in bold** indicate groups for whom more than 5% had a partial-MI score outside the 95%-CI of the sum score. A = Agreeableness. C = Conscientiousness. E = Extraversion. N = Neuroticism. O = Openness to experience. CI = Confidence interval.

[a] $n = 7504$.

[b] $n = 938$.

partial-MI scores were beyond the sum-score-95%-CI for the most test takers (conscientiousness: 6.39% of all test takers; extraversion: 5.77% of all test takers).

Irish women were the only group for which a considerable proportion (i.e., more than 5%) of test takers (6.29% – 7.78%) had partial-MI scores outside the sum-score-95%-CI for all Big-Five traits. The largest mismatch of partial-MI scores and sum-score-95%-CIs across all groups and all traits was observed for Swedish women regarding conscientiousness: 8.64% of test takers in this group had a partial-MI score outside the sum-score-95%-CI.

Discussion

The present master thesis delt with the practical relevance of accounting for partial-scalar MI in diagnostic personality assessments. If a measurement is not fully scalar invariant, partial-MI scores (i.e., factor scores computed from partial-scalar-MI models) have been suggested to allow for a fair comparison of individuals from different subpopulations

(McNeish, 2023). Whether partial-MI scores in fact alter the results of personality assessments as compared to sum scores has not been investigated so far.

Re-analysing the responses of n = 7504 women and men from four European countries (viz., Germany, Ireland, Sweden, and the Netherlands) to the Big-Five questionnaire IPIP-NEO-120, the present study had two objectives: First, the data was tested for (partial) scalar MI across nationality-by-gender groups via multigroup CFA (objective 1). Second, I compared partial-MI scores, computed from the partial-scalar-MI models fitted within objective 1, with sum scores (objective 2).

Objective 1: Testing for (Partial) Scalar Measurement Invariance

For all Big-Five traits, only partial scalar MI was supported across the nationality-by-gender groups. Following the notion that full scalar MI is a prerequisite for sum scores being comparable across subpopulations (e.g., Minkov et al., 2024), sum scores would be expected to be biased in the present sample for at least some of the groups. The lack of full scalar MI across nationalities and across gender is in line with previous personality research (Dong & Dumas, 2020).

It should be noted that the model fit of the configural-MI models was mediocre when compared with conventional cut-offs for CFI and SRMR. Configural MI was accepted nevertheless because such mediocre model fit is typical of established personality questionnaires (Hopwood & Donnellan, 2010). This typical mediocre model fit is not surprising when considering that personality questionnaires, Big-Five questionnaires in particular, have been found to be good predictors of diverse outcome variables (for an overview regarding the Big Five, see Soto, 2019). As pointed out by Revelle (2024), predictive power and factor-structure fit of an instrument are to some degree mutually exclusive. Furthermore, the mediocre fit of the configural-MI models can probably be attributed to the sample being a convenience sample, at least partly. An anonymous online questionnaire might provoke careless responding of test takers. Careless responding could be shown to deteriorate the model fit of personality inventories (Arias et al., 2020).

Objective 2: Comparing Partial-MI Scores with Sum Scores

Considering the sample as a whole (i.e., ignoring nationality and gender), partial-MI scores and sum scores of all Big-Five traits were almost identical on average. Regarding the trait openness to experience and irrespective of nationality, partial-MI scores were

systematically higher than sum scores for men and lower than sum scores for women. The mean difference between scores for each group was however not very large. Accordingly, results of diagnostic personality assessments would, on average, not vary much with the scoring method.

Relating partial-MI scores to the 95%-CI of the respective sum scores revealed that for up to n = 479 of the total N = 7504 test takers (i.e., 6.38%), the partial-MI score did not fall within the sum-score-95%-CI. The proportion varied considerably between nationality-by-gender groups, ranging from 0.53% (n = 5) regarding agreeableness scores of German men to 8.64% (n = 81) regarding conscientiousness scores of Swedish women. Consequently, using partial-MI scores would alter the results of diagnostic personality assessments considerably for quite a number of test takers, albeit not on average.

Whether the revealed differences in trait scores are deemed negligible or not, might depend on the reason for the personality assessment. In high-stakes situations (e.g., hiring someone or not), diverging inferences for roughly 6% of test takers might seem unacceptable. For a more leisurely measurement, the difference between partial-MI scores and sum scores might be judged less relevant.

Limitations

When interpreting the results of the present master thesis, it is important to highlight what the study aimed at and what not. To start with, this thesis did not strive for proving that the personality inventory IPIP-NEO-120 is measurement (non-)invariant in European samples – this is not possible since (non-)invariance is a property of concrete measurements and should be tested for each sample anew. Second, its results do not allow any conclusions on which of the examined scoring methods yields the least biased estimates of a person's standing on the latent personality trait – this question can be answered via simulation studies only. Instead, this thesis explored, based on an example dataset, the practical relevance of choosing partial-MI scores over sum scores in diagnostic personality assessment if full scalar invariance does not hold.

Besides, the present thesis is limited in several ways. These can be summarised under four points: (a) sample characteristics, (b) the ways/instruments used to measure the constructs of interest, (c) analytical approach, and (d) relevance for prediction.

Sample Characteristics

The sample re-analysed within the present thesis limits the generalisability of results in several ways. First, the present sample is a convenience sample. This means the sample is probably not representative of the present groups of test takers. Therefore, it must be cautioned against generalising the present study's results to all men and women from Germany, Ireland, the Netherlands, and Sweden aged 19 to 69.

Second, with only four European nationalities present, it would be far-fetched to conclude that the investigated scoring methods have a similar small effect on individual trait scores in a Europe-wide sample. Conceivably, estimated measurement-model parameters deviate more strongly for nationals from other European countries not considered. This would affect the results of MI testing and might result in partial-MI scores and sum scores deviating more strongly from each other.

Measurement of Personality, Nationality, and Gender

In the re-analysed dataset, personality was measured via the IPIP-NEO-120. This is only one of the multiple Big-Five questionnaires available. Since these questionnaires differ in what and how many items are used to measure personality traits, MI testing results would probably be different from the present analysis.

The national affiliation measured in the original data collection does neither necessarily reflect legal nationality nor cultural identity being a multifaceted construct. Furthermore, gender was measured as a binary construct (viz., female or male). This does not reflect the plurality of gender identities.

Analytical Approach

Due to the study's focus on the relevance for an applied context, widely used and easily accessible modelling techniques were preferred over novel or niche techniques. However, it stands to reason that the choice of analytical approach influenced results.

Alternative approaches to testing for MI – or its item-response-theory based relative, differential item functioning – are numerous (e.g., multiple group alignment [Asparouhov & Muthén, 2014] or Bayesian approximate measurement invariance [Muthén & Asparouhov, 2012], for an overview: Leitgöb et al., 2023). Likewise, different recommendations on when to reject a metric- or scalar-MI model exist (for an overview: Han et al., 2019; Putnick & Bornstein, 2016). Furthermore, there are yet no best-practice approaches to building partial-MI

models (for an overview: Han et al., 2019; Putnick & Bornstein, 2016). With reference to measurement models in general, there are several estimation techniques for model parameters (e.g., maximum likelihood or Bayesian; Revelle, 2024) and individual factor scores (e.g., with the Bartlett method or Thurstone's regression; DiStefano et al., 2009).

Another deliberate decision was to use facet means instead of items to model the Big Five. Summarising groups of items and using these groups as indicators of latent variables in measurement models – a practice known as item parcelling – is viewed controversially. Little et al. (2013) contrasted several of the down- and upsides of item parcelling: Two important downsides are that this procedure (a) can blur the meaning of constructs, and (b) potentially masks model misspecification. On the other hand, item parcels (a) exhibit greater reliability than items, (b) reduce sources of sampling error and the number of model parameters to be estimated, and (c) allow for treating the resulting indicators of the latent variable as continuous even if items are ordinal (e.g., likert-scale items). The latter point was a decisive advantage in the present thesis. Most of the recommended model-fit cut-offs for rejecting (MI) models – also those used in the present thesis – have been developed based on the ML estimator (for an overview: Han et al., 2019). This estimator requires continuous indicator variables. Whether the cut-offs work also with estimators for categorical/ordinal indicators (e.g., diagonally weighted-least squares) is yet unclear (Han et al., 2019).

Relevance for Prediction

The present study focused on diagnostic situations in which learning about an individual's personality was of genuine interest. In such situations, the practical relevance of accounting for partial scalar MI in score computation is reflected in the difference between partial-MI scores and sum scores. In other diagnostic contexts, the results of a personality assessment might be used to predict some variable of interest (e.g., job success). In such cases, the practical relevance of using partial-MI scores instead of sum scores would lie in whether the predictions based on these scores differ.

In the present study, partial-MI scores and sum scores were very similar on average. For those test takers whose partial-MI score fell outside the sum-score-95%-CI, using partial-MI scores as predictors would potentially change the predicted outcome.

However, should trait scores be used as predictors in statistical prediction models, a set of plausible values for individual trait scores would be more appropriate than point estimates (which sum scores and partial-MI scores in this study were; Asparouhov & Muthen, 2010). This

would allow for incorporating uncertainty of estimated trait scores into the prediction model. After all, trait scores based on measurement models are mere estimations of true trait scores.

Implications for Future Research

Examining the practical relevance of psychometric debates can help to bridge the gap between scientists and practitioners. Judging practical relevance can only happen in relation to concrete contexts. Hence, more research on the effects of accounting for partial scalar MI in score computation is needed, for example in clinical contexts or personnel selection. By showing the practical relevance of methodological debates, researchers can facilitate the advent of relevant scientifically rigorous methods in applied contexts.

Future research could draw on the limitations of the present study: It would be informative to repeat the present analyses with samples more diverse in terms of culture and of gender. A focus on vulnerable subpopulations at risk of discrimination (e.g., ethnic minorities) seems especially important.

Besides, future research could take a more technical stance to the comparison of sum scores and partial-MI scores by testing the effects of different modelling techniques. For example, personality can be modelled multidimensionally or with Bayesian approaches to estimating model parameters.

The generally assumed lack of bias in partial-MI scores needs more examination (cf. Lai & Tse, 2024). This would help to conclude which scoring method, sum scores or partial-MI scores, is less biased and hence more appropriate if full scalar MI is absent.

Finally, in many diagnostic settings, personality is used for predictions. For example, in work-related contexts, recruiters want to estimate future job performance from personality (Armoneit et al., 2020). Hence, future research should also assess whether the usage of either sum scores or partial-MI scores alters such predictions. This could illuminate whether the differences between scores found in the present study are large enough to impact actual selection decisions.

Practical Implications for Diagnostic Personality Assessment

The results of the present thesis suggest that accounting for partial scalar MI in score computation can alter the outcomes of a diagnostic personality assessment. Although the average difference between partial-MI scores and sum scores was close to zero for the whole sample, for some groups of test takers and some traits, the average difference was more

pronounced. Besides, for up to n = 479 of the total N = 7504 test takers, partial-MI scores did not even fall within the 95%-CI of the sum score.

Therefore, the study's results support the recommendation by APA (2020) to select inventories for assessments across subpopulations that have already been shown to be invariant across these subpopulations. If such inventories are not available, test users should be aware of this fact and not take the results of the personality assessment as carved in stone. To be able to make informed decisions on scoring and score interpretation if full scalar MI does not hold, practitioners should concern themselves with the topic of (partial) MI.

Conclusion

The rejection of full measurement invariance on the scalar level regarding all Big-Five personality traits suggest, together with prior cross-cultural and cross-gender personality research, that the finer structure of personality might be culture and gender dependent.

The comparison of two differently calculated trait scores revealed that accounting for partial scalar measurement invariance in score computation hardly affected the results of a personality questionnaire on average – though, for some subpopulations and personality traits, there was a systematic difference. The present study is limited in several ways. Especially, it should be noted that Big-Five data of men and women from only four northern European countries were analysed.

Whether accounting for partial scalar measurement invariance yields less biased trait scores than sum scores, remains to be investigated by future simulation studies. Test users should be aware of the potential differences between scoring methods, especially for vulnerable subpopulations. To that end, test users should concern themselves with measurement invariance testing and personality trait scoring so that they can make informed decisions in diagnostic personality assessment.

References

Achaa-Amankwaa, P., Olaru, G., & Schroeders, U. (2020). Coffee or tea? Examining cross-cultural differences in personality nuances across former colonies of the British Empire. *European Journal of Personality, 35*(3), 383-397. https://doi.org/10.1177/0890207020962327

American Psychological Association, APA Task Force on Psychological Assessment and Evaluation Guidelines. (2020). *APA Guidelines for Psychological Assessment and Evaluation*. Retrieved from https://www.apa.org/about/policy/guidelines-psychological-assessment-evaluation.pdf

Arias, V. B., Garrido, L. E., Jenaro, C., Martinez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods, 52*(6), 2489-2505. https://doi.org/10.3758/s13428-020-01401-8

Armoneit, C., Schuler, H., & Hell, B. (2020). Nutzung, Validität, Praktikabilität und Akzeptanz psychologischer Personalauswahlverfahren in Deutschland 1985, 1993, 2007, 2020: Fortführung einer Trendstudie. *Zeitschrift für Arbeits- und Organisationspsychologie, 64*, 67–82.

Asparouhov, T., & Muthén, B. (2010). *Plausible values for latent variables using Mplus*. [Unpublished manuscript]. Retrieved from http://www.statmodel.com/download/Plausible. pdf.

Asparouhov, T., & Muthén, B. (2014). Multiple-Group Factor Analysis Alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

Bleidorn, W., Schwaba, T., Zheng, A., Hopwood, C. J., Sosa, S. S., Roberts, B. W., & Briley, D. A. (2022). Personality stability and change: A meta-analysis of longitudinal studies. *Psychological Bulletin, 148*(7-8), 588-619. https://doi.org/10.1037/bul0000365

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling a Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Costa Jr, P. T., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the Revised NEO Personality Inventory. *Journal of Personality Assessment, 64*(1), 21-50. https://doi.org/10.1207/s15327752jpa6401_2

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

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied Researcher. *Practical Assessment, Research & Evaluation, 14*(20), 20. https://doi.org/10.7275/da8t-4g52

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

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

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*(1), 84-96. https://doi.org/10.1016/j.jrp.2005.08.007

Gonzalez, O., Georgeson, A. R., Pelham III, W. E., & Fouladi, R. T. (2021). Estimating classification consistency of screening measures and quantifying the impact of measurement bias. *Psychological Assessment, 33*(7), 596-609. https://doi.org/10.1037/pas0000938

Goodman, S. (2024, January 31). *5 Popular Pre-Employment Tests for job Applicants*. EPALE - European Commission. https://epale.ec.europa.eu/en/blog/5-popular-pre-employment-tests-job-applicants

Greiff, S., & Scherer, R. (2018). Still comparing apples with oranges? Some thoughts on the principles and practices of measurement invariance testing. *European Journal of Psychological Assessment, 34*, 141–144. https://doi.org/10.1027/1015-5759/a000487

Han, K., Colarelli, S. M., & Weed, N. C. (2019). Methodological and statistical advances in the consideration of cultural diversity in assessment: A critical review of group

classification and measurement invariance testing. *Psychological Assessment, 31*(12), 1481-1496. https://doi.org/10.1037/pas0000731

Hirschfeld, G., Brachel, R. v., & 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

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review, 14*(3), 332-346. https://doi.org/10.1177/1088868310361240

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

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409-426. https://doi.org/10.1007/BF02291366

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

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. 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

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research, 44*(3), 486-507. https://doi.org/10.1177/0049124114543236

Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford publications.

Lai, M. H., Kwok, O. M., Yoon, M., & Hsiao, Y. Y. (2017). Understanding the impact of partial factorial invariance on selection accuracy: An R script. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(5), 783-799. https://doi.org/10.1080/10705511.2017.1318703

Lai, M. H., Richardson, G. B., & Mak, H. W. (2019). Quantifying the impact of partial measurement invariance in diagnostic research: An application to addiction research. *Addictive Behaviors, 94*, 50-56. https://doi.org/10.1016/j.addbeh.2018.11.029

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

Leitgöb, H., Seddig, D., Asparouhov, T., Behr, D., Davidov, E., De Roover, K., Jak, S., Meitinger, K., Menold, N., Muthen, B., Rudnev, M., Schmidt, P., & van de Schoot, R. (2023). Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. *Social Science Research, 110*, Article 102805. https://doi.org/10.1016/j.ssresearch.2022.102805

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

Maassen, E., D'Urso, E. D., van Assen, M. A. L. M., 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://doi.org/10.1037/met0000624

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. https://doi.org/10.3758/s13428-022-02016-x

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods, 52*(6), 2287‑2305. https://doi.org/10.3758/s13428-020-01398-0

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525-543. https://doi.org/10.1007/BF02294825

Minkov, M., Vignoles, V. L., Welzel, C., Akaliyski, P., Bond, M. H., Kaasa, A., & Smith, P. B. (2024). Comparative culturology and cross-cultural psychology: How comparing societal cultures differs from comparing individuals' minds across cultures. *Journal of Cross-Cultural Psychology*, *55*(2), 164-188. https://doi.org/10.1177/00220221231220027

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and

utility of personality nuances. *Journal of Personality and Social Psychology, 112*(3), 474-490. https://doi.org/10.1037/pspp0000100

Mõttus, R., Wood, D., Condon, D. M., Back, M. D., Baumert, A., Costantini, G., Epskamp, S., 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

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313–335. https://doi.org/10.1037/a0026802

Pornprasertmanit, S. (2022). *A note on effect size for measurement invariance*. Retrieved from https://cran.r-project.org/web/packages/semTools/vignettes/partialInvariance.pdf

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

Revelle, W. (2024). The seductive beauty of latent variable models: Or why I don't believe in the Easter Bunny. *Personality and Individual Differences, 221*, Article 112552. https://doi.org/10.1016/j.paid.2024.112552

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. (2013). 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

Schmidt-Atzert, L., Krumm, S., Amelang, M. (2021). Durchführung einer diagnostischen Untersuchung und Gutachtenerstellung. In L. Schmidt-Atzert, S. Krumm, & M. Amelang (Eds.), *Psychologische Diagnostik* (pp. 477-525). Springer. https://doi.org/10.1007/978-3-662-61643-7_4

Society for Industrial and Organizational Psychology. (2018). *Principles for the validation and use of personnel selection procedures* (5th ed.). *American Psychological Association, 11*, 1-97. https://doi.org/10.1017/iop.2018.195

Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project. *Psychological Science, 30*(5), 711-727. https://doi.org/10.1177/0956797619831612

Terwee, C. B., Crins, M. H., Roorda, L. D., Cook, K. F., Cella, D., Smits, N., & Schalet, B. D. (2021). International application of PROMIS computerized adaptive tests: US versus country-specific item parameters can be consequential for individual patient scores. *Journal of Clinical Epidemiology, 134*, 1-13. https://doi.org/10.1016/j.jclinepi.2021.01.011

Welzel, C., Brunkert, L., Kruse, S., & Inglehart, R. F. (2021). Non-invariance? An overstated problem with misconceived causes. *Sociological Methods & Research, 52*(3), 1368-1400. https://doi.org/10.1177/0049124121995521

Widaman, K. F., & Revelle, W. (2023). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*, *55*(2), 788-806. https://doi.org/10.3758/s13428-022-01849-w

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

Yoon, M., & Kim, E.S. (2014). A comparison of sequential and nonsequential specification searches in testing factorial invariance. *Behavior Research Methods, 46*, 1199-1206. https://doi.org/10.3758/s13428-013-0430-2

Appendix A

Sample Characteristics

National affiliation	Gender [a]	Age		
		Min – Max	*M*	*SD*
Dutch	Male	19 – 65	29.4	9.8
	Female	19 – 60	27.0	9.0
German	Male	19 – 65	29.4	8.6
	Female	19 – 60	27.0	7.6
Irish	Male	19 – 65	28.9	9.5
	Female	19 – 65	27.5	8.5
Swedish	Male	19 – 63	29.5	9.1
	Female	19 – 66	27.4	9.4

Note. Displayed are the characteristics of the sample that was analysed in the present thesis, a subsample of the dataset collected by John Johnson (2014). In total, $N = 7504$ test takers were included, $n = 938$ men and $n = 938$ women per national affiliation.

[a] Gender was operationalised as a dichotomous variable in the original data collection.

Appendix B

Measurement Invariance Testing

Table B1

Measurement-Invariance Testing - Agreeableness

Model	χ^2 (*df*)	CFI	RMSEA (90% CI)	SRMR	Model comp	$\Delta\chi^2$ (Δdf)	ΔCFI	ΔRMSEA	ΔSRMR	Decision
M1	1556.73 * (72)	.785	.165 (.158 - .172)	.070	--	--	--	--	--	--
M2	1652.34 * (107)	.785	.136 (.130 - .142)	.073	M1	46.27 (35)	.001	.029	.003	Accept
M3	2521.86 * (142)	.691	.141 (.136 - .146)	.096	M2	962.49 * (35)	.094	.005	.023	Reject
M3a	2283.77 * (135)	.717	.139 (.133 - .144)	.091	M2	688.16 * (28)	.068	.003	.018	Reject
M3b	2035.70 * (128)	.745	.135 (.130 - .140)	.084	M2	398.99 * (21)	.040	.001	.011	Reject
M3c	1846.51 * (121)	.766	.133 (.128 - .138)	.077	M2	188.03 * (14)	.018	.003	.005	Reject
M3d	1713.57 * (114)	.779	.133 (.128 - .139)	.076	M2	54.06 * (7)	.006	.003	.003	Accept

Note. Rejection/Acceptance was based on CFI and SRMR. Reported are robust model fit indices. $N = 7504$. Eight groups á $n = 938$. M1 = Configural-MI model. M2 = Metric-MI model. M3 = Full scalar-MI model. M3a – M3d = Partial scalar-MI models.

* $p < .001$.

Table B2

Measurement-Invariance Testing - Neuroticism

Model	χ^2 (*df*)	CFI	RMSEA (90% CI)	SRMR	Model comp	$\Delta\chi^2$ (Δdf)	ΔCFI	ΔRMSEA	ΔSRMR	Decision
M1	784.95 * (72)	.948	.110 (.103 - .117)	.044	--	--	--	--	--	--
M2	948.97 * (107)	.940	.097 (.091 - .102)	.061	M1	151.31 * (35)	.008	.013	.017	Accept
M3	1700.83 * (142)	.895	.111 (.106 - .116)	.077	M2	839.77 * (35)	.045	.015	.016	Reject
M3a	1466.72 * (135)	.909	.106 (.101 - .111)	.074	M2	569.33 * (28)	.031	.009	.012	Reject
M3b	1291.25 * (128)	.920	.102 (.097 - .107)	.072	M2	369.38 * (21)	.021	.006	.010	Reject
M3c	1122.87 * (121)	.930	.098 (.093 - .103)	.067	M2	182.64 * (14)	.01	.002	.006	Accept

Note. Rejection/Acceptance was based on CFI and SRMR. Reported are robust model fit indices. $N = 7504$. Eight groups á $n = 938$. M1 = Configural-MI model. M2 = Metric-MI model. M3 = Full scalar-MI model. M3a – M3c = Partial scalar-MI models.
* $p < .001$.

Table B3

Measurement-Invariance Testing - Conscientiousness

Model	χ^2 (*df*)	CFI	RMSEA (90% CI)	SRMR	Model comp	$\Delta\chi^2$ (Δdf)	ΔCFI	ΔRMSEA	ΔSRMR	Decision
M1	907.12 * (72)	.926	.121 (.114 - .128)	.052	--	--	--	--	--	--
M2	1021.84 * (107)	.920	.103 (.097 - .109)	.064	M1	107.98 * (35)	.006	.018	.012	Accept
M3	1587.08 * (142)	.881	.109 (.104 - .114)	.078	M2	617.49 * (35)	.039	.006	.014	Reject
M3a	1351.34 * (135)	.898	.104 (.099 - .109)	.071	M2	342.13 * (28)	.022	< .001	.007	Reject
M3b	1205.23 * (128)	.909	.101 (.096 - .106)	.067	M2	179.99 * (21)	.011	.002	.003	Accept

Note. Rejection/Acceptance was based on CFI and SRMR. Reported are robust model fit indices. $N = 7504$. Eight groups á $n = 938$. M1 = Configural-MI model. M2 = Metric-MI model. M3 = Full scalar-MI model. M3a – M3b = Partial scalar-MI models.
* $p < .001$.

Table B4

Measurement-Invariance Testing - Extraversion

Model	χ^2 (*df*)	CFI	RMSEA (90% CI)	SRMR	Model comp	$\Delta\chi^2$ (Δdf)	ΔCFI	ΔRMSEA	ΔSRMR	Decision
M1	1340.20 * (72)	.901	.147 (.140 - .154)	.057	--	--	--	--	--	--
M2	1462.19 * (107)	.896	.123 (.118 - .129)	.068	M1	102.18 * (35)	.005	.023	.011	Accept
M3	2234.75 * (142)	.849	.129 (.125 - .134)	.087	M2	840.52 * (35)	.047	.006	.019	Reject
M3a	2013.51 * (135)	.862	.126 (.122 - .131)	.081	M2	589.11 * (28)	.034	.003	.013	Reject
M3b	1830.43 * (128)	.874	.123 (.119 - .129)	.076	M2	386.22 * (21)	.022	.001	.008	Reject
M3c	1656.93 * (121)	.885	.122 (.117 - .128)	.071	M2	195.47 * (14)	.011	.001	.003	Accept

Note. Rejection/Acceptance was based on CFI and SRMR. Reported are robust model fit indices. $N = 7504$. Eight groups á $n = 938$. M1 = Configural-MI model. M2 = Metric-MI model. M3 = Full scalar-MI model. M3a – M3c = Partial scalar-MI models.

* $p < .001$.

Table B5

Measurement-Invariance Testing - Openness to Extraversion

Model	χ^2 (*df*)	CFI	RMSEA (90% CI)	SRMR	Model comp	$\Delta\chi^2$ (Δdf)	ΔCFI	ΔRMSEA	ΔSRMR	Decision
M1	439.88 * (72)	.927	.078 (.071 - .085)	.038	--	--	--	--	--	--
M2	547.89 * (107)	.914	.069 (.063 - .075)	.047	M1	102.13 * (35)	.012	.009	.008	Accept
M3	2383.50 * (142)	.601	.129 (.125 - .134)	.098	M2	2511.6 * (35)	.313	.06	.051	Reject
M3a	1494.76 * (135)	.749	.105 (.100 - .110)	.072	M2	1164.5 * (28)	.165	.036	.025	Reject
M3b	921.85 * (128)	.850	.084 (.078 - .089)	.062	M2	425.68 * (21)	.064	.014	.015	Reject
M3c	685.42 * (121)	.892	.073 (.068 - .078)	.052	M2	149.49 * (14)	.022	.004	.005	Reject
M3d	576.06** (114)	.911	.068 (.063 - .074)	.048	M2	26.91 * (7)	.003	.001	.001	Accept

Note. Rejection/Acceptance was based on CFI and SRMR. Reported are robust model fit indices. N = 7504. Eight groups á n = 938. M1 = Configural-MI model. M2 = Metric-MI model. M3 = Full scalar-MI model. M3a – M3d = Partial scalar-MI models.
* $p < .001$.

Appendix C

Paired-Samples t-Tests Comparing Partial-MI Scores and Sum Scores for the Trait Openness to Experience

Nationality-by-gender group	Mean difference	Test statistic
German men	0.24	23.7 *
Swedish men	0.32	31.4 *
Irish men	0.11	10.3 *
Dutch men	0.28	27.9 *
German women	-0.20	-20.8 *
Swedish women	-0.17	-18.7 *
Irish women	-0.36	-34.9 *
Dutch women	-0.22	-23.8 *

Note. One-sided paired-samples t-tests were conducted. *n*s = 938 and *df*s = 937 for all tests. For all male groups: H1: Partial-MI score > sum score; H0: Partial-MI score ≤ sum score. For all female groups: H1: Partial-MI score < sum score; H0: Partial-MI score ≥ sum score.
* $p < .001$.