Journal of Research in Personality 113 (2024) 104551

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Brief Report

Comparability of personality facets between men and women: A test of measurement invariance in IPIP-NEO facets in 49 countries

Tim Temizyürek [a], George B. Richardson [b], Gillian R. Brown [a,*]

[a] School of Psychology & Neuroscience, University of St Andrews, UK
[b] School of Human Services, University of Cincinnati, USA

ARTICLE INFO

Keywords:
Gender
Sex
Measurement invariance
IPIP-NEO
Personality facets

ABSTRACT

Identifying whether men and women differ in their personalities, and whether such differences are robust across populations, requires researchers to consider measurement invariance (MI) when comparing between groups. Here, we examined thirty facets of the 120-item IPIP-NEO personality measure between genders (49 countries, N = 831,849). Confirmatory factor analyses (CFAs) revealed that only about half of the facets exhibited robust factor structure within each gender. Based on multi-group CFAs, some facets consistently exhibited scalar MI between the genders in the majority of countries, whereas others reached this MI level in few, or zero, countries. These findings suggest that caution is warranted when comparing personality between genders in cross-cultural datasets and that such comparisons might be more appropriate for some facets than others.

1. Introduction

Numerous studies have investigated whether, on average, men and women differ on personality measures and whether these gender differences in personality are exhibited universally across countries (e.g., Costa et al., 2001; Murphy et al., 2021). Some gender differences in personality traits appear to be relatively consistent across populations; for example, Schmitt and colleagues (2008) reported that, on average, men score higher than women on emotional stability, and women score higher than men on agreeableness, in the majority of sampled countries (N = 55; see also McCrae et al., 2005). Gender differences in personality have also been reported to be relatively stable across age groups (Kajonius and Johnson, 2018). Although these consistent gender differences have been suggested to reflect ‘innate’ differences in psychological traits between men and women (Schmitt et al., 2017), universal gender differences could potentially be underpinned by consistent gender-role socialisation across populations (Wood and Eagly, 2012). Therefore, the interpretation of gender differences in personality remains a topic of ongoing debate.

Yet, the basic idea of comparing personality scores across groups (e.g., genders) remains contentious, because such interpretations are based on strict assumptions about personality measurement that are not always acknowledged. These assumptions relate to factor structure of the questionnaire, concerning the sources of people’s responses to items (individual questions) and the information contained in composite scores when these items are combined (e.g., when responses are summed). Researchers have argued that assumptions must be checked to ensure that between-group comparisons of means or variances (e.g., between genders) are valid (Wang et al., 2018). Minimally, these assumptions are (a) configural invariance (the same items load onto the same factors, and the number of factors is the same across groups); (b) metric invariance (factor loadings are equal across groups); and (c) scalar invariance (item intercepts or thresholds are equal across groups). Configural invariance ensures that a factor score can be meaningfully discussed across the groups; metric invariance ensures it has the same unit of measurement so that group differences can be interpreted; and scalar invariance ensures the item points of origin are the same across groups. The method for checking these assumptions is called measurement invariance (MI) testing (Meredith 1993; Wang et al., 2018).

Measurement non-invariance (or partial invariance) can occur for several reasons (e.g., the effects of translating personality instruments into different languages), and the resulting systematic measurement errors, or ‘measurement biases’, can negatively impact the validity of between-group comparisons (van de Vijver and Tanzer, 2004). In addition to the type of confounds that might occur during testing (e.g., group versus individual assessment), participants in different settings might interpret items differently depending upon their point of reference; for example, responses might vary depending on whether

[*] Corresponding author at: School of Psychology & Neuroscience, University of St Andrews, South Street, St Andrews KY16 9JP, UK.
E-mail address: grb4@st-andrews.ac.uk (G.R. Brown).

https://doi.org/10.1016/j.jrp.2024.104551
Received 3 May 2024; Received in revised form 1 November 2024; Accepted 14 November 2024
Available online 16 November 2024

participants compare themselves to in-group members (own gender) or out-group members (opposite gender). Without measurement comparability, between-group comparisons of personality scores remain potentially compromised, and any findings should be treated with caution (Lasker et al., 2022). While some have argued that expecting scalar MI to be achieved is unrealistic, because the underlying constructs are not necessarily identical across groups or because the assumptions of multigroup confirmatory factor analysis (MGCFA) models are overly restrictive (Han et al., 2019; Welzel et al., 2023), the importance of establishing at least configural invariance is not controversial, given that configural non-invariance often means that scores summarize responses to different sets of items.

This study tested for MI between genders using a large database containing responses to the IPIP-NEO from multiple countries (Johnson, 2014). This personality instrument consists of five domains, each of which is comprised of six facets. Given that gender differences in personality are often more pronounced at lower levels (Kajonius and Johnson, 2018), examining personality at the facet level has the potential to increase the scope and precision of personality research. By focusing on the facet level, our study provided a fine-grained perspective and allowed for the possibility that between-gender MI might be upheld for some facets, but not others, within a domain. By calculating between-gender MI in a large number of countries, we were able to examine whether certain facets performed consistently better than others internationally, in terms of MI across the genders, and whether any differences in between-gender facet performance varied systematically across the domains. In summary, using data for thirty IPIP-NEO facets, we conducted i) confirmatory factor analyses (CFA) to investigate facet structure within groups (i.e., within each gender in each country), and ii) MGCFAs to estimate MI, and thus comparability, across the genders in each country.

2. Methods

We used an open-access personality dataset collated by Johnson (2014) that contains responses to the IPIP-NEO. For each of the five domains, the six associated facets are: i) Agreeableness: *Altruism*, *Cooperation*, *Morality*, *Modesty*, *Sympathy*, *Trust*; ii) Conscientiousness: *Achievement-striving*, *Cautiousness*, *Dutifulness*, *Orderliness*, *Self-discipline*, *Self-efficacy*; iii) Extraversion: *Activity level*, *Assertiveness*, *Cheerfulness*, *Excitement-seeking*, *Friendliness*, *Gregariousness*; iv) Neuroticism: *Anger*, *Anxiety*, *Depression*, *Immoderation*, *Self-consciousness*, *Vulnerability*; and v) Openness: *Adventurousness*, *Artistic interest*, *Emotionality*, *Imagination*, *Intellect*, *Liberalism*.

The IPIP-NEO data were collected online, and respondent information included 'sex' (female/male, hereafter 'gender'), age (years) and country of origin. The full dataset ($n = 926,463$ respondents) contains both 120-item and 300-item versions, which have been cross-validated with each other (Johnson, 2014), and, as in Kaiser (2019), and both versions were combined by retaining only the items from the 120-item version. For the current study, 1268 entries were removed due to unclear country names (e.g., 'Micronesia'), and 58,356 entries from respondents younger than 16 years of age were removed. The original IPIP-NEO database had already undergone systematic screening for protocol validity to remove duplicates and inattentive responding (Johnson, 2005). We imputed single missing responses (0.002 % of items) with facet means and omitted 7869 participants who left two or more items per facet unanswered. For our analyses, 17,142 entries were removed by including only countries with > 500 participants (Hirschfeld et al, 2014), although most countries far exceed 500 participants. Prior to the analyses, models were checked for negative latent or observed variances, in order to identify potential Heywood cases; where Heywood cases were identified across all analyses, the data from the relevant countries were removed (Egypt, Philippines and Thailand). The final dataset consisted of 831,849 participants from 49 countries (Table S1).

All statistical analyses were performed in R, and CFAs were run with the 'lavaan' software. Data were treated as categorical in all analyses, and the MLR estimator was used, because fit index behaviour is best understood for MLR estimators. We tested each facet individually, where each facet model contained one latent variable and four observed variables (items). The first item, as identified in the item key, served as the reference indicator (see *Data availability statement* for information about open access research protocols, data and R script).

The first part of the analyses involved generating CFA models to examine facet structure (unidimensionality) for i) male datasets, and ii) female datasets, separately for each country. Goodness-of-fit criteria for CFA models were based on values of the root mean square error of approximation (RMSEA) and comparative fit index (CFI) (Cheung and Rensvold, 2002). Models were considered acceptable with an RMSEA value of ≤ 0.08 and a CFI value of > 0.9. A small number of countries that produced Heywood cases in the CFA models were removed (male datasets: Hong Kong, Lebanon, United Arab Emirates; female datasets: China, India, Indonesia, Taiwan). A Pearson correlation coefficient was computed to compare the performance of facets in the male and female datasets. A two-way ANOVA was used to examine i) the number of facets that passed these thresholds within the male and female datasets, and ii) the number of facets that passed the threshold within each Big Five domain.

The second part of the analyses involved generating MGCFA models to compare factor structures across gender within each country. We ran configural, metric and scalar invariance models. Goodness-of-fit was based on the values of RMSEA (≤ 0.08) and CFI (> 0.9), and a drop of $>$ 0.01 in CFI (from configural to metric, and metric to scalar) was considered indicative of non-invariance (Cheung and Rensvold, 2002). In keeping with the nested analysis structure, scalar invariance was only accepted for any given comparison if configural and metric models were also acceptable according to these criteria. A small number of countries were producing Heywood cases in the MGCFAs and were excluded (China, Hong Kong, India, Indonesia, Lebanon, Taiwan, United Arab Emirates). For the majority of countries in the dataset, the sample sizes for each gender were similar (Table S1); although sample sizes were unbalanced by gender in some countries (e.g., India = 64 % men), MGCFA has been shown to be robust to these levels of imbalance (Yoon and Lai, 2018). A one-way ANOVA was used to compare the number of facets in each domain that reached scalar invariance across the genders.

3. Results

3.1. Facet structure within each gender in each country

The CFA models revealed that, on average, the fit of the single-factor model was adequate for around half of the facets in the male datasets and around half of the facets in the female datasets (see Supplementary Material: Table S1). Across countries, the average number of facets that reached the goodness-of-fit criteria under a unidimensional model did not differ between male datasets (mean = 52 % of facets) and female datasets (mean = 53 %; $F_{1,54} = 0.00$, $p = 0.98$). Those facets that exhibited adequate fit in male datasets were also likely to exhibit adequate fit in female datasets, i.e., the number of countries in which a specific facet exhibited adequate fit in male datasets correlated with the number of countries in which the same facet exhibited adequate fit in female datasets (Fig. 1; $t_{28} = 8.43$, $p < 0.001$, $r = 0.85$).

Some facets reached the CFA goodness-of-fit threshold in several countries in both the male and female datasets, whereas other facets reached the threshold in few countries (Fig. 1; see full model outputs in *Supplementary Material*), and the number of facets that reached the threshold did not vary according to Big Five domain ($F_{4,54} = 0.64$, $p = 0.64$). Of note, the number of facets that reached the criteria threshold in male and female datasets was not higher in countries with the largest sample sizes (see Supplementary Material: Table S1).

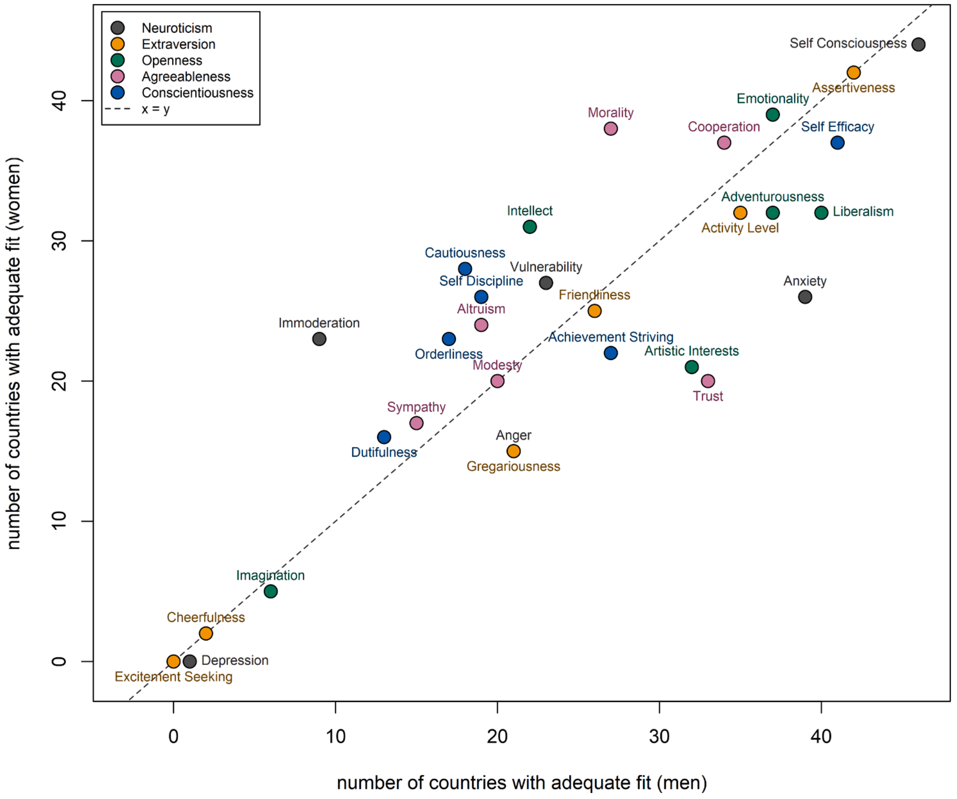

Fig. 1. Number of countries with adequate model fit for men (x-axis) and women (y-axis). Facets are labelled in the plot and coloured based on Big Five domain (note, that *Anger* and *Gregariousness* overlap in the plot).

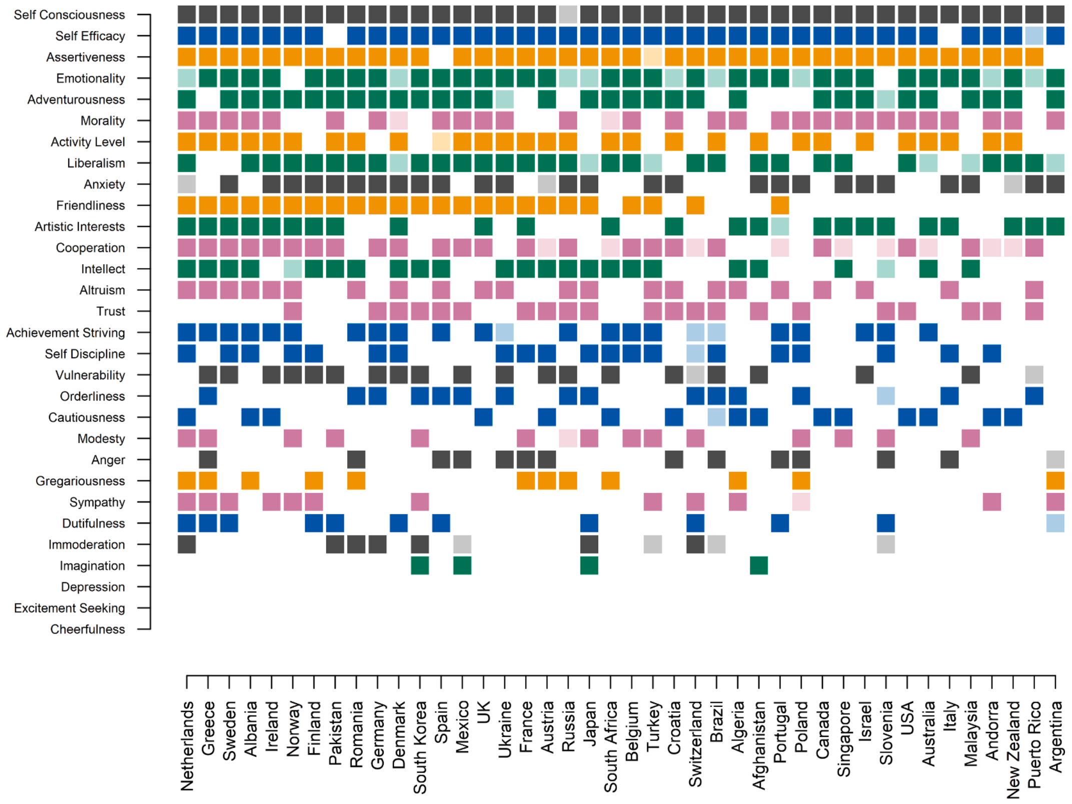

Fig. 2. Results of the MGCFA (within country, across sex) analysis, facets (y-axis) against countries (x-axis). If a facet failed to reach configural invariance no box is drawn, if configural invariance is reached a transparent box is drawn, and if scalar invariance is reached a solid box is drawn. Colour indicates Big Five domain (see Fig. 1). Facets and countries are ordered by decreasing performance, so that the top most facet reached scalar invariance in most countries in the dataset. Similarly, countries to the left reached scalar invariance for more facets than countries to the right.

3.2. Facet structure between genders within countries

The specific facets that reached the MGCFA goodness-of-fit criteria for configural and scalar MI are shown for each country in Fig. 2. To reach scalar MI, facets also had to reach the criteria for configural and metric invariance; approximately half of the facets (48 %) reached scalar MI across the countries (see Supplementary Material: Table S1).

Some facets performed above the goodness-of-fit threshold for scalar MI in the majority of countries (Fig. 2; see full model outputs in *Supplementary Material*); for example, *Self-consciousness* reaches the criteria threshold in 41 out of 42 countries, and both *Self-efficacy* and *Assertiveness* reached the threshold in 40 out of 42 countries. In other words, for those countries where scalar MI criteria are met, comparisons between male and female datasets for these facets can be considered to reflect comparisons between the underlying latent variables. In contrast, some facets performed poorly and reached the goodness-of-fit threshold in few, or zero, countries; for example, *Depression, Excitement-seeking* and *Cheerfulness* did not reach the criteria threshold in any country. Thus, any within-country comparisons between male and female datasets for these facets cannot be considered robust. The relative performance of facets across countries does not appear to be related to geographical location, and the performance of facets is not related to the Big Five domains ($F_{4,25} = 0.22$, $p = 0.93$; see Supplementary Material: Figure S1).

4. Discussion

This study revealed that around half of the facets exhibited robust factor structure within each gender, with the same facets tending to have a robust factor structure in both men and women. These findings suggest that additional efforts might be required to determine the internal structure of the remaining facets. Only around one third of the facets exhibited MI in at least half of the countries. MI was upheld across multiple countries for certain personality facets but not for other facets, and some were not comparable across the genders in any country. Facet performance did not differ according to their placement in the Big Five domains. While our findings raise questions about the comparability of the genders in terms of many personality facets, they do not necessarily imply that comparisons cannot be made. Rather, they suggest researchers need to investigate invariance in their datasets and address it as needed to ensure that they do not draw spurious inferences.

Our analyses bolster calls for caution when making between-group comparisons without first examining measurement invariance (Wang et al., 2018). Kajonius and Johnson (2018) compared the mean scores across genders for the US sample only from Johnson's (2014) IPIP-NEO dataset. Although the authors reported alpha reliability scores for each facet, CFAs were not presented to test whether between-gender latent facet structures were comparable. In our analyses, only eleven out of thirty facets passed goodness-of-fit MI thresholds in the US dataset. Kajonius and Mac Giolla (2017) conducted domain-level CFAs for 22 countries from the IPIP-NEO dataset, but without comparing between-gender factor structure and reported below-threshold CFI scores in many instances. Kaiser (2019) reported acceptable RMSEA values, but unacceptable CFI values, for MI comparisons between genders using the full IPIP-NEO dataset, without reporting comparisons for individual countries. Our results suggest that the conclusions from these previous studies should be interpreted cautiously.

Where violations of MI occur, between-group differences cannot be assumed to be attributable to group differences in the latent constructs, and groups may not be comparable at all if configural invariance does not hold. However, non-invariance is not black-and-white and does not necessarily thwart scientific progress. Where metric, but not scalar, invariance is established, specific comparisons can sometimes still be made (e.g., Ikizer and colleagues (2022) compared correlations between personality scores and stress levels without attempting to interpret between-group mean differences). Researchers can also proceed with partially invariant models when feasible (e.g., when the number of groups is small and non-invariant parameters can be identified and freed), as well as examining the practical significance of ignoring non-invariance (e.g., Lai et al., 2019). Moreover, evidence for non-invariance can lead to important refinements of instruments or novel findings. Further studies looking at the effect sizes underlying MI violation (e.g., Nye and Drasgow, 2011) will help identify factors that differentially affect responses to scale items across groups (e.g., response styles and item interpretation); such factors can be addressed via scale redevelopment efforts. The dimensionality of targeted latent constructs might differ between groups (possibly indicating construct bias); thus, examining the correlates of MI violations might provide insights into between-group phenomena.

A number of limitations can be raised. Although the IPIP-NEO database has numerous benefits (large overall sample size and number of countries), the database is under-represented with data from Africa, Asia and South America. As the questionnaire was administered in English, many participants will not have responded in their first language, which can contribute to MI failure. The original questionnaire asked about sex by providing a binary choice, which is not inclusive for all participants. To maximise the number of countries that were included, a minimum requirement of 500 participants per country was used; future studies could investigate whether MI outputs are influenced by sample sizes. Future studies could also examine the potential impact of other sample variables (e.g., age, socioeconomic status, educational background) that were not considered in our study and examine between-country variables (e.g., Human Development Index) that might correlate with the patterning of facets that reach scalar invariance.

A more general limitation of MGCFA approaches is that expecting scalar MI to hold might be an ideal that is difficult to reach. While some have argued that MGCFA approaches, or MI principles in general, could be eliminated altogether (e.g., Funder & Gardiner, 2024; Welzel et al., 2023), others have provided strong rebuttals in defence of establishing MI (e.g., Meuleman et al., 2023; Lasker, 2024). There should be little disagreement that at least configural invariance is important; yet, using more liberal fit cut-offs than some researchers, we still found that about half of facets did not demonstrate configural invariance, a finding of significance for researchers studying gender differences in personality. Alternative statistical approaches (e.g., alignment methods) that are beyond the scope of this paper address some of the criticism of traditional MI testing, each of which has merits and constraints. Generating numerous, individual CFA models will increase the likelihood that some factor loadings will differ across groups due to stochastic variation, which remains a potential limitation of the current study. Although the analyses indicated that the same personality constructs were being measured in male and female participants for some specific facets, the findings warrant replication using other facet-level datasets or personality measures.

The current study conducted MI testing on facet-level personality scores from male and female participants in a large-scale dataset. The findings suggest that investigating MI at the facet level provides novel insights into the comparability of personality facets between genders. For the roughly half of facets for which scalar MI was upheld, the analyses suggest that the measurement processes were valid and that these facet-level personality construct shows structural equivalence between the genders within the specific country-level datasets. These findings also suggest a form of universality within these psychological phenomena. For roughly half of facets for which MI was not upheld, the findings suggest that either measurement biases influence how male and female participants respond on the items within the facets, potentially compromising inferences based on scores, or suggest that the actual facet-level concepts do not have the same meaning between the genders within the sampled populations. Given the under-usage of MI testing, this study provides empirical support to calls for greater consideration of MI within the psychological sciences.

CRediT authorship contribution statement

Tim Temizyürek: Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **George B. Richardson:** Writing – review & editing, Supervision, Methodology. **Gillian R. Brown:** Writing – original draft, Supervision, Methodology, Conceptualization.

Data availability statement

All original data and analysis scripts are available at https://doi.org/10.17630/2e63ae0e-bc46-4c8d-8c6f-503f06b8ec82. The original data were extracted from an open access dataset (http://ipip.ori.org/).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jrp.2024.104551.

References

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Costa, 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*, 322–331. https://doi.org/10.1037/0022-3514.81.2.322

Funder, D. C., & Gardiner, G. (2024). MIsgivings about measurement invariance. European Journal of Personality, 38, 889-895. DOI: 10.1177/08902070241228338.

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*, 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

Ikizer, G., Kowal, M., Aldemir, I. D., Jeftić, A., Memisoglu-Sanli, A., Najmussaqib, A., Lacko, D., Eichel, K., Turk, F., Chrona, S., Ahmed, O., Rasmussen, J., Kumaga, R., Iddin, M. K., Reynoso-Alcántara, V., Pankowski, D., & Coll-Martín, T. (2022). Big Five traits predict stress and loneliness during the COVID-19 pandemic: Evidence for the role of neuroticism. *Personality and Individual Differences, 190*, Article 111531. https://doi.org/10.1016/j.paid.2022.111531

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality, 39*, 103–129. https://doi.org/10.1016/j.jrp.2004.09.009

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

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

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., & Mac Giolla, E. (2017). Personality traits across countries: Support for similarities rather than differences. *PLoS One1, 12*. https://doi.org/10.1371/journal.pone.0179646

Lai, M. H. C., 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

Lasker, J. (2024). Measurement invariance testing works. *Applied Psychological Measurement, 48*, 257–275. https://doi.org/10.1177/01466216241261708

Lasker, J., Haltigan, J. D., & Richardson, G. B. (2022). Measurement issues in tests of the socioecological complexity hypothesis. *Evolutionary Psychological Science, 8*, 228–239. https://doi.org/10.1007/s40806-021-00301-0

McCrae, R. R., Terracciano, A., & 78 Members of the Personality Profiles of Cultures Project. (2005). Universal features of personality traits from the observer's perspective: data from 50 cultures. Journal of Personality and Social Psychology, 88, 547-561. https://psycnet.apa.org/doi/10.1037/0022-3514.88.3.547.

Meuleman, B., Żółtak, T., Pokropek, A., Davidov, E., Muthén, B., Oberski, D. L., Billiet, J., & Schmidt, P. (2023). Why measurement invariance is important in comparative research. A response to Welzel et al. (2021). Sociological Methods and Research, 52, 1401-1419. https://psycnet.apa.org/doi/10.1177/00491241221091755.

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

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

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology, 96*, 966–980. https://doi.org/10.1037/a0022955

Schmitt, D. P., Long, A. E., McPhearson, A., O'Brien, K., Remmert, B., & Shah, S. H. (2017). Personality and gender differences in a global perspective. *International Journal of Psychology, 52*, 45–56. https://doi.org/10.1002/ijop.12265

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology, 94*, 168–182. https://doi.org/10.1037/0022-3514.94.1.168

van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 54*, 119–135. https://doi.org/10.1016/j.erap.2003.12.004

Wang, S., Chen, C., Dai, C., & Richardson, G. B. (2018). A call for, and beginner's guide to, measurement invariance testing in evolutionary psychology. *Evolutionary Psychological Science, 4*, 166–178. https://doi.org/10.1007/s40806-017-0125-5

Welzel, C., Brunkert, L., Kruse, S., & Ingelhart, R. F. (2023). Non-invariance? An overstated problem with misconceived causes. *Sociological Methods and Research, 52*, 1368–1400. https://doi.org/10.1177/0049124121995521

Wood, W., & Eagly, A. H. (2012). Biosocial construction of sex differences and similarities in behavior. *Advances in Experimental Social Psychology, 46*, 55–123. https://doi.org/10.1016/B978-0-12-394281-4.00002-7

Yoon, M., & Lai, M. H. C. (2018). Testing factorial invariance with unbalanced samples. *Structural Equation Modeling, 25*, 201–213. https://doi.org/10.1080/10705511.2017.1387859