methods, data, analyses | Vol. 18(1), 2024, pp. 59-78 DOI: 10.12758/mda.2024.01

Migrant Health Inequalities or Unequal Measurements? Testing for Cross-cultural and Longitudinal Measurement Invariance of Subjective Physical and Mental Health

Manuel Holz & Jochen Mayerl

Chemnitz University of Technology, Faculty of Behavioral and Social Sciences

Abstract

Background: The aim of the study is to investigate the longitudinal and cross-cultural measurement invariance of the Short-Form 12-Item Health Survey (SF-12) between Native Germans, European migrants and Non-European Migrants. Further, we test for differences in latent means dependent on invariance restrictions.

Methods: We include 7 waves (2006-2018) from a representative panel study in Germany. We apply Multigroup Confirmatory Factor Analysis via a Structural Equation Modelling approach. Finally, we compare gender and age adjusted latent means between different settings of invariance assumptions.

Results: The decrease in model fit measures by increasing equality constraints on the SF-12 factor structure of both physical and mental health between origin groups and across time is within common thresholds for good model fit. Latent means of both health factors differ, dependent on whether scalar invariance is set longitudinally and cross-culturally, or only longitudinally.

Conclusion: We conclude acceptable longitudinal and cross-cultural measurement invariance of the SF-12 for a period of 12 years. Yet, ignoring multigroup scalar invariance constraints produces bias in the latent means of both health factors, where migrant health is shown to be overestimated, especially for Non-European migrants if indicator intercepts are not sufficiently constrained.

Keywords: measurement invariance, cross-cultural comparison, longitudinal study, health inequality, migration, structural equation modelling, SF-12

The study of migrant health inequalities is a crucial and timely issue in post-industrial countries.

The complex nature of health inequalities in migrants is influenced by both subjective and objective factors. In terms of objective measures, migrants often exhibit a higher prevalence of chronic conditions like cardiovascular disease and obesity compared to the native populations (Raza et al., 2017; Rellstab et al., 2016). However, depending on how comparison groups are defined, e.g. with respect to duration of stay, results might differ. It was shown that recent migrants may actually show health advantages in chronic conditions, a phenomenon known as the "Healthy Migrant Effect" (HME) (McDonald et al., 2004). When comparing different countries of origin, variations in prevalence levels and differences compared to native populations have been observed in metrics like obesity (Campostrini et al. 2019), adverse cholesterol levels (Hergenç et al., 1999) and mortality rates (Weitoft et al., 1999).

The examination of subjective measures of health adds further complexity to the picture. On the one hand, there is evidence that newly arrived migrants experience health advantages in terms of subjective physical and mental health scores (Holz, 2022). On the other hand, when all migrants are compared with the native population, only minimal differences in physical and mental health scores persist (Metzing et al., 2019; Wengler, 2011). In particular, migrants from Western countries (Europe, Canada, the United States, etc.) tend to report higher self-rated health outcomes than their counterparts from non-Western countries (Acevedo-Garcia et al., 2010; Holz, 2022).

However, assessing subjective health measures in a cross-cultural context raises certain methodological challenges. Comparative social research has extensively demonstrated the impact of cultural contexts on cognition (Schwarz et al., 2010). Culture variant elements such as value orientations (e.g., individualism vs. collectivism) and other contextual information are strongly linked to cognitive processes during the survey response phase (Schwarz et al., 2010; Sudman et al., 1996; Tourangeau et al., 1988) and can therefore potentially induce bias, leading to variations in the interpretation of results of survey data.

In order to draw valid conclusions about differences in aspects of health between respondents from different cultural contexts, two important aspects need to be considered: firstly, the potentially different ways in which issues of illness, health and disease are expressed need to be taken into account. Secondly, it is necessary to test whether respondents consider the same aspects with the same

Direct correspondence to

Manuel Holz, Chemnitz University of Technology, Faculty of Behavioral and Social Sciences, Institute for Sociology, Thüringer Weg 9, 09126 Chemnitz, Germany
E-mail: manuel.holz@hsw.tu-chemnitz.de

importance and meaning when confronted with a particular object of thought. The existence of differences in meanings, cognition and response behavior can be empirically demonstrated by testing for measurement invariance (also known as measurement equivalence) (Cheung et al., 2000).

This article contributes to the field of comparative social research by addressing a crucial question: whether subjective health measures are genuinely comparable across groups and time periods in Germany. More specifically, our study focuses on assessing the longitudinal and cross-cultural measurement invariance of the Short-Form 12-Item Health Related Quality of Life Questionnaire (SF-12) in its physical and mental health components. The study spans 12 years, from 2006 to 2018, and includes three different groups of origin: European migrants, non-European migrants and native-born Germans without a migration background.

Conceptual Background

Culture, Health and Bias

The formation of health attitudes is significantly influenced by differences in cognition and cultural factors, as they are strongly determined by information from the social, institutional and media environment (Bakanauskas et al., 2020). This influence can lead to differences in attitudes, their conceptualization and survey response behavior. For example, the attribution of causes of disease and illness differs between ‘Western’ and non-‘Western’ populations. The Western perspective tends to follow the biomedical model, emphasizing individual responsibility and secular empirical explanations in the field of health and illness. In contrast, non-Western societies often additionally draw on socio-environmental explanations (‘holistic’ approaches) and may include magico-religious thinking (Bates et al., 1993; Anderson, 1999; Lee et al., 1996).

More precisely, cultural differences play an important role, e.g. in the conceptualization of chronic pain. Hispanic respondents have been shown to be more likely to perceive chronic pain as being beyond the individual’s control, whereas non-Hispanic Caucasians, Italians, French Canadians, Irish or Polish respondents tend to believe that the variation of chronic pain can be influenced by the individual (Bates et al., 1993).

Religion, as a cultural factor, introduces additional bias in the pattern of missing values in survey responses on individual health levels. For example, some highly religious respondents in rural Lebanon refused to rate their future health using the SF-36 questionnaire (a related questionnaire to the SF-12) because it was considered blasphemous to make predictions about the future (Sabbah et al., 2003).

Furthermore, migration-specific issues can bring additional challenges. Migration to post-industrial countries is characterized by positive self-selection in terms of health (Holz, 2022), but variations in general health levels exist among different countries of origin (Jürges, 2007). This raises the issue of social comparison, where individuals assess their level of health based on the strategy of comparison used – whether they compare themselves to those who are better off or those who are worse off, potentially biasing self-rated health upwards or downwards (Beaument et al., 2004).

Measurement Invariance and Subjective Health

When conducting the test for measurement invariance, researchers examine the factor structure of latent constructs not only across groups but also over time (Cheung et al., 2000; Seddig et al., 2018). Only when a latent construct successfully passes the test for measurement invariance can latent mean differences be attributed to real differences between groups or time points, rather than being influenced by variations in the aforementioned contextual factors (Leitgöb et al., 2022).

The status of cross-cultural measurement invariance for subjective health measures remains unclear, with some authors affirming measurement invariance (Schulz, 2012), while others identify differences in factor structures based on cultural or ethnic background (Desouky et al., 2013; Fleishman et al., 2003; Lam et al., 2005). Longitudinal evidence for the invariance of subjective health measures is even more rare, although there is evidence for valid measures of subjective physical health over a period of up to four years (Cernat, 2015; Lynch et al., 2021).

Interest in the SF-12 scale as an instrument has been in both cross-cultural (Holz, 2022) and longitudinal contexts (O’Kelly et al., 2022; Teachman, 2011). However, most evidence for the measurement invariance of the SF-12 has come from separate investigations of the temporal and cultural/ethnic dimensions. To our knowledge, our study is the first to combine a cross-cultural and longitudinal examination of physical and mental health measurement. Based on our findings, we can provide evidence on whether the construction of additive indices or the application of the widely used scoring algorithm (Ware, 2007) leads to unbiased results in longitudinal and comparative studies.

Furthermore, we examine health differences between groups of origin and over time in models where measurement equivalence is partially ignored, in order to explore possible outcome bias due to violation of the invariance assumption. Although our case is limited to Germany, we believe that the results and issues addressed in this paper are transferable to other social and national contexts.

Data and Methods

Participants

We use secondary data from the German Socio-Economic Panel (GSOEP) (Liebig et al., 2021), a representative longitudinal survey of over 12,000 private households in Germany, conducted annually since 1984 by the German Institute for Economic Research (DIW). The survey modes used include CAPI, PAPI, CAWI and CASI, depending on the survey year (Deutsches Institut für Wirtschaftsforschung (DIW Berlin), 2023). Data from this panel is particularly well suited for Structural Equation Modelling, mainly due to its large sample size (more than 12,000 households), which increases the likelihood of detecting potential measurement biases (Meade & Lautenschlager, 2004). In addition, the panel is advantageous due to its deliberate oversampling of migrant respondents from (South) Eastern Europe and Southwest Asia (Herbert Brücker et al., 2014). Respondents were aged 17 and over. Health variables as repeated measures are available in the biennial survey waves of 2002, 2004, 2006, 2008, 2010, 2012, 2014, 2016 and 2018. In order to increase sample sizes for each migration group, waves 2002 and 2004 were excluded from the final sample, mitigating panel attrition concerns associated with a longer observation period.

SF-12

We use both the physical health scale and the mental health scale of the Short-Form 12-Item Health-Related Quality of Life Questionnaire (SF-12). The former is measured by six items: general health, limitations in climbing stairs and performing daily activities, presence of severe bodily pain in the past 4 weeks, limitations in performance due to physical health, and general limitations due to physical health (see Table 1 in the Supplementary Appendix for exact wording and scales). Mental health is measured by six items: frequency of feeling rushed and pressed for time, feeling down and gloomy, feeling calm and relaxed, feeling energetic, having achieved less than desired and doing tasks less thoroughly.

The debate over whether variables used for Health-Related Quality of Life (HRQoL) are reflective or formative indicators is critical (Testa et al., 2021). We argue for treating HRQoL indicators as reflective for the following reasons: firstly, the majority of items (7 out of 12) explicitly tie the health state of respondents as the cause of the health issues (for example: *"When you have to climb several flights of stairs on foot, does your health limit you greatly, somewhat or not at all".)* Secondly, we believe physical health issues cause pain and difficulty in climbing. For objective physical health problems, we argue these problems cause pain, not the reverse. Thirdly, the criterion for formative constructs, that a change in the latent variable has low or no influence on indicators (Diamantopoulos et al., 2021; MacKenzie, 2003) does not apply; as subjective physical health declines, all indicators should tend to decline. Lastly, the widely-used SF-12, treated as reflective, consistently produced reliable results (Schulz, 2012; Kilbourne et al., 2008; Forero et al., 2018).

Migration Background

In our study, migrants are defined as respondents who were not born in the Federal Republic of Germany. Native Germans are identified when both the respondents and their parents were born in Germany. We do not consider indirect migration background or second generation migrants, where only one parent was born abroad or the respondent was born to foreign born parents in Germany, in this analysis. Additionally, we categorize migrants into European and Non-European groups based on the United Nations Statistics Division-Standard Country and Area Codes Classification (United Nations, 2013), utilizing the respondent's country of origin (country of birth) information.

For our focus on longitudinal effects, we only include cases with sufficient panel participation, excluding individuals with more than a total of 20 missing values across the 12 health indicators over all 7 waves. The final sample comprises data from waves 2006 to 2018, consisting of 8,922 cases. Among them, 8,427 are Native Germans (53.0% female, mean age=49.6 (sd=14.79) years), 485 are European migrants (57.4% female, mean age=49.5 (sd=14.51) years), and 164 are Non-European migrants (50.6% female, mean age=44.1 (sd=12.87) years).

In our sample, over 60% of European migrants predominantly originate from Eastern Europe (Poland, Russia, Czech Republic, Romania, Ukraine) and Southern Europe (Italy, Spain, Greece). Meanwhile, the majority (over 50%) of Non-European migrants are from Turkey.

Statistical Methods

The study aims to investigate the extent of measurement invariance in the SF-12 instrument across subgroups of European migrants, Non-European migrants, and native Germans over time, utilizing Multigroup Confirmatory Factor Analysis (MGCFA) within the framework of Structural Equation Modeling (SEM) (Kline, 2016). The procedure involves fitting a baseline model (configural model) where all factor loadings and intercepts are freely estimated across subgroups and waves (Model 0). Subsequent models progressively impose restrictions on factor loadings (metric invariance: Model 1 and Model 3) and intercepts (scalar invariance: Model 2 and Model 4) to be equal across subgroups and waves of the configural model. Measurement invariance is concluded when increasing constraints do not substantially decrease model fit. Because in this analysis the focus lies on migrant health inequalities, multigroup invariance is tested before longitudinal invariance. If the construct does not pass, further invariance steps are unnecessary.

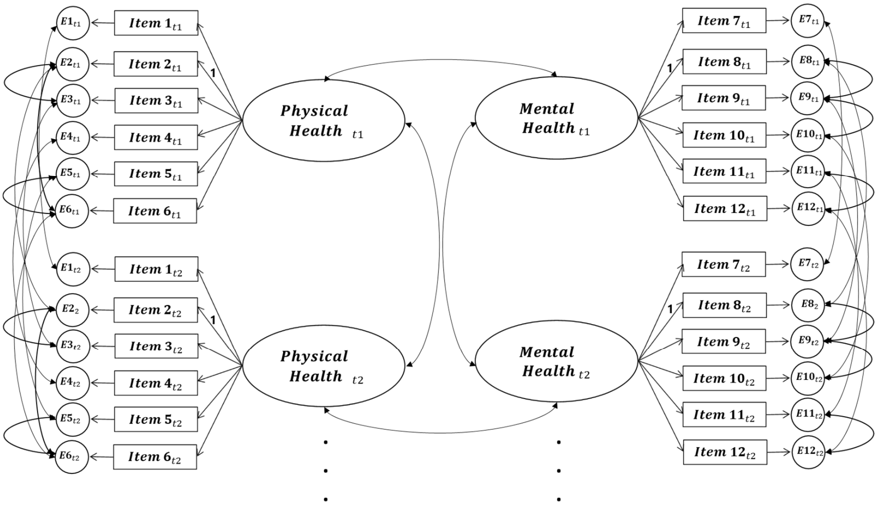
The criteria for establishing invariance include a lack of statistically significant increase in the model Chi-square value, a Comparative Fit Index (CFI) difference smaller than 0.01, a Root Mean Square Error of Approximation (RMSEA) difference smaller than 0.015 with overlapping 95% confidence intervals, and a Standardized Root Mean Square Residual (SRMR) difference smaller than 0.03 (Chen, 2007; Ploubidis et al., 2019). Single model fit criteria include a CFI above 0.95, and RMSEA and SRMR below 0.05 (Kline, 2016; Marsh et al., 2009).

Measurement invariance allows for meaningful comparison of latent factor means across groups and time without construct bias. This ensures that any observed differences in latent factor means (physical health and mental health) are attributable to real differences in the latent factors rather than variations in the properties of the dimensions (factor loadings and item intercepts) (Davidov et al. 2014; Mayerl 2016). Measurement error invariance testing is omitted due to the expected minimal impact on latent means (Joo & Kim, 2019).

Models 0 to 4 depict the primary invariance tests, wherein latent means are restricted to 0. Models 5, 6, and 7 illustrate the potential outcomes for unconstrained latent means in the absence of adequately established scalar invariance. The study calculates latent means adjusted for age and gender for each year by migration group (Model 5), using the Native German group in the first wave (year 2006) as a reference. In the context of SEM, by adjusted latent means we refer to the intercepts of the latent means after controlling for age and gender (both grand mean centered) in the regression (regression coefficients are set equal between origin groups). Potential consequences of insufficient invariance are explored, examining biased latent means due to non-equivalence of intercepts across groups (Model 6) or time (Model 7).

The analysis employs the Full Estimation Maximum Likelihood estimator (FIML) for its efficiency in handling missing values, conducted in RStudio (Version 2022.07.1) and lavaan (Version 06.-12).

Figure 1 illustrates the measurement model for physical and mental health, depicting indicators for each health construct. The model accounts for autocorrelation of error terms across survey years, autocorrelation for both health constructs, three contemporary error correlations per construct, and contemporary correlations between the latent factors physical and mental health. Item 2 in physical health and Item 8 in mental health serve as reference indicators with factor loadings set to 1.00. For brevity, the figure displays the model for the first two time points (t1 and t2), with subsequent waves up to 2018 following the same structure. E1, E2, etc., represent error terms or residuals of the indicators at specific time points. Additionally, contemporary correlations between the latent factors are included. Item



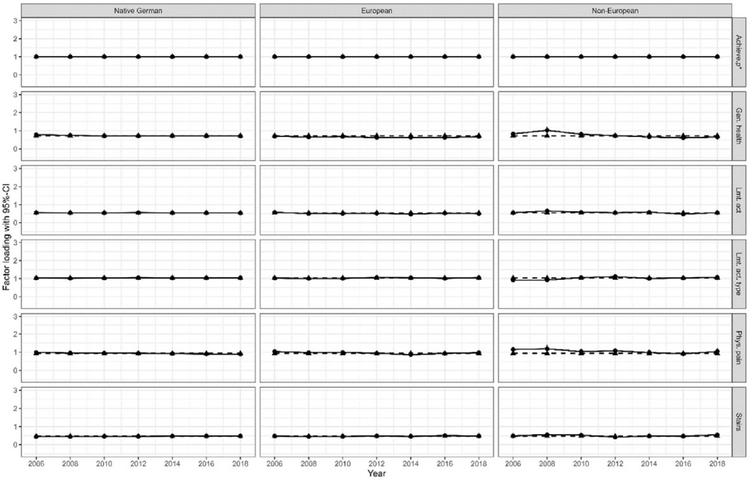
Note: Item wording and response scales can be found in Table 1 and in the Supplementary Material

Figure 1 Measurement model of the SF-12 physical and mental health component

wording details, as well as descriptive statistics can be found in the Supplementary Material (Table 1 and 2).

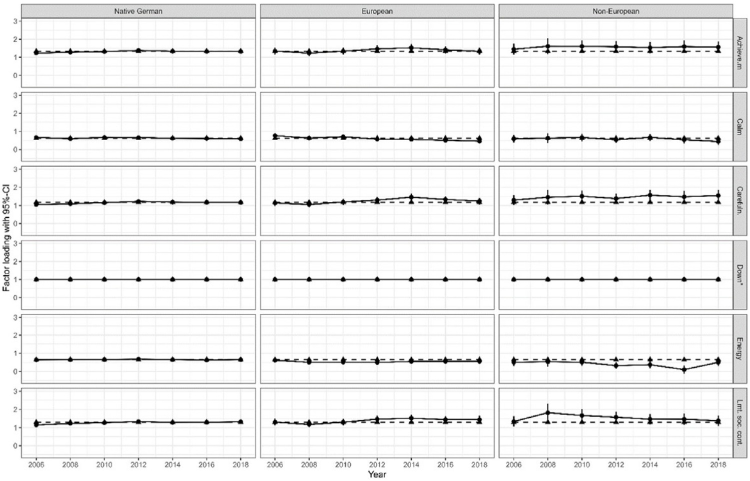
Results

Figures 2 and 3 depict unstandardized factor loadings over time for each migration group under the configural model (Model 0) and the full invariance model (Model 4). The straight line represents Model 0, while the dashed line shows factor loadings from the longitudinal and multigroup invariance model (Model 4). A closer alignment indicates a better fit. Results show that ‘physical health’ (Figure 2) remains consistent across survey years for each origin group, with minimal differences from the invariance model, as almost all factor loadings align and all confidence intervals overlap. In native Germans (Figure 3), ‘mental health’ exhibits no substantial differences between freely estimated factor loadings and metric invariance. However, Non-European migrants show more pronounced variations over time. In the Supplementary Material (Figure 1), standardized factor loadings are sufficiently high in physical health indicators over time, exceeding the 0.5 threshold. The ‘mental health’ indicator (Figure 2) shows weaker performance, especially in later survey waves (2012 to 2018).



Note: Achieve.p*=achieved less due to physical health; Gen.health=General health status; Lmt.act..= limited amount of activities due to physical health, Lmt.act.type=limited in type of activities due to physical health, Phys.pain=Physical pain; Stairs.=problems going up staris due to physical health; *Reference Indicator with factor loading set to 1.00; See Supplementary Material Table 1 for wording and scales

Figure 2 Unstandardized Factor loadings of physical health over time (Model 0 vs. Model 4)



Note: Achieve.m=achieved less due to mental health; Calm=felt calm; Carefuln.= work less thoroughly, Down*=felt down, Energy=felt energetic; Lmt.soc.cont.=limite social contatcs due to mental health; *Reference Indicator with factor loading set to 1.00; See Supplementary Material Table 1 for wording and scales

Figure 3 Unstandardized Factor loadings of mental health over time (Model 0 vs. Model 4)

Table 1 displays data fit measures for each step of invariance restriction. The ‘x’ in each row signifies parameters that were constrained to be equal and whether latent means were computed (in Models 0 to 4 latent means are constrained to 0). It is s worth noting that a sufficient model fit cannot be achieved without including three additional error correlations per construct (see Supplementary Material – Note to 4 for further explanation). At each invariance step, there is a notable rise in chi-square values. However, given that chi-square differences tend to be significant in larger sample sizes, closer scrutiny and detailed discussion are devoted to fit measures. Across all waves, both health constructs exhibit satisfactory fit indices in the configural model (Model 0) with RMSEA = 0.031, SRMR = 0.059, and CFI = 0.958. When factor loadings are restricted across groups (Model 1), the Chi-square value increases significantly, but other fit measures remain almost unchanged (RMSEA = 0.030, SRMR = 0.059, CFI = 0.958). The same holds for Model 2, where intercepts of indicators are set equal across origin groups, with minimal changes in fit indices except for the Chi-square value. In Model 3, setting factor loadings equal across waves results in no difference in RMSEA (0.030) and CFI (0.958), but an increase in SRMR by 0.001 (0.060). The final invariance step, constraining indicator intercepts over time (Model 4), leads to an RMSEA increase of 0.002 (0.032), an SRMR increase of 0.001 (0.061), and a CFI decrease of 0.006 (0.952).

Models 6 and 7 do not establish full scalar invariance, complicating the estimation of latent means for comparing health measures between groups and over time. Comparing Model 5 (full scalar invariance) with Model 6 (no scalar invariance between groups) or Model 7 (no scalar invariance over time) allows us to assess potential outcome bias in health differences when scalar invariance is not fully specified (as in models 6 and 7).

In Figures 4 and 5, latent factor means of health constructs (controlled for gender and age) are presented, categorized by model restriction (Model 5 vs. Model 6 vs. Model 7). When comparing Model 5 and Model 6, differences in the trajectory of the latent construct ‘physical health’ for both migrant groups are evident. In Model 5 (full scalar invariance), European migrant health aligns with Native German health, while Non-European migrants consistently fall below both groups. In Model 6 (no multigroup scalar invariance), both migrant groups nearly follow the same trend, often lacking statistical significance compared to the reference group (Native Germans in the survey year 2006). Figure 5 illustrates that in Model 6, the trajectories for mental health almost align, indicating minimal negative slope. Full scalar invariance in Model 5 produces a similar trend as in the physical health trajectory, where European migrant health approximates Native German levels, and Non-European migrants consistently fall below, suggesting a slight decreasing tendency.

Table 1 Fit measures

Model	Metric MG Invar.	Scalar MG. Invar.	Metric Longit. Invar.	Scalar Longit. Invar.	Latent Means	Chisq (Δ Chisq)	df (Δdf)	CFI	RMSEA (95%-CI)	SRMR
Testing for measurement invariance										
0						34297.22	9051	0.958	0.031 (0.030 - 0.031)	0.059
1	x					34580.80 (283.58***)	9191 (140)	0.958	0.030 (0.030 - 0.031)	0.059
2	x	x				34878.02 (297.22***)	9359 (168)	0.958	0.030 (0.030 - 0.031)	0.059
3	x	x	x			35273.93 (395.91***)	9419 (60)	0.957	0.030 (0.030 - 0.031)	0.060
4	x	x	x	x		38486.67 (3212.7***)	9491 (72)	0.952	0.032 (0.032 - 0.032)	0.061
Calculation of latent means in different invariance settings										
5	x	x	x	x	x	40017.79 (1531.1***)	9951 (460)	0.950	0.032 (0.032 - 0.032)	0.060
6	x		x	x	x	39878.92 (138.87***) [a]	9927 (24)	0.951	0.032 (0.030 - 0.032)	0.060
7	x	x	x		x	38488.25 (1529.5***) [b]	9883 (68)	0.953	0.031 (0.031 - 0.032)	0.061

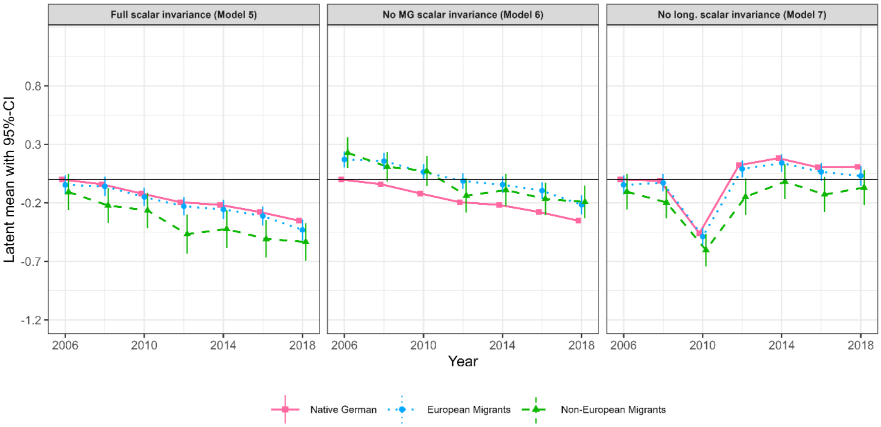
Note: ‘x’ in each row indicates which parameters were restricted to be equal and if latent means were calculated.

MG.: Multigroup, Longit.: Longitudinal; Invar.: Invariance, Chisq: Chi-Square test value, df: degrees of freedom, RMSEA: Root Mean Square Error of Approximation, SRMR: Standardized Root Mean Squared Error, CFI: Comparative FIt Index;

Δ Chisq , Δ df, and Chisq significance difference tests always refer to values compared to the previous model.

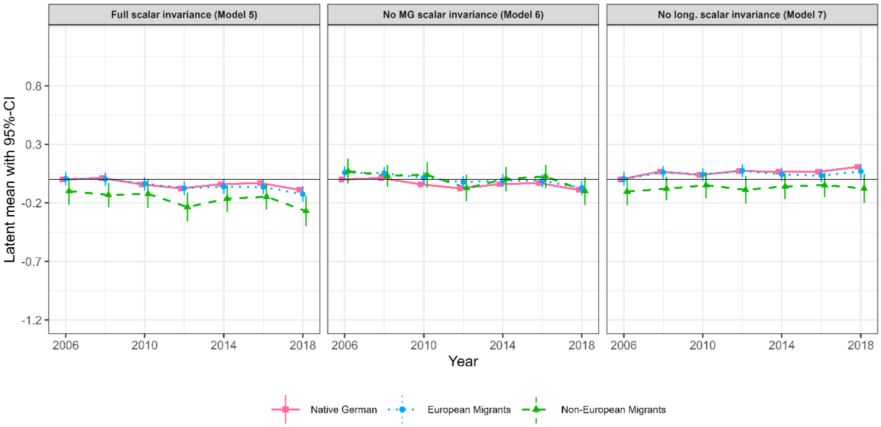
[a] Model 6 is tested against Model 5; [b] Model 7 is tested against Model 5.

p-levels: $p \leq 0.000$: ****, $p \leq 0.001$: ***, $p \leq 0.01$: *



Note: Reference group: Native German in the year 2006 – effects controlled for age and gender (grand mean-centered)
MG: Multigroup; Longit.: Longitudinal; solid symbol: statistically significant to reference group with p ≤ 0.05; empty symbol: statistically non-significant to reference group with p > 0.05

Figure 4 Latent means of physical health by scalar invariance restrictions



Note: Reference group: Native German in the year 2006 – effects controlled for age and gender (grand mean-centered)
MG: Multigroup; Longit.: Longitudinal; solid symbol: statistically significant to reference group with p ≤ 0.05; empty symbol: statistically non-significant to reference group with p > 0.05

Figure 5 Latent means of mental health by scalar invariance restrictions

Further comparisons reveal differences in latent means between Model 5 and Model 7 (no longitudinal scalar invariance). By not setting intercepts equal across waves in both health constructs, the actual downward trend is not captured. Notably, in the physical component, there is a conspicuous abrupt decline in latent means in the year 2010 (Model 7).

Discussion

We can affirm that achieving acceptable metric and scalar measurement invariance is attainable for the latent constructs 'physical health' and 'mental health' of the SF-12 in a German panel survey across diverse groups and over the observation period, as per the established invariance criteria (Chen, 2007). Increasing restrictions on model parameters increases the deviation of the observed and expected matrices in the form of increasing Chi-square values. Nevertheless, the fit measures consistently signal satisfactory model performance (Kline, 2016; Marsh et al., 2009). It is important to highlight that achieving satisfactory data fit relies on incorporating additional error correlations. The improved data fit is presumed to result from factors such as question wording, position, and format.

Our findings align with the current literature (Ploubidis et al., 2019) and SF-12 research in Germany (Schulz, 2012). What sets our study apart is its contribution in integrating both longitudinal and cross-cultural dimensions of measurement invariance. While the invariance of the SF-12 has been examined in a more limited temporal context (≤ 4 years) in previous studies (Cernat, 2015; Lynch et al., 2021) our research extends this examination, affirming the functionality of the SF-12 over a more extensive 12-year time span.

We noted a slightly heightened efficacy of the SF-12 survey for native Germans, with more consistent factor loadings over time, while other groups display more longitudinal variation, e.g. in the mental construct ('energy' in Figure 3)[1]. Despite literature highlighting cultural differences in self-rated health measures (Crockett et al., 2005; Desouky et al., 2013), our study aligns with global fit standards (Schulz, 2012). Limited German language proficiency may contribute to migrants showing disruptions in mental health factor loadings. Prior studies indicate that mental health indicators are prone to Different Item Functioning among ethnic groups (Crockett et al., 2005; Desouky et al., 2013; Fleishman et al., 2003), possibly stemming from diverse interpretations of mental illness (Crockett et al.,

[1] We acknowledge that separate calculations for each group and a comparison of fit indices is needed to deliver an empirical test for differing functionality. We assume that a low level of variation of factor loadings over time is a sign for consistency of the construct and thus a sufficient but not necessary condition for functionality of a questionnaire.

2005; Roberts et al., 1992). Culturally distinct cognitive processes and response styles may also play a role. Investigating nuances like middle category or extreme responding is crucial in measurement invariance research (Weijters et al., 2008). However, merging respondents from different continents into the "Non-European" category may potentially lower the quality of correlational relationships between indicators and factors.

Moreover, we identified significant differences in latent factor means based on whether full scalar invariance between groups and/or time was specified. When the intercepts of the indicators are set equal only across waves, but not across origin groups (Model 5 vs. Model 6), the health of migrants is prone to substantial overestimation. In Model 6 (depicted in Figure 4), an initial physical health advantage of migrants over Native Germans endures over time, with European and Non-European migrants appearing almost indistinguishable. However, when intercepts are additionally set equal across groups (Model 5), the scenario changes markedly. Non-Europeans now exhibit a persistent health disadvantage over time, while the health trajectory for European migrants closely mirrors that of Native Germans. Beyond considerations related to survey response (Fleishman & Lawrence, 2003; Weijters et al., 2008), the potential overestimation of migrant health levels might be attributed to the positive selection of healthier individuals participating in large household surveys (Saß et al., 2015).

Latent means for the 'mental health' construct (refer to Figure 5) also vary depending on the invariance setting. In the 'softer' invariance model, Model 6, we observe minimal differences in latent means between migrants and native Germans, both over time and in terms of longitudinal trends. However, when implementing longitudinal multigroup scalar invariance (Model 5), the scenario changes, revealing that Non-European migrants consistently score below both Native Germans and European migrants. Disregarding longitudinal scalar invariance (Model 7) results in a sudden drop in all latent means of physical health in the year 2010. We attribute this to a potential mode effect, as the composition of survey modes became more reliant on Computer-Assisted Personal Interviews (CAPI) from 2010 onward (Deutsches Institut für Wirtschaftsforschung (DIW Berlin), 2023). The mental health trajectory of Native Germans and European migrants remains almost identical, displaying only a slight decreasing tendency.

This finding contradicts the cross-sectional results of Schulz (Schulz, 2012), where no significant mean differences between origin groups were identified. However, it aligns with the results of Fleishman et al. (2003), where adjustments for Different Item Functioning (DIF) reduced ethnic minority health advantages. Unlike the cross-sectional study by Fleishman et al. (2003), our study reveals changes in latent means in both dimensions of the SF-12 (physical and mental) after imposing invariance constraints. We attribute these differing findings primarily to our longitudinal approach and our focus on (first generation) migration status rather than ethnic minority status.

We conducted a robustness check by recalculating the entire invariance test using the Weighted Least Squares with Mean and Variance adjustment (WLSMV) estimator, applying the threshold invariance approach for ordered-categorical variables as suggested by Liu et al. (2017). Furthermore, we recalculated the model by using the Robust Maximum Likelihood estimator. The results from the threshold invariance and the robust analysis consistently supports our findings (see Supplementary Material). Despite this, for consistency with approaches in the literature (Schulz, 2012; Testa et al., 2021; Anagnostopoulos et al., 2009), we maintain the FIML estimation in our primary analysis.

In addition to our contributions, the analysis comes with certain limitations. We faced a trade-off between the number of waves and sample sizes across various origin groups. Given our specific focus on the consistency concept of 'health' over time, delving deeper into more specific regions of origin was unfeasible due to compromised sample sizes. Future studies could explore this by utilizing a reduced number of later waves from the GSOEP and delving into country-specific differences, as demonstrated in Schulz (2012), while adopting a longitudinal approach.

Another issue dependent on sample size that we could not address is the categorization of migrant groups into recent and non-recent migrants, a crucial element for analyzing the Healthy Migrant Effect. As highlighted in the introduction, chronic health conditions vary based on migration status. Whether the SF-12 yields reliable and valid results when considering different cultural groups over time and under varying chronic conditions (objective health measures) is a question that requires exploration in future research.

Language poses another challenge. While there is some information available about whether a translation of the GSOEP questionnaire was used, the topic itself is intricate. Depending on the survey year, demand and costs; various translations, translation devices, aids, or in-person interpreters were available for the interview (Liebau et al., 2015). There is no information on the language version concerning the questionnaire language at the beginning of our observation period (wave 2006). Drawing valid conclusions about the influence of language on factor structures between groups necessitates further research.

Conclusion

Utilizing seven waves spanning from 2006 to 2018 of the GSOEP and employing a Structural Equation Modelling approach, we examined the intercultural and longitudinal measurement invariance of the SF-12 in both its physical and mental health components. Our findings provide empirical evidence that both scales achieve metric and scalar measurement invariance across native Germans without a migration background, European, and Non-European migrants over time. This finding supports the functionality of summative indices or the standard scoring algorithm (Ware, 2007). However, despite attesting measurement invariance, differences persist among these groups. It is crucial to note that the identification of measurement invariance does not imply that invariance steps can be overlooked in longitudinal multigroup analysis. Instead, we demonstrated that neglecting scalar invariance could lead easily to biased results in latent mean comparisons.

When the longitudinal latent mean difference between the investigated origin groups within a Structural Equation Modelling framework (using lavaan) is the estimand of interest, it is crucial that all models are constrained to full scalar invariance across groups *and* time[2]. Given that the values of latent means heavily rely on the intercept structure of the indicators, the health of migrants, especially Non-European migrants, is susceptible to overestimation if indicator intercepts are not equated.

Consequently, we advocate for the use of a Structural Equation Modelling approach when engaging in intercultural and longitudinal analyses of the SF-12. Special attention should be given to specifying metric and scalar invariance when the focus involves multigroup latent mean differences or health trajectories over time.

Note

The Online Appendix containing the results of robustness checks and analysis scripts can be retrieved from
https://doi.org/10.5281/zenodo.10521878

Statements and Declarations

No potential conflict of interest was reported by the authors.
The authors declare that no funds, grants, or other support were received during the preparation of this manuscript

Ethical approval

Not Applicable.

Informed Consent

Not applicable.

Consent for publication

Not applicable.

[2] The built-in invariance option of lavaan only equalizes the parameters of the grouping variable; if more time points are defined in the measurement model, these parameters have to be set equal manually.

Data availability

The data that support the findings of this study are available from German Institute for Economic Research (Deutsches Institut für Wirtschaftsforschung). Restrictions apply to the availability of these data, which were used under license for this study.

Contribution

Both authors contributed equally to this work and were involved in drafting the manuscript. Both authors read and approved the final manuscript.

References

Acevedo-Garcia, D., Bates, L. M., Osypuk, T. L., & McArdle, N. (2010). The effect of immigrant generation and duration on self-rated health among US adults 2003-2007. *Social Science & Medicine, 71 (*6), 1161–1172. DOI: 10.1016/j.socscimed.2010.05.034.

Anderson, C. A. (1999). Attributional Style, Depression, and Loneliness: A Cross-Cultural Comparison of American and Chinese Students. *Personality and Social Psychology Bulletin, 25(*4), 482–499. DOI: 10.1177/0146167299025004007.

Anagnostopoulos, F., Niakas, D., & Tountas, Y. (2009). Comparison between exploratory factor-analytic and SEM-based approaches to constructing SF-36 summary scores. *Quality of Life Research, 18,* 53-63.

Bakanauskas, A. P., Kondrotienė, E., & Puksas, A. (2020). The Theoretical Aspects of Attitude Formation Factors and Their Impact on Health Behaviour. *Management of Organizations: Systematic Research, 83* (1), 15–36. DOI: 10.1515/mosr-2020-0002.

Bates, M. S., Edwards, T. W., & Anderson, K. O. (1993). Ethnocultural influences on variation in chronic pain perception. *Pain, 52 (*1), 101–112.
DOI: 10.1016/0304-3959(93)90120-E.

Beaumont, J. G., & Kenealy, P. M. (2004). Quality of life perceptions and social comparisons in healthy old age. *Ageing and Society, 24 (*5), 755–769.
DOI: 10.1017/S0144686X04002399.

Campostrini, S., Carrozzi, G., Severoni, S., Masocco, M., & Salmaso, S. (2019). Migrant health in Italy: a better health status difficult to maintain-country of origin and assimilation effects studied from the Italian risk factor surveillance data. *Population Health Metrics, 17 (*1), 14. DOI: 10.1186/s12963-019-0194-8.

Cernat, A. (2015). Impact of mixed modes on measurement errors and estimates of change in panel data. *Survey Research Methods, 9 (*2), 83-99. DOI: 10.18148/srm/2015.v9i2.5851.

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.
DOI: 10.1080/10705510701301834.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing Extreme and Acquiescence Response Sets in Cross-Cultural Research Using Structural Equations Modeling. *Journal of Cross-Cultural Psychology, 31 (*2), 187–212. DOI: 10.1177/0022022100031002003.

Crockett, L. J., Randall, B. A., Shen, Y.-L., Russell, S. T., & Driscoll, A. K. (2005). Measurement equivalence of the center for epidemiological studies depression scale for Latino and Anglo adolescents: a national study. *Journal of consulting and clinical psychology, 73 (*1), 47–58. DOI: 10.1037/0022-006X.73.1.47.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement Equivalence in Cross-National Research. *Annual Review of Sociology, 40* (1), 55–75. DOI: 10.1146/annurev-soc-071913-043137.

Desouky, T. F., Mora, P. A., & Howell, E. A. (2013). Measurement invariance of the SF-12 across European-American, Latina, and African-American postpartum women. *Quality of life research: an international journal of quality of life aspects of treatment, care and rehabilitation, 22 (*5), 1135–1144. DOI: 10.1007/s11136-012-0232-5.

Deutsches Institut für Wirtschaftsforschung (DIW Berlin). (2023). Survey Concepts and Modes. Retrieved from https://companion.soep.de/Survey%20Design/Survey%20Concepts%20and%20Modes.html

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of marketing research, 38 (*2), 269–277.

Fleishman, J. A., & Lawrence, W. F. (2003). Demographic variation in SF-12 scores: true differences or differential item functioning? *Medical care,* III75-III86.

Herbert Brücker, Ingrid Tucci, Simone Bartsch, Martin Kroh, Parvati Trübswetter, & Jürgen Schupp. (2014). Neue Muster der Migration. *DIW Wochenbericht, 81 (*43), 1126–1135. Retrieved from http://hdl.handle.net/10419/104052.

Hergenç, G., Schulte, H., Assmann, G., & Eckardstein, A. von. (1999). Associations of obesity markers, insulin, and sex hormones with HDL-cholesterol levels in Turkish and German individuals. *Atherosclerosis, 145 (*1), 147–156.
DOI: 10.1016/S0021-9150(99)00027-1.

Holz, M. (2022). Health inequalities in Germany: Differences in the 'Healthy migrant effect'of European, non-European and internal migrants. *Journal of Ethnic and Migration Studies, 48 (*11), 2620–2641.

Joo, S.-H., & Kim, E. S. (2019). Impact of error structure misspecification when testing measurement invariance and latent-factor mean difference using MIMIC and multiple-group confirmatory factor analysis. *Behavior research methods, 51 (*6), 2688–2699.
DOI: 10.3758/s13428-018-1124-6.

Jürges, H. (2007). True health vs response styles: exploring cross-country differences in self-reported health. *Health economics, 16 (*2), 163–178. DOI: 10.1002/hec.1134.

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling (Fourth; T. D. Little, Ed.).* New York (UK): The Guilford Press.

Lam, C. L. K., Tse, E. Y. Y., & Gandek, B. (2005). Is the standard SF-12 health survey valid and equivalent for a Chinese population? *Quality of life research: an international journal of quality of life aspects of treatment, care and rehabilitation, 14 (*2), 539–547.
DOI: 10.1007/s11136-004-0704-3.

Lee, F., Hallahan, M., & Herzog, T. (1996). Explaining Real-Life Events: How Culture and Domain Shape Attributions. *Personality and Social Psychology Bulletin, 22 (*7), 732–741. DOI: 10.1177/0146167296227007.

Leitgöb, H., Seddig, D., Asparouhov, T., Behr, D., Davidov, E., Roover, K. d., et al. (2022). Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. *Social Science Research.*
DOI: 10.1016/j.ssresearch.2022.102805.

Liebau, E., & Tucci, I. (2015). Migrations- und Integrationsforschung mit dem SOEP von 1984 bis 2012: Erhebung, Indikatoren und Potenziale. *Berlin: Deutsches Institut für Wirtschaftsforschung (DIW) (SOEP Survey Papers, 270).* Retrieved from
https://www.econstor.eu/handle/10419/111916.

Liebig, S., Goebel, J., Schröder, C., Grabka, M., Richter, D., Schupp, J., et al. (2021). Sozio-oekonomisches Panel, Daten der Jahre 1984-2019 (SOEP-Core, v36, EU Edition). *Berlin: Kantar Deutschland GmbH.*

Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological methods, 22* (3), 486.

Lynch, C. P., Cha, E. D. K., Mohan, S., Geoghegan, C. E., Jadczak, C. N., & Singh, K. (2021). Two-year validation and minimal clinically important difference of the Veterans RAND 12 Item Health Survey Physical Component Score in patients undergoing minimally invasive transforaminal lumbar interbody fusion. *Journal of neurosurgery. Spine*. DOI: 10.3171/2021.6.SPINE21231.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory Structural Equation Modeling, Integrating CFA and EFA: Application to Students' Evaluations of University Teaching. *Structural Equation Modeling: A Multidisciplinary Journal, 16 (*3), 439–476.
DOI: 10.1080/10705510903008220.

Mayerl, J. (2016). Environmental concern in cross-national comparison: Methodological threats and measurement equivalence. In: *Green European: Routledge,* 210–232.

McDonald, J. T., & Kennedy, S. (2004). Insights into the 'healthy immigrant effect': health status and health service use of immigrants to Canada. *Social science & medicine, 59 (*8), 1613–1627. DOI: 10.1016/j.socscimed.2004.02.004.

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo Study of Confirmatory Factor Analytic Tests of Measurement Equivalence/Invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 11 (*1), 60–72. DOI: 10.1207/S15328007SEM1101_5.

Metzing, M., & Schacht, D. (2019). Gesundheitliche Situation der Bevölkerung mit Migrationshintergrund in Deutschland - Sonderauswertung für die Bundesintegrationsbeauftragte 2019. *SOEP Survey Papers*. Retrieved from
http://hdl.handle.net/10419/196880.

O'Kelly, B., Vidal, L., Avramovic, G., Broughan, J., Connolly, S. P., Cotter, A. G., et al. (2022). Assessing the impact of COVID-19 at 1-year using the SF-12 questionnaire: Data from the Anticipate longitudinal cohort study. *International journal of infectious diseases: IJID: official publication of the International Society for Infectious Diseases, 118,* 236–243. DOI: 10.1016/j.ijid.2022.03.013.

Ploubidis, G. B., McElroy, E., & Moreira, H. C. (2019). A longitudinal examination of the measurement equivalence of mental health assessments in two British birth cohorts. *Longitudinal and Life Course Studies, 10 (*4), 471–489.
DOI: 10.1332/175795919X15683588979486.

Raza, Q., Nicolaou, M., Dijkshoorn, H., & Seidell, J. C. (2017). Comparison of general health status, myocardial infarction, obesity, diabetes, and fruit and vegetable intake between immigrant Pakistani population in the Netherlands and the local Amsterdam population. *Ethnicity & health, 22 (*6), 551–564. DOI: 10.1080/13557858.2016.1244741.

Roberts, R. E., & Sobhan, M. (1992). Symptoms of depression in adolescence: A comparison of Anglo, African, and Hispanic Americans. *Journal of Youth and Adolescence, 21 (*6), 639–651. DOI: 10.1007/BF01538736.

Sabbah, I., Drouby, N., Sabbah, S., Retel-Rude, N., & Mercier, M. (2003). Quality of life in rural and urban populations in Lebanon using SF-36 health survey. *Health and Quality of Life Outcomes, 1 (*1), 30. DOI: 10.1186/1477-7525-1-30.

Sara Rellstab, Marco Pecoraro, Alberto Holly, Philippe Wanner, & Karine Renard. (2016). The Migrant Health Gap and the Role of Labour Market Status: Evidence from Switzerland. *IRENE Working Paper.* Retrieved from http://hdl.handle.net/10419/191494

Saß, A.-C., Grüne, B., Brettschneider, A.-K., Rommel, A., Razum, O., & Ellert, U. (2015). Beteiligung von Menschen mit Migrationshintergrund an Gesundheitssurveys des Robert Koch-Instituts. *Bundesgesundheitsblatt, 58 (*6), 533–542. DOI: 10.1007/s00103-015-2146-1.

Schulz, M. (2012). Messartefakte bei der Erfassung der Gesundheit von Migranten in Deutschland: Zur interkulturellen Äquivalenz des SF-12-Fragebogen im Sozio-oekonomischen Panel (SOEP). *SOEPpapers on Multidisciplinary Panel Data Research*. Retrieved from http://hdl.handle.net/10419/59013.

Schwarz, N., Oyserman, D., & Peytcheva, E. (2010). Cognition, communication, and culture: Implications for the survey response process. In: *Survey methods in multinational, multiregional, and multicultural contexts,* 175–190.

Seddig, D., & Leitgöb, H. (2018). Approximate measurement invariance and longitudinal confirmatory factor analysis: concept and application with panel data. *Survey Research Methods, 12 (*1), 29-41. DOI: 10.18148/srm/2018.v12i1.7210.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology.* Jossey-Bass.

Teachman, J. (2011). Are veterans healthier? Military service and health at age 40 in the all-volunteer era. *Social Science Research, 40 (*1), 326–335. DOI: 10.1016/j.ssresearch.2010.04.009.

Testa, S., Di Cuonzo, D., Ritorto, G., Fanchini, L., Bustreo, S., Racca, P., & Rosato, R. (2021). Response shift in health-related quality of life measures in the presence of formative indicators. *Health and Quality of Life Outcomes, 19,* 1-11.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103 (*3), 299–314. DOI: 10.1037/0033-2909.103.3.299.

United Nations. (2013). United Nations Statistics Division-Standard Country and Area Codes Classifications.

Ware, J. E. (2007). *User's manual for the SF-12v2TM health survey.* Lincoln: QualityMetric Incorporated.

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science,36* (3), 409–422. DOI: 10.1007/s11747-007-0077-6.

Weitoft, G. R., Gullberg, A., Hjern, A., & Rosén, M. (1999). Mortality statistics in immigrant research: method for adjusting underestimation of mortality. *International Journal of Epidemiology, 28* (4), 756–763. DOI: 10.1093/ije/28.4.756.

Wengler, A. (2011). The health status of first- and second-generation Turkish immigrants in Germany. *International journal of public health, 56 (*5), 493–501. DOI: 10.1007/s00038-011-0254-8.