REFUGEE MIGRATION TO GERMANY:

A MAGNIFYING GLASS FOR BROADER PUBLIC HEALTH CHALLENGES

PH-LENS

WORKING PAPER SERIES

NO. 04
NOVEMBER 2022

Measurement Invariance of central Social Science and Health Constructs among Syrian, Iraq and Afghan Refugee Groups in Germany

Jasmin Kadel, Natalja Menold & Hagen von Hermanni

Kadel, J., Menold, N. & von Hermanni, H. (2022). Measurement Invariance of central Social Science and Health Constructs among Syrian, Iraq and Afghan Refugee Groups in Germany. Bielefeld: PH-LENS Working Paper Series No. 4, 10/2022. DOI: https://doi.org/10.4119/unibi/2967485

Jasmin Kadel, Natalja Menold & Hagen von Hermanni

Measurement Invariance of central Social Science and Health Constructs among Syrian, Iraq and Afghan Refugee Groups in Germany

PH-LENS Working Paper Series No. 4, Version 1.0
Bielefeld, Dezember 2022
General Editors: Razum, O.; Bozorgmehr, K.; von Unger, H.
ISSN: 2702-2676, DOI: https://doi.org/10.4119/unibi/2967485

Gefördert durch DFG Deutsche Forschungsgemeinschaft

PH-LENS is funded by the German Research Foundation (DFG).

Note for Publication (PH-LENS Working Papers are refereed scholarly papers. Submissions are reviewed by the general editors before a final decision on publication is made.)

Note to Readers (The Working Paper Series is a forum for presenting works in progress. Readers should communicate comments on the manuscript directly to the author(s).)

The PH-LENS Working Paper Series is intended to aid the rapid distribution of work in progress, research findings and special lectures by researchers and associates of PH-LENS. Papers aim to stimulate discussion among the worldwide community of scholars, policymakers and practitioners. They are distributed free of charge in PDF format.

The papers can be downloaded from the project website: https://www.uni-bielefeld.de/fakultaeten/gesundheitswissenschaften/forschungsgruppen/ph-lens/

PH-LENS „Refugee migration to Germany: a magnifying glass for broader Public Health challenges“
University of Bielefeld
School of Public Health
P. O. Box 100131
33501 Bielefeld, Germany
Phone.: +49 521 106-3837
Email: oliver.razum@uni-bielefeld.de

UNIVERSITÄT BIELEFELD Fakultät für Gesundheitswissenschaften
UNIVERSITÄT BIELEFELD Faculty of Psychology and Sports Science
UNIVERSITÄTS KLINIKUM HEIDELBERG
TECHNISCHE UNIVERSITÄT DRESDEN
CHARITÉ UNIVERSITÄTSMEDIZIN BERLIN
ASH Berlin Alice Salomon Hochschule Berlin University of Applied Sciences
IKG Institut für interdisziplinäre Konflikt- und Gewaltforschung
DIW BERLIN
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Measurement Invariance of central Social Science and Health Constructs among Syrian, Iraq and Afghan Refugee Groups in Germany

Jasmin Kadel, Natalja Menold, Hagen von Hermanni

Abstract

(Forced) migration is a major issue all over Europe. To assess requirements and needs of refugees, self-reports such as surveys are a relevant data collection method. Nevertheless, most measurement instruments are developed and validated in Western contexts. Due to i.e. cultural and language differences, comparability cannot be taken for granted. With the help of multi-group confirmatory analyses and here measurement invariance analysis, it can be assessed if data fulfil prerequisites for statistical comparisons. We test four frequently used scales within social science and public health research (Brief-Resilient Coping-Scale, Attitudes towards Democracies, Loneliness, Locus of Control) with regard to their comparability amongst the three major refugee groups in Germany (Syria, Iraq, Afghanistan). Using data from the refugee sample of the Socioeconomic Panel (IAB-BAMF-SOEP) we investigate if configural, metric and scalar invariance – the three in context of survey research relevant types of measurement invariance – can be supported. For three out of four constructs, we fail to reproduce the factor solutions, that were suggested in the literature and found therefore limited configural invariance. Improvement could be reached by weakening theoretical assumptions about factorial structure. Scalar invariance – a prerequisite for meaningful mean comparison – was supported for none of the models. We conclude that statistical comparisons between the different populations are not given or at least restricted in the SOEP refugee sample.

Public Significance Statement

This study presents results on the possibilities and limitations of comparing several scores derived from latent scales across the three major refugee groups in Germany. We show that a direct comparisons of refugees from Afghanistan, Iraq and Syria might be biased, due to a lack of measurement invariance.

1. Introduction

Migration in general and especially forced migration is and will remain a major issue all over Europe and especially in Germany due to several causes such as demographic change, war, climate crisis and relating thereto environmental damages (Missirian & Schlenker, 2017; Myers, 2017). Thus, integration and enabling social participation of refugees within all the different societal spheres such as access to the education system, job market, or health care system is an important and everlasting challenge and opportunity. For this, the process of integration of refugees, that is their introduction in the new symbolic and value systems such as language, governmental and institutional functioning, democratic values and new social and professional roles are crucial (e.g. 1992; Esser, 2006). In this context, data on individual values, knowledge and behaviour of migrants in a country and their change over time are a relevant source of information, which can be obtained from surveys. Through this, researcher found, for example, that many refugees in Germany have experienced violence, are traumatized (Schröder et al., 2018), suffer more from loneliness and depression (Brücker et al., 2016) and show a higher need for psychological assistance (Metzing et al., 2020) than Germans.

In order to make such comparisons validly between different populations, it is necessary that the "principle of equivalence" (Jowell, 1998, p. 169) receives attention within the whole research process. This includes among others that the measurement instruments are comparable between the different populations (Hubley & Zumbo, 2011). A basic prerequisite for valid cross-group comparisons is a bias free and equivalent measurement of the concepts under investigation. As required by Scheuch (1993) not only the comparability of question wording or visual design but also a comparable suitability of data for the analysis of interest should be given. When comparing measurements of different concepts, the bias would stem from different sources such as the translation mistakes, different understanding or use of survey questions or absence of comparable concepts in the target language (Stathopoulou et al., 2019). Taking the modern measurement theory into account, measurement can be defined as structured observation of so-called latent variables (Mellenbergh, 1994; Markus & Borsboom, 2013). This means, that a concept or attribute of interest cannot be observed directly. Instead, researcher develop observable indicators also called manifest variables (e.g. questions, tests or items) and statistically test measurement assumptions. Item responses are then explained as causes of variation it the latent variable (reflective model) (Mellenbergh, 1994; Weiber & Mühlhaus, 2014).

Nevertheless, despite 'principle of equivalence' is demanded through the whole research process, in behavioural research a strong bias toward Western, educated, industrialized, rich and democratic societies has been discussed (Henrich et al., 2010). This implies that most of the survey instruments are developed and validated in these contexts and thus may not apply to non-Western settings. Therefore, measurement equivalence cannot be taken for granted and it is supposed to be assessed. One method to assess the comparability of data of latent constructs between different populations ex-post is the evaluation of variance, co-variance and mean structures, which is known as measurement invariance analysis (Jöreskog, 1971; Meredith, 1993). Testing measurement invariance has been broadly used in

psychological assessment and has been becoming increasing popularity in other disciplines such as sociology or social sciences (Meuleman et al., 2022; Roberts et al., 2020.

Against this backdrop we examine if measurement invariance is supported between different populations among diverse concepts available in public data sets which are available for researcher for secondary analyses. More specifically, we use data of the refugee sample from German Socio-Economic Panel (IAB-BAMF-SOEP) which is one of the most important data sources of refugees in Germany. We investigate if important concepts from psychological, social science and public health research (Loneliness, Resilience, Locus of Control and Attitudes toward Democracy) are equivalent between the three major refugee groups in Germany. Large groups of refugees included in the SOEP database originate from Syria, Afghanistan and Iraq. Therefore, through the systematic testing of different measurement instruments following the same procedure within one data set, our results portray a broader view about comparability amongst the major groups of refugees in Germany.

2. Cross-Population Comparability and Measurement Invariance

Lacking comparability can be caused by many sources (Stathopoulou et al., 2019) such as differences in response process (Tourangeau et al., 2000) leading to differential item functioning (DIF) (Kline, 2016). However, the cause may already exist at a deeper level. The concept of interest may not exist in every population, which might especially occur between strongly different populations. Moreover, it is possible, that distinct indicators are not associated with the concept in the same way in every population. For example, as Putnick and Bornstein (2016) describe illustratively, the "frequency of crying, weight gain, and feelings of hopelessness are indicative of the severity of depression in women, but only feelings of hopelessness are indicative of the severity of depression in men" (p. 72). If one takes these three described observable indicators (= manifest variables) to compare the severity of depression between men and woman, this will be misleading as long as the first two indicators are not associated with the latent construct 'depression' for men. On the basis of these circumstances, the equivalence of the measurement between populations has to be evaluated before examining communalities and differences. In general, measurement invariance means that group membership does not bias the results of comparisons and that across compared groups respondents with the same true trait level have the same probabilities of providing pre-specified responses on individual questions (e.g., Meredith, 1993; Millsap, 2011).

The concept of measurement invariance has been developed within the frame of Latent Variable Modeling (LVM) (Mellenbergh, 1994) and can be evaluated by means of Structural Equation Modeling (SEM; Jöreskog, 1971). With SEM, it is possible to test theoretical assumptions with respect to measurement by empirical evidence. Particularly with a measurement model, the associations between the latent variable under investigation and their observed manifestations can be tested and compared between different groups. There are different levels of measurement invariance and the support of a certain level allows bias-

free comparing either correlations or means between groups as follows (Meredith, 1993; Millsap, 2011; Hox et al., 2015):

The lowest level of measurement invariance is referred to 'configural'. Configural invariance is given, if the basic factor model can be reproduced within the different groups. This means that both, the number of factors – which represent the latent construct – and the allocation of the indicator questions (manifest variables) to the factor(s) correspond between the groups. The next level is so called 'metric invariance'. It is is given, if, additionally, the co-variances between the manifest and latent variables are equal in the groups under consideration. This means that the factor loadings of the manifest variables are equal across groups. This level of invariance is a perquisite to compare correlations between the latent variables or summarized scores. The next level, scalar invariance, is given, if additionally, the items intercepts (means) are equal. If scalar invariance is achieved, valid comparisons between the means of the latent constructs can be made.

3. Methods

3.1. Data

As data basis we used the first three waves of the IAB-BAMF-SOEP refugee sample which were conducted between 2016 and 2019. It is an annual survey in which participants are interviewed repeatedly. Individuals who entered Germany from January 2013 to June 2019 and have applied for asylum, as well as their household members, are interviewed. The sample was drawn from the German central foreign registry (Kroh et al., 2016) which allows making references to the refugee population in Germany. The interviews were conducted face-to-face. In a first step, we identified all respondents stemming from Syria, Iraq and Afghanistan through a corresponding variable in the data set. Second, we restricted it to people who speak Arabic or Dari respectively[1]. Third, as long as the official survey language is German for any respondent, we further restricted our sample to those who used either 'written translation assistance' to respond to at least 50 per cent of the questionnaire or had 'help by an interpreter' or some 'other person' or both (see Table 1). With this procedure we intend to ensure, that the respondents used the questionnaire in their mother tongue or had an interpreter respectively or both. For example, 1408 respondents used the written translation assistance for each question, but had no help by an interpreter or another person.

[1] While Arabic is the predominant language in Syria and Iraq, Dari, which is itself Persian dialect and derived from Farsi, is spoken by the absolute majority of Afghani refugees in Germany and our sample, respectively.

Table 1 Translation Assistance

	Interpreter			
	Yes, professional Interpreter.	Yes, another Person.	No.	**Total**
Use of Written Translation Assistance				
For each question	25	683	1408	**2116**
About two thirds of the questions	3	136	259	**398**
About the half of the questions	1	90	151	**242**
Less than half of the questions	7	63	-	**70**
For none of the questions	7	34	-	**41**
Total	**43**	**1006**	**1818**	**2867**

Thus, our final dataset comprises 2,867 cases. Additionally, we deleted listwise in case of missing data. Therefore, specific group sizes vary for each scale and are shown separately below (Tables 2-5).

3.2. Latent constructs under investigation

We selected measurement instruments for the analyses based on the following considerations. First, each construct under investigation has to have a relevance in social science or public health research, especially within the refugee context. Second, each construct viz. latent variable is supposed to be measured with at least four manifest variables to properly conduct our analyses.

In the following section, we provide an overview about the concepts and instruments we used in our analysis. More detailed information can be retrieved from the IAB-BAMF-SOEP Scales Manual (Jacobsen et al., 2017).

3.2.1. Brief Resilient Coping Scale (BRCS-4)

Within psychology, resilience refers to origins of stress (i.e. trauma) and the ability to cope with these situations (Kocalevent et al., 2015). The BRCS-4 scale was developed to cope with stress

in an adaptive manner (Sinclair & Wallston, 2004). It consists of one factor which is measured by four manifest variables. Jacobsen et al. (2017) report a Cronbach's Alpha of .60 amongst the general refugee survey of the SOEP, which is inacceptable low. We achieved a Cronbach's Alpha of .64 within our subsample.

Table 2 Brief Resilient Coping Scale

No	Question Wording	Response Options	$N_{Afg.}$	N_{Iraq}	N_{Syria}	N_{total}
1	I try to think of how I can change difficult situations.	1 (Totally disagree) to 7 (Totally agree)	268	411	1926	1717
2	No matter what happens to me, I think I have my reactions under control.	1 (Totally disagree) to 7 (Totally agree)	381	402	1932	2715
3	I think I can develop further if I deal with difficult situations.	1 (Totally disagree) to 7 (Totally agree)	376	405	1886	2667
4	I actively seek ways to balance out the losses that have affected my life.	1 (Totally disagree) to 7 (Totally agree)	377	402	1923	2702

3.2.2. Attitudes toward Democracies

The construct is supposed to measure democratic values through the (dis-)agreement to different goals of a democracy. The items used within the IAB-BAMF-SOEP survey are derived from the World Value Survey (WVS) but were adapted by SOEP. Thus, the scale used within the SOEP survey consists of six items and the one by WVS only of four items (Items 1-4). The theoretical background or development procedure has not been documented (cf. Jacobsen et al., 2017). Furthermore, the response categories were adapted by SOEP to align them with the style of the survey. SOEP reports a Cronbach's Alpha .61, whereas our sub-sample solely shows a Cronbach's Alpha of .54 for the same 6-Item-Version and .53 for the WVS 4-Item-Vesion. In our analyses, we tested both, the original four and the six-item version from SOEP.

Table 3 Attitudes towards Democracy

No	Question Wording	Response Options	$N_{Afg.}$	N_{Iraq}	N_{Syria}	N_{total}
Intro	**Do you think that the following things are what should happen in a democracy or not?**					
1	**The government taxes the rich and supports the poor**.	0 (Should definitely not happen in a democracy.) to 10 (Should definitely happen in a democracy.)	320	403	181	2540
2	**The people choose their government in free elections.**	0 (Should definitely not happen in a democracy.) to 1 0 (Should definitely happen in a democracy.)	341	419	1864	2624
3	**Civil rights protect the people from government oppression**.	0 (Should definitely not happen in a democracy.) to 10 (Should definitely happen in a democracy.)	293	400	1803	2496
4	**Women have the same rights as men**.	0 (Should definitely not happen in a democracy.) to 10 (Should definitely happen in a democracy.)	344	418	1907	2669
5	Religious leaders ultimately determine the interpretation of laws.	0 (Should definitely not happen in a democracy.) to 10 (Should definitely happen in a democracy.)	293	385	1699	2377
6	Minorities are protected.	0 (Should definitely not happen in a democracy.) to 10 (Should definitely happen in a democracy.)	302	412q	1822	2536

3.2.3. Loneliness

Loneliness is defined as perceived isolation which captures the difference of "desired and actual social relationships" (Luhmann & Hawkley, 2016, p.943). The Loneliness-Scale used in the IAB-BAMF-SOEP comprises of one dimension with four manifest variables. The first three items are based on the short scale of Loneliness by Hughes et al. (2004),[2] who reports an internal consistency of Cronbach's Alpha of .72. The fourth item was added especially for the

[2] Please note that the question-wording in English differ between the original scale and the items used in the SOEP.

Refugee-Survey by the SOEP (Jacobsen et al., 2017) who report a Cronbach's Alpha of .70 for the 4-items versions of the scale. Our subsample Cronbach's Alpha .78 for the three-item scale and .69 with additional item.

Table 4 Loneliness

No	Question Wording	Response Options	$N_{Afg.}$	N_{Iraq}	N_{Syria}	N_{total}
Intro	How often do you...					
1	...miss the company of others?	1 (Very often); 2 (Often); 3 (Sometimes); 4 (Occasionally); 5 (Never)	399	431	1956	2786
2	...feel like an outsider?	1 (Very often); 2 (Often); 3 (Sometimes); 4 (Occasionally); 5 (Never)	368	430	1968	2766
3	...feel socially isolated?	1 (Very often); 2 (Often); 3 (Sometimes); 4 (Occasionally); 5 (Never)	376	435	1977	2788
4	...miss people from your country of origin?	1 (Very often); 2 (Often); 3 (Sometimes); 4 (Occasionally); 5 (Never)	401	439	1986	2826

3.2.4. Locus of Control

Originally, the concept of Locus of control captures the extent of which individuals believe that their lives are the result of their own behavior or that they believe that the things that happen to them lay outside their control (Rotter, 1966). By Rotter, internal-external Locus of Control was considered as a one-dimensional construct. The scale used in the SOEP is based on the scale developed by Nolte et al. (1997). This scale captures the construct in a broader sense and differentiates four dimensions: internal locus of control (3 items), external locus of control (5 items), attitudes about justice (1 item) and individual vs. collective orientation (1 item). Instead, Specht et al. (2013) state that seven (1-2 & 6-10) of the ten Items belong to one dimension whereas Jacobsen et al. (2017) differentiate between the internal (Item 1-5) and (Item 6-10) external locus of control. Nevertheless, they only report a Cronbach's Alpha of .18 for the internal and .45 for the external dimension, which is inacceptable low. Within our analyses we test the one-dimensional approach proposed by Specht et al. (2013) and the two-dimensional approach claimed by Jacobsen and colleagues. Within our sub-sample Cronbach's Alpha of the 7-item- version of .47 and .27 for the 5 internal dimensions and .39 for the external one.

Table 5 Locus of Control

No	Question Wording	Response Options	$N_{Afg.}$	N_{Iraq}	N_{Syria}	N_{total}
1	**My life's direction depends on me.**	1 do not agree at all 7 totally agree	353	382	1819	2554
2	**When I encounter difficulties in life, I often doubt my abilities.**	1 do not agree at all 7 totally agree	349	391	1880	2620
3	You must work hard to achieve success.	1 do not agree at all 7 totally agree	387	421	1954	2762
4	If you are socially or politically active, you can influence social circumstances.	1 do not agree at all 7 totally agree	334	369	1778	2481
5	The abilities we have are more important than the efforts we make.	1 do not agree at all 7 totally agree	363	387	1834	2584
6	**In comparison with others, I haven't achieved what I deserved to achieve.**	1 do not agree at all 7 totally agree	325	380	1812	2517
7	**What can be achieved in life is mainly a result of fate or luck.**	1 do not agree at all 7 totally agree	373	397	1888	2658
8	**I often find that other people dictate my life.**	1 do not agree at all 7 totally agree	357	403	1912	2672
9	**The options that I have in life are determined by social circumstances.**	1 do not agree at all 7 totally agree	342	379	1805	2526
10	**I don't have much control over what happens in my life.**	1 do not agree at all 7 totally agree	343	392	1855	2590

3.3. Data Analysis

All analyses followed the same procedure and were conducted using Mplus 8.1 (Muthén & Muthén, 2017): For each construct under investigation, we set out to replicate the modelling approach(es) described above via multi-group confirmatory factor analysis (MG-CFA) using maximum likelihood estimation (MLR) due to non-normality of data. More specifically, each model was identified by freely estimating item intercepts, factor loadings and residual variances, while setting the latent factor means and variances to 0 and 1 respectively (Byrne, 2011; Kline 2016). Thus, within that first step we examined configural invariance through examining if the variance in the manifest variables is explained through the same latent factor in each group. The model fit of MG-CFAs was evaluated using the chi-square test (CMIN), the

Root-Mean-Square Error of Approximation (RMSEA), and the Comparative Fit Index (CFI) (Beauducel & Wittmann, 2005). As the chi-square statistic is prone to be oversensitive in case of larger sample sizes, RMSEA and CFI are additionally used as absolute fit indexes (cf. Kline ,2016). The CFI should be 0.95 or higher, while an RMSEA of 0.08 or less indicates an acceptable fit (Hu & Bentler, 1999). The Robust Maximum Likelihood estimator (MLR) was used due to the ordinal nature and non-normality of the data (Muthén & Muthén, 2014). The configural model as baseline model should provide an acceptable goodness of fit statistic (GOF) to evaluate the next level of measurement invariance. Metric invariance was evaluated by restricting the loadings to be equal among evaluated refugee groups, whereas scalar measurement invariance is evaluated by introducing equality of intercepts into the metric model. Test of measurement invariance is accessed by model difference test between the configural and metric as well as metric and scalar invariance respectively (e.g. Meredith, 1993; Millsap, 2011). The differences between models were evaluated by difference in CMIN (Satorra & Bentler, 2010) and the changes in RMSEA and CFI. A significant change of chi-square (Meredith, 1993) or a change of $\Delta CFI \geq .010$ and $\Delta RMSEA \geq .015$ indicate significant differences in model fit (Chen, 2007), and thus lack of measurement invariance for the nested models. In the case the configural invariance was violated, we inspected misspecifications looking at the Modification Indeces to be able to proceed and evaluate next levels of measurement invariance. MIs describe the decrease of CMIN, if a modification that is a deviation from the initial model, is introduced. This procedure has been proposed e.g. by Byrne (2011). However, modified models implicate violations from the initially stated models so that modifications were used only to be able to inspect potential additional comparability bias due to the differences in loadings or intercepts.

4. Results

4.1. Brief Resilient Coping Scale (BRCS-4)

While testing configural invariance, Chi-Square Test is significant, which was expectable due to large sample size. CFI and RMSEA point to the acceptable model fit and we assume configural invariance as established. When restricting loadings to be equal, the change of Chi Square Test (ΔCMIN) is not significant. The model fit improves when looking at the change of CFI and RMSEA. Although this change is above the benchmark, the improved model fit point to given metric invariance. Thus, also metric invariance is established. When constraining the intercepts to be equal across groups, the model fit decreases. The change in the Chi-Square value is significant and changes in CFI is strongly above the cut off point. Due to significant change in two of three fit statistics and particularly that of CMIN we reject the scalar model. This decision is supported when looking at the model parameters provided in Appendix. According to MIs, significantly differing intercepts are given for three items of four between Afghan and Syrian samples.

Table 6 Brief Resilient Coping Scale (BRCS-4)

Model	CMIN (df)	ΔCMIN (Δdf)	CFI	ΔCFI	RMSEA Estimate	ΔRMSEA	
Configural Model	34.651 (6)***		0.952		0.072		accepted
Metric Model	33.318 (12)***	4.951 (12.994)	0.965	-0.013	0.044	-0.028	accepted
Scalar Model	56.205 (18)***	25.135 (6.958)***	0.936	0.029	0.048	0.004	rejected

Note. ΔCMIN: Sattora Bentler corrected; ***p < .001, **p < .01, *p < .05.

To sum it up, the concept under investigation exists across all three groups. Due to metric invariance, measurement bias can be excluded as explanation for comparisons of correlations. Comparisons of the means of latent or summarized scores could not be conducted bias free.

4.2. Attitudes towards Democracy

4.2.1. 6-Items Version (SOEP)

First, we set out the 6-item solution proposed by SOEP. The Chi-Square Test is significant and CFI failed the cut-off points for goodness of fit. However, RMSEA point to acceptable fit. Standardized factor loadings revealed, that Item 5 (see Appendix and Table 3) does not seem to be a good indicator for the latent construct for any group (factor loading within the groups are -.247 (p<.01) for Afghans, -.127 (p>.05) for Iraquis, and -.161 (p<.001) for Syrian). With regard to the content of the item ("Religious leaders ultimately determine the interpretation of laws") it is conceivalbe that rather religious convictions than attitudes towards democracies might cause the anwsers to that item. Thus, we exclude this item and test our new model. Both, CFI and RMSEA improve. Within that model, RMSEA is accaptalbe and as long as CFI only narrowly miss our cut-off-points, we just accept the configural model. Therefore, we proceed in our analysis and constrain the factor loadings to be equal. Changes within Chi-Square are signifiacant, which shows, that the nested model fits the data significantly worse than the configural. Furthermore, changes in CFI do not fit the cutt-off points. Looking at the parameters in Appendix, signficant differences of two of six loadings could be found. Thus, we reject the metric model. Since proceeding to the evaluation of scalar invariance assumes metric invariance, we refrain from its evaluation.

Table 7 Attitudes towards Democracy - 6 Item Version (SOEP)

Model	CMIN (df)	ΔCMIN (Δdf)	CFI	ΔCFI	RMSEA Estimate	ΔRMSEA	
Configural Model	111.0520 (27)***		0.890		0.059		rejected
Config without I5	58.1120 (15)***		0.922		0.056		Just accepted
Metric	78.4690 (23)***	71.227 (8.661)**	0.900	0.022	0.052	-0.004	rejected

Note. ΔCMIN: Sattora Bentler corrected; ***p < .001, **p < .01, *p < .05.

Our analyses reveals, that the concept of “attitudes towards democracies” exist across all groups. Nevertheless, inferential statistical conclusions cannot be made between the groups as long as metric invariance could not be established.

4.2.2. 4-Item Version (WVS)

Additional to the SOEP-Version of the scale, we also examine the 4-item version by WVS. Thus, we further exclude item 5 (“Minorities are protected.”). The absolute fit indices (CFI and RMSEA) are acceptable and thus, configural MI is given. Nevertheless, based on changes in Chi-Square and CFI, we have to reject the metric model again.

Table 8 Attitudes towards Democracy - 4 Item Version (WVS)

Model	CMIN (df)	ΔCMIN (Δdf)	CFI	ΔCFI	RMSEA Estimate	ΔRMSEA	
Configural Model	18.2740 (6)**		0.966		0.048		accepted
Metric Model	35.5790 (12)***	17.594 (16.334) **	0.934	0.032	0.047	-0.001	rejected

Note. ΔCMIN: Sattora Bentler corrected; ***p < .001, **p < .01, *p < .05.

In summary, the construct exists in all groups. The lack of comparability in response behavior between the groups in relation to higher levels of MI could be due to the different experiences of political situation in their countries of origin. In order to improve the MI, one suggestion would be to state even more clearly that respondents should indicate what should take place in an "ideal democracy". Another problem could be that respondents are afraid to express political attitudes (in a questionnaire) because of the political situation in their countries of origin.

4.3. Loneliness

With regard to the construct loneliness, the value of CFI is accaptalbe, whearas the Chi-Square is significant and RMSEA failed the cutt-off-points. Modicifcaion indices suggest that our model could be improved by allowing a shared covariance of item 2 and 3 for all three groups. With regard to the similarity of the content of the two itmes ("feel like an outsider" and "feel socially isolated", see Table 4), this is conceivable and thus we re-specified our model. In this model, the Chi-Square Test is not significant. Furthermore, CFI improved and the RMSEA is acceptable. Thus, the configural invariance is establised and we proceeded with our analyses by testing metric invariance for the re-specificed model. Changes of the Chi-Square are not significant. Additional, the changes in CFI and RMSEA are not significant. Thus, we evalute metric invariance as established and proceed testing the scalar invariance. In the scalar model, changes of all GOF statistics are remarcable and singificant, so that the scalar invarince is not exhibited.

Table 9 Loneliness

Model	CMIN (df)	ΔCMIN (Δdf)	CFI	ΔCFI	RMSEA Estimate	ΔRMSEA	
Configural Model	69.6880 (6)***		0.967		0.106		rejected
Configural ECOV 2/3	6.1010 (3)		0.998		0.033	-0.073	accepted
Metric ECOV 2/3	19.0360 (9)*	12.931 (6.255)	0.995	0.003	0.034	0.001	accepted
Scalar Model ECOV 2/3	198.3160 (15)***	197.600*** (5.254)	0.905	0.093	0.114	0.080	rejected

Note. ΔCMIN: Sattora Bentler corrected; ***$p < .001$, **$p < .01$, *$p < .05$.

To sum it up, our re-specified model leads to a reasonable model so that confgural invariance is partially established. For metric model estimated on the basis of re-specified configural model, the comparability of factor loadings is given. Hence, scalar invariance is strongly violated and comparison of latent or summarised mean cannot be assumed to be bias free.

4.4. Locus of Control

4.4.1. Locus of Control

With regard to the 7-Item-Version, which captures Locus of Control as unidimensional construct, neither Chi-Square Test, nor CFI or RMSEA are acceptable. Standardized factor loadings indicate, that Item 1 does not seem to be a good indicator for the latent construct – across all groups (standardized factor loading within the groups are -.051 (p>.05) for Afghans,

-.019 (p>.05) for Iraqis, and -.036 (p>.5) for Syrian). With regard to the content of the item ("My life's direction depends on me.") this result is quite unexpected as long as this item presents straightforward the core of the concept (see 3.2.4.).

The reason why exactly this item does not load on the construct might be related to our specific populations. All groups share the characteristic that they are refugees. Since flight is a life-changing and externally cause experience, it is reasonable that this item does not represent the same latent construct as items as i.e. "You must work hard to achieve success" within this population. Based on this considerations, we exclude Item 1. The Chi-Square Test is still significant in our re-specified model, but CFI and RMSEA improved. RMSEA is now acceptable. As long as CFI only slightly miss our cut-off points, we just accept this model and take configural invariance as partly established. Thus, we set out the metric model. Differences in Chi-Square are not significant. Additionally, changes in CFI and RMSEA are acceptable. This indicates that our metric model fits the data not worse than the configural one. Thus, we accept the metric model. Based on the results of Chi-Square difference Test, the significant difference in CFI and numerous indications to non-comparable intercepts (Appendix) we reject the Scalar Model for seven items representing a latent dimension.

Table 10 Locus of Control

Model	CMIN (df)	ΔCMIN (Δdf)	CFI	ΔCFI	RMSEA Estimate	ΔRMSEA	
Configural Model	712.888 (63)***		0.811		0.056		rejected
Configural Model without Item 1	74.0280 (27)***		0.920		0.043		Just accepted
Metric Model without Item 1	87.9780 (37)***	16.270 (13.697)	0.913	0.007	0.039	-0.004	accepted
Scalar Model without Item 1	138.3540 (47)***	53.282 (9.955)***	0.844	0.069	0.046	0.007	rejected

Note. ΔCMIN: Sattora Bentler corrected; ***p < .001, **p < .01, *p < .05.

4.4.2. External Locus of Control

Next, we intend to test the External Locus of Control Model independently. The value of RMSEA is good and as long as CFI misses our cut-off point only slightly, we just accept the configural model. Testing the metric model leads to significant increase of CMIN and big changes in CFI. Thus, we reject the metric model.

Table 11 External Locus of Control

Model	CMIN (df)	ΔCMIN (Δdf)	CFI	ΔCFI	RMSEA Estimate	ΔRMSEA	
Configural Model	33.6090 (15)**		0.942		0.037		Just accepted
Metric Model	64.9630 (23)***	30.997 (9.016)***	0.870	.072	0.044	.007	rejected

Note. ΔCMIN: Sattora Bentler corrected; ***p < .001, **p < .01, *p < .05.

4.4.3. Internal Locus of Control

Last but not least, we set out the 5-item solution which captures Internal Locus of Control as a distinct construct. The Chi-Square Test was significant and CFI failed the cut-off points for goodness of fit. Thus, we have to reject this model even though RMSEA point to acceptable fit. Standardized factor loadings revealed, that Item 2 does not seem to be a good indicator for the latent construct (factor loading within the groups are .088 (p>.05) for Afghans, .076 (p>0.05) for Iraquis, and .012 (p>.5) for Syrian). Comparing this items with the others shows that it is reversed. Thus, we decided to exclude this item. With our new model, we achieved perfect fit for CFI and RMSEA and non significant result for the Chi-Square Test. Thus, configural invariance is established. For the metric model, changes in Chi-Square are significant. Furthermore, the changes in CFI and RMSEA are too strong. Thus, we had to reject to metric model.

Table 12 Internal Locus of Control

Model	CMIN (df)	ΔCMIN (Δdf)	CFI	ΔCFI	RMSEA Estimate	ΔRMSEA	
Configural Model	62.0230 (15)***		.742		0.058		rejected
Configural Model without Item 2	3.7430 (6)		1.000		0.000		accepted
Metric Model without Item 2	17.9140*** (12)	14.047 (7.738)**	.955	.045	0.023	-.023	rejected

Note. ΔCMIN: Sattora Bentler corrected; ***p < .001, **p < .01, *p < .05.

5. Discussion

Based on data from the IAB-BAMF-SOEP refugee sample, we tested measurement invariance between refugees from Syria, Afghanistan and Iraque of the four constructs Resilience, Attitudes towards Democracy, Loneliness and Locus of Control. As long as different factorial structures were proposed within the literature for some of the constructs, we set out not only four, but seven models with different factorial structures: one for resilience and one loneliness, two for attitudes towards democracy and three for the concept Locus of Control. For these seven models, even the lowest level of MI (configural), was established only for the Brief-Resilience-Coping-Scale, the WVS-Version of Attitudes towards Democracy and External Locus of Control. Thus, only for these three models, we can assume the existence of comparable concepts across the refugee populations. With regard to the other models, we had to reject the comparability of the concepts between our groups. Metric invariance was only established for the BRCS and had to be rejected for Attitudes towards Democracy. Thus, only for BRCS, meaningful comparisons of correlations can be made. We could not establish scalar invariance for any of the models derived through literature. Thus, comparisons of means amongst the three major refugee groups in Germany, as they are often used within analyses, might not be valid for any of the instruments

We examined non significant factor loadings and modification indices to detect potential causes of the lacking invariance based on our data and re-specified some of the configural models. By this procedure, we were able to establish configural invariance for the modified models for all scales. Causes and thus re-specifications were different and comprises the following. With regard to Loneliness, we allowed a correlated covariance between two items, based on modification indices and a content-related considerations. For the SOEP-version of Attitudes towards democracy, Locus of Control and Internal Locus of Control, we excluded items with low or non-significant factor loadings. With regard to Attitudes towards democracies and Locus of Control, we assume that the content of the respective item is not appropriate for the latent construct within the refugee populations. For the removed item of the Internal Locus of Control Scale, we assume a methodological cause. This is because the problematic item is reversed in comparison with all other items. Surprisingly, the reversely coded item exhibited problems within the Internal Locus of Control Scale, but not in the entire Locus of Control Scale. Vice versa, the item which we had to exclude from the global Locus of Control Scale, were not associated with anomalies within the Internal Locus of Control Scale.

The problems with respect to the configural models are rather general and comparable across all refugee groups. Thus, there has been no specific problems caused by language or culture which resulted in common re-specifications in all groups. The models derived through the literature do not suit every group in comparable manner. A reason might be, that all the constructs were developed and validated within Western, educated, industrialized, rich and democratic societies (Henrich et al., 2010) and thus are not easily applicable to other contexts or populations. As long as we had no appropriate control group to test this hypothesis, it should be addressed in further research. Additionally, based on the data structure, we were able to discover if certain items do not fit the construct. An analysis of whether additional indicators

should be included to more suitable represent the construct within the populations cannot be performed through MI analyses. This should be addressed in further research.

To sum it up, the IAB-BAMF-SOEP is one of the important data sources for the research on refugees in Germany. Established constructs from various scientific disciplines are surveyed. The data is freely available and thus often used by researcher. Results of analyses are used for as basis for political decisions. As our analyses show however that results for comparisons when using these data might be limited doe the bias in measurement.

6. Literature

Bentler, P. M. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences*, *42*(5), 825–829. https://doi.org/10.1016/j.paid.2006.09.024

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural equation modeling, 12*(1), 41-75.

Brücker, H., Rother, N., Schupp, J., Babka von Gostomski, C., Böhm, A., Fendel, T., Friedrich, M., Giesselmann, M., Kosyakova, Y., Kroh, M., Kühne, S., Liebau, E., Richter, D., Romiti, A., Schacht, D., Scheible, J., Schmelzer, P., Siegert, M. Sirries, S., ... Vallizadeh, E. (2016). *IAB-BAMF-SOEP Refugee Survey: Forced migration, arrival in Germany, and first steps toward integration*. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-67552-6

Byrne, B. (2011). *Structural Equation Modeling with Mplus: Basic Concepts, Applications, and Programming (Multivariate Applications)*. Taylor & Francis.

Campbell-Sills, L., Liverant, G. I., & Brown, T. A. (2004). Psychometric evaluation of the behavioral inhibition/behavioral activation scales in a large sample of outpatients with anxiety and mood disorders. *Psychological Assessment*, *16*(3), 244–254. https://doi.org/10.1037/1040-3590.16.3.244

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

Esser, H. (2006). *Sprache und Integration*. Campus Verlag.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hox, J. J., De Leeuw, E. D., & Zijlmans, E. A. (2015). Measurement equivalence in mixed mode surveys. *Frontiers in psychology, 6*, 87.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*(2), 219–30. https://doi.org/10.1007/s11205-011-9843-4

Hughes, M. E., Waite, L. J., Hawkley, L. C., & Cacioppo, J. T. (2004). A Short Scale for Measuring Loneliness in Large Surveys: Results from Two Population-Based Studies. *Research on Aging, 26*(6), 655–672. https://doi.org/10.1177/0164027504268574

Jacobsen, J., Klikar, J., & Schupp, J. (2017). Scales Manual IAB-BAMF-SOEP Survey of Refugees in Germany – revised version. *SOEP Survey Papers, 475*(Series C). DIW/SOEP.

Jöreskog, K. G. (1971). Statistical Analysis of Sets of Congeneric Tests. *Psychometrika 36,* 109-133.

Jowell R. (1998). How Comparative Is Comparative Research? *American Behavioral Scientist*, *42*(2), 168–177. https://doi.org/10.1177/0002764298042002004

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. Guilford Press.

Kroh, M., Brücker, H., Kühne, S., Liebau, E., Schupp, J., Siegert, M., & Trübswetter, P. (2016). Das Studiendesign der IAB-BAMF-SOEP-Befragung von Geflüchteten. In H. Brücker, N. Rother,

& J. Schupp (Eds.), IAB-BAMF-SOEP-Befragung von Geflüchteten: Überblick und erste Ergebnisse, IAB Forschungsbericht (4)

Kocalevent, R. D., Zenger M., Heinen, I., Dwinger S., Decker, O., Brähler, E. (2015). Resilience in the General Population: Standardization of the Resilience Scale (RS-11). *PLoS ONE 10*(11), e0140322. https://doi.org/10.1371/journal.pone.0140322

Luhmann, M., & Hawkley, L. C. (2016). Age differences in loneliness from late adolescence to oldest old age. *Developmental Psychology*, *52*(6), 943–959. https://doi.org/10.1037/dev0000117

Markus, K. A. & Borsboom, D. (2013). *Frontiers of Test Validity Theory. Measurement, Causation, and Meaning*. Routledge.

McNeish, D., An, J., & Hancock, G. R. (2018). The Thorny Relation Between Measurement Quality and Fit Index Cutoffs in Latent Variable Models. *Journal of Personality Assessment*, *100*(1), 43–52. https://doi.org/10.1080/00223891.2017.1281286

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, *29*(3), 223-236.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543.

Metzing, M., Schacht, D., & Scherz, A. (2020). Psychische und körperliche Gesundheit von Geflüchteten im Vergleich zu anderen Bevölkerungsgruppen. *DIW Wochenbericht*, *87*(5), 63–72. https://doi.org/10.18723/diw_wb:2020-5-1

Meuleman, B., Żółtak, T., Pokropek, A., Davidov, E., Muthén, B., Oberski, D. L., Billiet, J., & Schmidt, P. (2022). Why Measurement Invariance is Important in Comparative Research. A Response to Welzel et al. (2021). *Sociological Methods & Research, 0*(0). https://doi.org/10.1177/00491241221091755

Millsap R. E. (2011). *Statistical approaches to measurement invariance.* Routledge.

Missirian, A., & Schlenker, W. (2017). Asylum applications respond to temperature fluctuations. *Science, 358*(6370), 1610-1614. https://doi.org/10.1126/science.aao0432

Muthén, L. K., & Muthen, B. (2017). *Mplus user's guide: Statistical analysis with latent variables, user's guide.* Muthén & Muthén.

Nolte, H., Weischer, C., Wilkesmann, U., Maetzel, J., & Tegethoff, H. G. (1997). Kontrolleinstellungen zum Leben und zur Zukunft. Auswertung eines neuen, sozialpsychologischen Itemblocks im Soziooekonomischen Panel. Diskussionspapiere Aus der Fakultät für Sozialwissenschaft Ruhr-Universität Bochum 97 – 06, ISSN 0943 - 6790

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

Roberts, C., Sarrasin, O., & Ernst Stähli, M. (2020). Investigating the Relative Impact of Different Sources of Measurement Non-Equivalence in Comparative Surveys: An Illustration with Scale Format, Data Collection Mode and Cross-National Variations. *Survey Research Methods, 14*(4), 399–415. https://doi.org/10.18148/srm/2020.v14i4.7416

Rotter, J. B. (1996). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied, 80*, 1-28.

Satorra, A., & Bentler, P. M. (2010). Ensuring possitiveness of the scaled chi-square test statistic. *Psychometrika*, *75*(2), 243–248.

Scheuch, E. K. (1993). The cross-cultural use of sample surveys: problems of comparability. *Historical Social Research, 18*(2), 104-138. https://doi.org/10.12759/hsr.18.1993.2.104-138

Schröder, H., Zok, K., & Faulbaum, F. (2018). Gesundheit von Geflüchteten in Deutschland–Ergebnisse einer Befragung von Schutzsuchenden aus Syrien, Irak und Afghanistan. *Wido Monit, 15*(1), 1-20. Sinclair, V.G., & Wallston, K.A. (2004). The Development and Psychometric Evaluation of the Brief Resilient Coping Scale. *Assessment, 11*(1), 94-101. https://doi.org/10.1177/1073191103258144

Specht, J., Egloff, B., & Schmukle, S. C. (2013). Everything under control? The effects of age, gender, and education on trajectories of perceived control in a nationally representative German sample. *Developmental Psychology, 49*, 353-364.

Stathopoulou, T., Krajčeva, E., Menold, N., & Dept, S. (2019). Questionnaire design and translation for refugee populations: Lessons learnt from the REHEAL Study. *Journal of Refugee Studies, 32(Special_Issue_1)*, i105-i121. https://doi.org/10.1093/jrs/fez045

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.

Weiber, R., & Mühlhaus, D. (2014). *Strukturgleichungsmodellierung. Eine anwendungsorientierte Einführung in die Kausalanalyse mit Hilfe von AMOS, SmartPLS und SPSS* (2nd ed.). Springer Gabler.

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, *51*, 409–428. https://doi.org/10.3758/s13428-018-1055-2

Appendix

Standardized Model Parameters of the Configural Model by Group

	Afghanistan		*Iraq*		*Syria*	
Indicators	*λ*	*τ*	*λ*	*τ*	*λ*	*τ*
BRCS-4S						
1	.595***	**5.983*** **	.549***	4.967***	.538***	5.213***
2	.456***	**4.762*** **	.382***	4.055***	.405***	**4.001*** **
3	.868***	5.304***	.712***	5.502***	.741***	5.412***
4	.746***	**5.835*** **	.660***	5.153***	.597***	**5.935*** **
AtD						
1	.571***	1.997***	.285***	4.098***	**.246*** **	3.448***
2	**.428*** **	**5.167*** **	**.675*** **	9.128***	.696***	10.113***
3	.511***	3226***	.620***	7.314***	.657***	6.556***
4	**.722*** **	**2.425*** **	.512***	7.635***	**.455*** **	7.438***
6	.689***	**2.626*** **	**.691*** **	10.979***	.719***	8.625***
AtD WVS						
1	.661***	2.004***	.316***	**4.095*** **	.222***	3.444***
2	.467***	**5.170*** **	.718***	9.120***	.810***	10.157***
3	.452***	3.263***	.631***	7.270***	.626***	6.553***
4	.649***	**2.428*** **	.451***	7.622***	.361***	7.439***
LONE						
1	.901***	**2.099*** **	.930***	2.290***	.872***	2.266***
s2	.580***	2.598***	.629***	2.784***	.595***	2.859***

3	**.407*** **	**2.483*** **	.557***	2.888***	.553***	**2.951*** **
4	.345***	**2.024*** **	.346***	**1.821*** **	**.252*** **	**1.701*** **
LOC						
2	.275***	**2.862*** **	.307**	**2.900*** **	**.335*** **	**2.868*** **
6	.358***	**2.227*** **	.395***	**2.009*** **	**.262*** **	**1.986*** **
7	.383***	**2.480*** **	.561***	**2.278*** **	**.358*** **	**2.036*** **
8	.465***	**1.493*** **	.149	**1.223*** **	**.356*** **	**1.220*** **
9	.597***	**1.981*** **	.415**	**1.668*** **	**.499*** **	**1.534*** **
10	.632***	**2.038*** **	.314*	**1.941*** **	**.408*** **	**1.862*** **
LoCe						
6	.393***	2.233***	.386***	2.010***	.311***	1.986***
7	.350***	2.482***	**.679*** **	2.281***	.400***	2.035***
8	**.499*** **	1.489***	**.080**	1.223***	.315***	1.219***
9	.279**	2.862***	.336***	2.900***	.358***	2.868***
10	.611***	2.037***	.195*	1.941***	.351***	1.862***
LoCi						
1	**.427** **	5.414***	.706***	5.422***	**.853*** **	5.631***
3	.365**	6.804***	.586***	6.392***	.285***	6.494***
4	.967***	5.879***	.493*	5.119***	.853***	5.320***
5	.710***	5986***	.562**	5.273***	.319***	5.200***

Note. λ loadings; τ intercepts; f: Factor; &: correlation; ***p < .001, **p < .01, *p < .05. Bold: equality of parameters is associated with MIs > 3.84 (Critical Value of CMIN for df = 1).

Notes:

Funding: funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – FOR 2928 PH-LENS

Conflict of interest: Authors state no conflict of interest.

Ethical statement: Neither primary data for human nor for animals were collected for this research work.