Menold, Natalja

Article — Published Version

Effect of Cognitive Pretests on Measurement Invariance and Reliability in Quality of Life Measures: An Evaluation in Refugee Studies

Social Indicators Research

Provided in Cooperation with:
Springer Nature

Suggested Citation: Menold, Natalja (2025) : Effect of Cognitive Pretests on Measurement Invariance and Reliability in Quality of Life Measures: An Evaluation in Refugee Studies, Social Indicators Research, ISSN 1573-0921, Springer Netherlands, Dordrecht, Vol. 179, Iss. 1, pp. 527-548, https://doi.org/10.1007/s11205-025-03622-w

This Version is available at:
https://hdl.handle.net/10419/330607

Social Indicators Research (2025) 179:527–548
https://doi.org/10.1007/s11205-025-03622-w

ORIGINAL RESEARCH

Effect of Cognitive Pretests on Measurement Invariance and Reliability in Quality of Life Measures: An Evaluation in Refugee Studies

Natalja Menold[1]

Accepted: 3 May 2025 / Published online: 24 May 2025

Abstract
The issue evaluated was how cross-cultural cognitive pretests affect the reliability and comparability of data when studying refugees and using cross-language comparisons. Three instruments employed to assess general and health-related quality of life were revised based on the findings of cognitive pretests. The versions before and after cognitive pretests were randomly assigned to respondents in two web survey studies. The first study recruited Arabic and Dari speaking refugees ($N=3448$) via Facebook. The second study used a random sample ($N=610$) of Arabic-speaking refugees and the German-speaking population. The comparability of data was evaluated by means of two popular measurement invariance analysis methods (exact and Bayesian measurement invariance analysis). These methods rely on factor analysis and thus enable the evaluation of the comparability of factor structure (configural invariance), the loadings or weights an item has on the factor (metric invariance), and the intercepts of an item on latent factor (scalar invariance). The revision of instruments on the basis of cognitive pretests primarily yielded positive outcomes with regard to reliability and cross-language comparability with respect to the factorial structure and loadings. However, the outcomes pertaining to scalar invariance were more equivocal. Cognitive pretests can therefore help to improve measurement quality, but further research is needed on the cross-cultural questionnaire design to make revisions with respect to cross-cultural comparability more conclusive.

Keywords Methodology of cross-cultural measurement · Refugees · Comparability bias · Measurement invariance · Composite reliability

1 Introduction

Comparisons of concepts between different groups, e.g. by country, nationality, gender, age or other defined groups, are essential for the research in quality of life. Studying the effects of migration on quality of life and health is a relevant research area (e.g., Di Barbiano

✉ Natalja Menold
natalja.menold@tu-dresden.de

[1] Institute of Sociology, Dresden University of Technology, D-01062 Dresden, Germany

Belgiojoso et al., 2022; Feliciano, 2020; Giovanis, 2022), where language, cultural and cognitive aspects need to be taken into account in measurement instruments. Not only do instruments frequently require adaptation to the languages of the refugees. Additionally, different patterns of understanding questions and responding to the subjective well-being measures has the potential to introduce bias in the comparisons. The present article focuses on data presumptions of the measurement in comparative research with the aim of evaluating how data quality can be improved when piloting measurement instruments for cross-language comparisons.

Which requirements must data fulfil in order to facilitate valid comparisons? The primary requirement for measurement is reliability, which can be described as the degree to which unsystematic measurement errors (or random fluctuation) are small and can be disregarded. Reliability is defined as the proportion of true variance to the obtained variance (e.g. Lord & Novick, 1968; Raykov & Marcoulides, 2011). Non-reliable measurements can be compared to measurements made with a length of rubber band: a method that makes it impossible to identify true differences.

The second requirement is that measurement in the groups being compared is parallel – that is free of comparability bias. The comparability bias in the data is defined as different responses provided by the respondents at the same level on the concept being measured (van de Vijver, 2018). Comparability bias in the measurement of general and health related quality of life when studying refugees may result from different patterns of symptom expression, different attributions of the causes of disease, e.g. as magic or as punishment for sins, and from different knowledge of and experiences with the health system and medical treatment (Kiesel, 1995).

Cognitive pretests were suggested as a method to increase reliability (Rammstedt et al., 2015), or to prevent cross-cultural bias at the questionnaire development, translation and adaptation stage (Willis, 2015; Willis & Miller, 2011). Cognitive pretests evaluate the load on cognitive response processes in terms of understanding questions, retrieving relevant information, forming judgements und providing responses (Tourangeau et al., 2000). Qualitative data are analysed to identify problematic wording, overly difficult questions or confusing parts of a questionnaire. Although cross-cultural cognitive pretests were proposed as a method to prevent cross-cultural bias (Willis, 2015; Willis & Miller, 2011), these were often not used in the cross-cultural research. Moreover, as cognitive pretests require additional budget and time to administer, there should be evidence of their positive impact on data quality. However, there is a paucity of studies that evaluate cognitive pretests for their potential to reduce comparability bias.

We address the question of the impact of cognitive pretests on the two prerequisites – high reliability and data comparability – for cross-cultural, cross-language measurement. We focus on studying refugees and use health-related and general Quality of Life (QoL) instruments in German and their translations into Arabic and Dari. To evaluate the effect of cognitive pretests, an experimental comparison was made of the instrument versions before and after the cognitive pretests. The present article has implications for research and practice, as it presents a study on evaluation of a method that has been proposed to improve data quality for cross-language comparisons, but which has rarely been evaluated to achieve this goal.

In the following, we review the theoretical and empirical background to the study of comparability bias and discuss the research that has evaluated or used cognitive pretests. We formulate our research question and hypotheses on this basis. Further sections describe the method used to evaluate the effect of cognitive pretests on reliability and comparability bias and the results obtained. The final section discusses the results and draws conclusions.

1.1 Comparability Bias in Cross-Cultural Research

According to van de Vijver (e.g., 2018), comparability bias can be due to different concepts being measured in different groups (construct bias), to methods that introduce incomparability (method bias), and finally to the formulation of single items that produces systematic differences in meaning for individuals from different groups (item bias). Examples of concept bias would be cross-cultural differences in the definitions of future, neighbourhood, family or household (Benítez et al., 2022; Hoffmeyer-Zlotnik & Warner, 2014; Scheuch, 1993). An example of method bias would be different response formats or visual layouts that may lead to different response behaviours (Höhne et al., 2021). An example of item bias was provided by Meitinger (2017), who found that the term "social security system" as used in the International Social Survey Program (ISSP) was understood as pension insurance in the U.S. but as health insurance in Spain and as welfare in Germany.

Research on comparability bias uses a relatively new method – so called measurement invariance analysis. This method is briefly introduced here to provide the reader with the background necessary to understand the research on comparability bias. More details are provided in the method section. Measurement invariance analysis has been proposed in the context of a modern measurement theory called Latent Variable Measurement (Muthén, 2002). Within this framework, confirmatory factor analysis (CFA) is used to specify a measurement model that allows measurement properties and substantive measurement results to be evaluated separately. Measurement invariance (alternatively referred to as invariance in the remainder of this paper) is defined as psychometric equivalence and ensures that individuals at the same level on the latent variable perceive and respond to the corresponding survey questions in the same way (Meade & Lautenschlager, 2004).

Invariance analysis has become increasingly popular and is used in the development or adaptation of instruments. However, especially in cross-national or cross-language comparisons, limited invariance has been reported in numerous studies (Davidov et al., 2008; Dong & Dumas, 2020; Wu et al., 2007; Zercher et al., 2015), demonstrating the reduced parallelism of translations and adaptations in many large-scale survey data. It appears to be particularly difficult to obtain invariant data for groups defined by ethnicity. For example, in their meta-analysis of measurement invariance of personality measures, Dong and Dumas (2020) found that it was largely not achieved when comparing ethnic groups. Similarly, limited invariance has been found in the adaptations of instruments to refugee populations in the German Socio-Economic Panel (GSOEP) (Tibubos & Kröger, 2020). These results show that comparability bias can be a serious problem in cross-cultural research. That the results of measurement invariance analyses are indeed related to potential comparability bias has been shown in simulation studies (Muthén & Asparouhov, 2012; Pokropek et al., 2019) and in mixed-method studies in which cognitive pretests were contrasted with the results of measurement invariance analyses (Benítez et al., 2022; Meitinger, 2017). Therefore, methods are needed to improve the comparability of measurements, and the author evaluates cognitive pretesting as such method.

1.2 Research on the Impact of Cognitive Pretests on Data Quality

Cognitive interviewing is a qualitative method of questionnaire development and piloting. Probes have often been used in cognitive pretests, i.e. follow-up questions about how

questions were understood, or questions asking for explanations of how a respondent understood a term or why he/she gave a particular answer.

Various researchers have found cognitive interviews to be conclusive in terms of identifying cognitive response problems (Lenzner & Neuert, 2017; Meitinger & Behr, 2016; Rothgeb et al., 2007). Some studies have also evaluated the effect of cognitive pretests on reliability and validity. Validity is defined as empirical evidence that an instrument actually measures the concept under investigation (see e.g., Rammstedt et al., 2015). Yan et al. (2012) compared cognitive interviews with several other methods (expert reviews, the Survey Quality Predictor (SQP), error rates from latent class analysis) and found that cognitive pretests best predicted validity. Maitland and Presser (2018) compared the Question Understanding AID, SQP, expert reviews, checklists, and cognitive interviews and found expert reviews and cognitive interviews to be the most predictive in terms of identifying problematic questions or reliability issues.

Benítez et al. (2022) evaluated the measurement invariance of life satisfaction for Dutch and Spanish in several large-scale cross-cultural population surveys and analysed non-invariant questions using cognitive pretests. The results explain non-invariance by differences in understanding of items and concepts. Similar results were obtained by Meitinger (2017), who used web probing (an adaptation of cognitive pretests to the web situation) and measurement invariance analysis for selected questions in the ISSP. However, these studies were not designed to show whether it is possible to improve the instruments on the basis of findings of the cognitive pretests.

The potential effect of the cognitive pretests and other piloting methods (web probing and expert review) on the comparability bias and reliability was evaluated by Menold et al. (2025). The results were mixed, particularly with respect to the cognitive pretests, which sometimes reduced data comparability. However, the authors used opinion instruments with some very low initial reliability, which could be systematically improved by means of piloting methods. In addition, the results were limited to German and U.S. English and did not take into account questionnaire adaptations for refugee groups. Nevertheless, the study shows that questionnaire piloting methods, including cognitive pretests, had a positive effect on the reliability of instruments where it was initially insufficient.

1.3 Research Question and Hypotheses

The research question is how cognitive pretests impact on measurement quality when using multi-language measures on the QoL. As the reliability of an instrument should be sufficiently high to meet the aim of comparability, it is treated as a quality criterion. Cognitive pretests are expected to increase reliability and have been shown to do so in some previous research. The author expects to replicate this result as follows:

> *Hypothesis H1*: Instruments that have been revised on the basis of cognitive pretests show higher reliability than instruments before the revision.

The next expectation is that the revision of the instruments to address the problems found in the cognitive pretests will reduce the comparability bias. This is the issue addressed by the second hypothesis:

Hypothesis H2: The results of the measurement invariance analysis will be more tenable when the instruments are revised based on cognitive pretests than when versions of the instruments before the revisions are used.

2 Method

To respond to the research question and to evaluate the hypotheses, an experimental study was carried out to compare the versions of instruments before (before CP) and after the cognitive pretests (after CP). We used three instruments relevant to refugees' health and quality of life that are included in the GSOEP refugee samples and for which Arabic, Farsi and German versions were available (Andersen et al., 2007; Tibubos and Kröger, 2020). Table 1 provides an overview of the instruments before CP. Online Source S1-S3 provides full information about instruments before and after CP. We adapted the Farsi versions to Dari using two professional translators and the Translation, Review, Adjudication, Pretesting, and Documentation (TRAPD) approach (Harkness, 2003). This method relies on a team of interpreters translating questionnaires independently, discussing the differences in a facilitated meeting and agreeing on the optimal translations. A professional cognitive pretest laboratory in Germany at the GESIS-Leibniz Institute for the Social Sciences was contracted to conduct cognitive pretests for Arabic and Dari versions and to provide the research team with a detailed protocol of the results. In addition, the pretest report was published by GESIS as conductor (Hadler et al., 2021). The author did not conduct the cognitive pretests and did not provide findings to increase the objectivity of the evaluation of cognitive pretests.

2.1 Instruments

2.1.1 Short Form Health Survey-12 (SF-12)

The first instrument was the Short Form Health Survey-12 (SF-12, 12 items, Ware et al., 1996). The concept covered is health-related quality of life, consisting of physical (PH) and mental (MH) health summaries, six items each (e.g., Ware et al., 1996; Table 1). The GSOEP version differs slightly from the SF-12 by Ware et al. (1996) in that it includes an additional indicator (rushed (SF-10 in Table 1) (cf. Tibubos & Kröger, 2020). The present research used all 13 GSOEP indicators.[1] The factorial structure of PH and MH is not very clear, because, in the literature, the general health indicator is assigned to PH and the item on restricted social activities due to health problems is assigned to MH (Ware et al., 1996; Tibubos and Kröger, 2020). The reliability (test–retest, composite reliability and Cronbach's Alpha) and validity of SF-12 (factorial, construct and criterion) were found by previous work to be satisfactory (e.g., Fong et al., 2010; Ware et al., 1996). However, some previous studies indicated potential problems in measurement invariance, as the factorial structure of SF-12 has been found to vary across language and ethnic groups (Fleishman & Lawrence, 2003; Fong et al., 2010; Tibubos and Kröger, 2020). Other studies documented non-invariance for the SF-12 for age and ethnic groups (Desouky et al., 2013; Fleishman

[1] The general health indicator was administered by respondents (also after CP), but was not included in the scoring in the present research, as it cannot be univocally assigned to either PH or MH.

Table 1 Instruments before Cogntive Pretests (CP)

Item	Wording
SF-12	
SF-0 overall	How would you describe your current state of health?
SF-1 pain	How often in the last four weeks, did you suffer from severe physical pain?
SF-2 stairs	If you have to climb stairs, i.e. walk up several floors: Does your state of health restrict you a lot, a little, or not at all?
SF-3 hinder	And what about other strenuous activities in everyday life, e.g. when you have to lift something heavy or need to be mobile: Does your state of health restrict you a lot, a little or not at all?
SF-4 less done physical	How often in the last four weeks, due to health problems of a physical nature, did you achieve less in your work or everyday activities than you actually intended?
SF-5 less content physical	How often in the last four weeks, due to health problems of a physical nature, have you been restricted in the type of tasks you can perform in your work or everyday activities?
SF-6 energy	How often in the last four weeks did you feel full of energy?
SF-7 relax	How often in the last four weeks did you feel calm and balanced?
SF-8 less social	How often in the last four weeks, due to health or psychological problems, have you been restricted in terms of your social contact to for example friends, acquaintances or relatives?
SF-9 down	How often in the last four weeks, did you feel in low spirits and melancholy
SF-10 rushed	How often in the last four weeks, did you felt rushed and under time pressure?
SF-11 less done mental	How often in the last four weeks, due to psychological or emotional problems, did you achieve less in your work or everyday activities than you actually intended?
SF-12 less content mental	How often in the last four weeks, due to psychological problems or emotional problems, did you perform your work or everyday activities less carefully than usual?
RHS-15	
	Please indicate the degree to which you had a symptom over the past month
1	Muscle, bone, joint pains
2	Feeling down, sad, or blue most of the time
3	Too much thinking or too many thoughts
4	Feeling helpless
5	Suddenly scared for no reason
6	Faintness, dizziness, or weakness
7	Nervousness or shakiness inside
8	Feeling restless, can't sit still
9	Must cry easily
	The following symptoms may be related to traumatic experiences during war and migration. How much in the past month have you
10	had the experience of reliving the trauma; acting or feeling as if it were happening again
11	Been having body reactions (for example, break out in a sweat, heart beats fast) when reminded of the trauma
12	Felt emotionally numb (for example, feel sad but can't cry, unable to have loving feelings)

Table 1 (continued)

Item	Wording
13	Been jumpier, more easily startled (for example, when someone walks up behind you)
14	Please provide for each oft he following statements how much it does apply to you I am feeling that I am able to handle (cope with) anything that comes my way I am feeling that I am able to handle (cope with) most things that come my way I am feeling that I am able handle (cope with) some things, but not able to cope with other things I am feeling that I am unable to cope with most things I am feeling that I am unable to cope with anything
15	*Distress Thermometer* Please provide how much distress you have been experiencing in the past week including today (0–100)
LS	
	How satisfied are currently with the following aspects of your life?
1	How satisfied are you with your life in general?
2	How satisfied are you with your health?
3	How satisfied are you currently with your personal income?
4	How satisfied are you in general with your current living arrangements?

Source for English wording Jacobsen et al. (2017)

and Lawrence, 2003) and suggested cognitive pretests to evaluate potential reasons for non-invariance.

2.1.2 Refugee Health Screener–15 (RHS-15)

Refugee Health Screener–15 (RHS-15) is a screening instrument for emotional distress and potential posttraumatic stress disorder in refugee groups (Hollifield et al., 2016). RHS-15 consists of 13 symptom items, 1 coping item and a distress thermometer (Table 1). RHS-15 has a one-factor structure, high to excellent reliability (Cronbach's Alpha between 0.80 and 0.95) as well as acceptable to excellent criterion validity (e.g., Borho et al., 2022; Hollifield et al., 2016). However, there were differences in the factor structure between refugee groups, and respondents also had some understanding problems (Hollifield et al., 2016). No studies were available on the measurement invariance of the RHS-15.

2.1.3 Life Satisfaction (LS)

The third instrument was the Satisfaction with Life Scale (LS, Table 1), which is a cognitive measure of individual well-being. The GSEOP instrument selected for the study includes a general life satisfaction item and three domain-specific satisfaction indicators (Schimmack et al., 2009). The one-dimension instrument aims to capture an evaluation of life that corresponds to a standard or an individual ideal (Schimmack et al., 2009). Test–retest reliability was found to be sufficient in the GSOEP (0.77, Richter et al., 2017). Studies provide indications of construct and criterion validity of

the German version (Schimmack et al., 2009). However, with respect to cross-cultural comparability, Benítez et al. (2022) found a high-degree of non-invariance for the general life satisfaction item, which is also included in the evaluated instrument.

2.1.4 Revisions using Results of Cognitive Pretests

The cognitive pretests were carried out in Arabic and Dari with 18 refugees (Syrian, Iraqi and Afghan) of different genders, age and education (Hadler et al., 2021). In general, instructions and selected items were explicitly examined rather than the whole instrument. The introduction and seven items were tested for the SF-12, five items were tested for the RHS-15, and two items were tested for the LS. Emerging probing was also used, as the problems and comments on the non-purposively probed items were elicited, documented and analysed. For further details of the cognitive pretests, see Hadler et al. (2021).

An overview of cognitive problems described by Hadler et al. (2021) and of revised instruments after CP is provided in Online Source S1-S3. The cognitive pretests for the SF-12 showed that respondents were confused about whether an item related to physical, mental or general health. Another difficulty with this instrument, but also in RHS-15, was the lack of clarity in the time references. Refugees also tended to understand "everyday activities" in the SF-12 as professional work. In the SF-12, the different number of response categories, with three categories for some indicators and five for others, also led to confusion. In the SF-12 and particularly in the RHS-15, some adjectives for feeling sad or depressed were not understood by refugees as symptoms, e.g., "to feel numb" or "to cry easily". Respondents were also confused by questions that contained enumerations or more than one stimulus to be evaluated in one question (Double Barreled Questions or DBQs).

For the LS, the two tested items (general life satisfaction and satisfaction with living situation) were found to be translated too similarly. The term "living arrangements" was taken to mean the entire living situation, so that respondents – in both languages—could not distinguish between these items.

The instruments were revised on the basis of suggestions made by the GESIS cognitive pretest laboratory (Hadler et al., 2021) and taking into account the cognitive problems described in the report. The revisions are documented in Online Source S1-S3. These were particularly pronounced for the SF-12 and RHS-15, but less so for the LS. In the SF-12, additional indicators were developed to balance physical and mental health and to split DBQs. Time references were highlighted. To handle the problem of adjectives as emotional symptoms in SF-12 and RHS-15, several alternatives for depression or pain indicators were provided. For this, we used hints from the cognitive interviews and incorporated wording provided by interviewees. In addition, response options were unified in the SF-12. The revisions for the LS involved including a definition of "living arrangements" and were therefore minor.

The revised version of the SF-12 consisted of 16 items, 12 of which had to be selected, which was implemented by the subsequent measurement invariance analyses (see Section Data Analysis). The revised version of the RHS-15 consisted of 16 indicators with room for 15 indicators to be selected depending on the results of subsequent quantitative analyses. All revised versions were provided in German and the adapted parts were translated into Arabic and Dari by professional translators using the TRAPD method.

2.2 Experimental Design and Samples

The versions before and after CP were evaluated in two different studies using random assignment. The studies were conducted in 2022 and 2023 in Germany. The questionnaires were available to a language group in the respective language only.

In the first study (Study 1), Arabic-speaking Syrian and Iraqi refugees and Dari-speaking Afghan refugees were recruited via Facebook to take part in an online survey. A screening process was carried out to ensure that the participants were refugees who spoke these languages. The realized sample (N= 3448) consisted of N= 1925 Arabic and N= 1523 Dari-speaking refugees (Online Source Table S9 for sample statistics). Depending on the instrument and due to data missing, n = 520 to n = 680 Arabic-speaking respondents vs. n = 331 to n = 519 Dari-speaking respondents used the versions before CP; The versions after CP were used by n = 527 to n = 772 Arabic-speaking vs. by n = 319 to n = 570 Dari-speaking individuals.

In the second study (Study 2) a probability sample was recruited from the German-speaking population, Arabic-speaking Syrian and Iraqi refugees and Dari-speaking Afghan refugees in three large cities of a German federal state. The addresses of sampled persons, their gender, age and country of origin were provided by the municipal registration offices. We implemented a tailored design method to increase response rates (Dillman et al., 2014). This involved using up to three postal contacts and administering a web survey. A link and a QR code leading to the survey were provided in the mail invitation and the reminders. Response rates (RR6, AAPOR, 2023) were 28.42% for German-speaking residents (N= 270) and 19% for Arabic-speaking refuges (N= 340). Response rates for Dari-speaking refugees were 16% (N= 141). The net sample size for Arabic- and German-speaking residents was N= 610 (Online Source Table S10 shows sample statistics). In the realized net sample, there were no significant deviations from the overall provided sample with regard to gender and age. This means that non-participants did not differ from participants in age and gender and therefore the realized sample was not biased with respect to these characteristics.

Instruments before CP were used by n= 119 to n = 132 German- and n= 142 to n =148 Arabic-speaking respondents. After CP, the samples for included insturments were composed of n= 122 to n =129 German- and n= 142 to n =158 Arabic-speaking individuals. The sample sizes varied due to data missing. For Dari, the group of respondents who used either version was too small (maximum n= 57 and n= 56 respectively) and excluded from the analyses. Demographic characteristics did not differ significantly between the experimental groups (before and after CP) in either study ($p > 0.10$).

2.3 Data Analysis

We used factor analysis-based estimation of latent *Composite Reliability* (RO, Raykov & Marcoulides, 2011, p. 161), which was calculated from the tenable configural models (confm, Table 3, see below in this section). RO is based on the so-called congeneric measurement model and does not assume equal factor loadings or uncorrelated error term variances. We used the estimation of RO for the general structure (Raykov, 2023) to obtain one score if the multifactorial structure was given. Correlated error terms are included in the error variance. The calculation of RO (ρ) is shown in Eq. 1.

$$\rho = \frac{\left(b_{11} + \ldots b_{1p}\right)^2 + \left(b_{21} + \ldots b_{2p}\right)^2 + 2cov\left(b_{11} + \ldots b_{1p}\right)\left(b_{21} + \ldots b_{2p}\right)}{\left(b_{11} + \ldots b_{1p}\right)^2 + \left(b_{21} + \ldots b_{2p}\right)^2 + 2cov\left(b_{11} + \ldots b_{1p}\right)\left(b_{21} + \ldots b_{2p}\right) + \theta_{11} + \ldots \theta_{2p} + 2psi}, \quad (1)$$

where $b_{11},\ldots, b_{1p}$ are the factor loadings of the factor 1, $b_{21},\ldots, b_{2p}$ are loadings of the factor 2, $\theta_{11},\ldots, \theta_{2p}$ are the error variances of the items of the two factors, *cov* is the covariance between the factors and *psi* is a correlated error term.

Measurement invariance was evaluated by so called exact Multi-Group Confirmatory Factor Analysis (MG-CFA). As shown in Eq. 2, in the MG-CFA, a CFA model is specified to predict the observed scores Y on an indicator of individual i within group j (Meredith, 1993; Millsap, 2011):

$$Y_{ij} = \tau_j + \Lambda_j \eta_{ij} + e_{ij} \quad (2)$$

where τ_j is the matrix of intercepts Λ_j is the matrix of factor loadings, and η_{ij} *and* e_{ij} are common factor scores and residuals for the individual i in group j, respectively.

Three kinds of measurement invariance were evaluated, starting with a more general and continuing to more specific level. Configural invariance – the first level—obtains, if the model specified in the Eq. 2 results in tenable model fit. If configural invariance is accepted, it does not mean that comparability bias is absent. Therefore, configural invariance serves as a first necessary step, but does not allow us to rule out comparability bias as alternative explanation for the results.

If $\Lambda_{j=} \Lambda$ holds, metric or weak invariance – the second level – obtains. To analyse metric invariance, equality constraints are placed on loadings. Metric invariance holds, if the restricted model does not significantly decrease the model fit statistics, as compared with the configural model. If metric invariance is obtained, bias can be ruled out as an explanation for the results for correlations (Meredith, 1993; Millsap, 2011).

Scalar or strong invariance as the third level obtains, if $\tau_{j=} \tau$ holds. To analyse scalar invariance, equality constraints are placed on intercepts. Scalar invariance holds, if the constraints on the intercepts to be equal among groups do not significantly decrease the fit of the metric model. Scalar invariance ensures that comparability bias does not confound the results when comparing latent or summarized group means or other population parameters (Meredith, 1993; Millsap, 2011).

In addition to exact MG-CFA, which is the most commonly used method to evaluate measurement invariance (e.g. Zercher et al., 2015), alternative methods are available, such as approximate Bayesian Multi-Group Factor Analysis (MGBSEM; B. Muthén & Asparouhov, 2012) and the Alignment Method (Asparouhov & Muthén, 2014).

The MGBSEM assumes that the differences in loadings and intercepts as parameters follow normal distributions with zero mean and pre-specified prior variance. It has been suggested that it should be used if MG-CFA rejects measurement invariance when there are no large, but many small (and potentially negligible) differences in loadings or intercept parameters between groups (Kim et al., 2017; Zercher et al., 2015). Such differences in parameters found in the data has been referred to as "approximate" invariance, in contrast to the "exact" invariance, which is given if MG-CFA provides tenable results (Kim et al., 2017; Pokropek et al., 2019). As it is not known beforehand, which differences in parameters (exact or approximate) will be present in the data, the author considered MGBSEM in addition to the exact MG-CFA. In the MGBSEM, the prior variance was set to 0.01, as the power is higher at this level of prior variance (e.g., Muthén & Asparouhov, 2013). The

author did not consider the Alignment method, because it was found to be less sensitive to misspecifications, which can be associated with the cross-cultural bias (Meitinger, 2017).

The analyses were conducted with Mplus 8.7. For MG-CFA, factor variance was fixed to 1 to allow all loadings to be compared. The author has implemented pairwise exclusion of missing values. Differences from the sample statistics reported in online source S9 and S10 are due to missing values. As more indicators were used in the revised versions of SF-12 and RHS-15, the author selected indicators from the alternatives that fitted the MG-CFA models best. Modification indexes (MODIND) were used to locate small (MODIND equal to or greater than 3.84 and less than 10; Kevin, 2015) and large (MODIND $\geq$ 10, e.g. Raykov & Marcoulides, 2011) non-invariant parameters.

The goodnes of fit (GOF) in MG-CFA was evaluated using a common approach by including the chi-square test (CMIN), the Root-Mean-Square Error of Approximation (RMSEA), and the Comparative Fit Index (CFI) (Beauducel & Wittmann, 2005). The CFI should be 0.95 or higher, while an RMSEA of 0.08 or less indicates an acceptable fit (Hu & Bentler, 1999). Hence, CFI of 0.90 have also been considered reasonable (Iacobucci, 2010). The Robust Maximum Likelihood estimator (MLR) was used due to the ordinal nature and non-normality of the data (L. Muthén & Muthén, 2024). For the MG-CFA, a significant change of chi-square (Meredith, 1993) and a change of ΔCFI $\geq$ - 0.010 or ΔRMSEA $\geq$ 0.015 indicate significant differences in model fit (Chen, 2007), and thus lack of exact measurement invariance. Accomponished with non-significant change of RMSEA and/or non-signficant change of chi-square, ΔCFI on the border of - 0.010 was accepted as a support for measurement invariance. The GOF of MGBSEM was evaluated by means of Bayesian posterior predictive p-value (ppp, e.g., B. O. Muthén & Asparouhov, 2013; Zercher et al., 2015) and 95% confidence limits (CI) for deviation between observed and predicted values, where a value of ppp close to zero and CI not including zero are interpreted as model misfit. RMSEA and CFI were used as additional GOF statistics for MGBSEM.

3 Results

3.1 Reliability

The results are presented in Table 2. A significant difference is obtained when the 95% confidence intervals (CI) of the RO do not overlap or overlap slightly. According to Kline (2016), reliability scores lower than 0.70 were considered inadequate, scores between 0.70 and 0.80 were considered adequate, scores between 0.80 and 0.90 were considered very good or high, and scores equal to or greater than 0.90 were considered excellent reliability. We first present the results for the first study, in which Arabic and Dari versions were compared and continue with the second study for the comparison between Arabic and German in a different sample.

1) Study 1

   The reliability of the SF-12 before CP was adequate in Arabic and high in Dari. It increased significantly after CP for Arabic and reached a high level. For Dari, there was an increase to an excellent level. For the RHS-15, reliability was excellent before CP and did not change thereafter. For LS, Arabic and Dari show opposite results. The high level of reliability for the Arabic version before CP decreased to adequate reliability

Table 2 Reliability RO (95% CI) before and after CP by instrument and sample

Instru-ment	Before CP				After CP			
	Study 1		Study 2		Study 1		Study 2	
	Arabic	Dari	Arabic	German	Arabic	Dari	Arabic	German
SF-12	.78 [.75;.81]	.86 [.83;.90]	.76 [.71;.84]	.77 [.70;.84]	**.86 [.83;.87]**	**.90 [.88;.93]**	**.85 [.80;.90]**	.80 [.76;.85]
RHS-15	.91 [.90;.92]	.92 [.90;.93]	.89 [.86;.92]	.85 [.81;.89]	.91 [.90;.93]	.92 [.91;.93]	**.93 [.91;.95]**	.86 [.80;.91]
LS	**.85 [.83;.87]**	.79 [.76;.82]	**.82 [.77;.88]**	.67 [.57;.78]	.77 [.74;.80]	**.82 [.80;.85]**	.74 [.67;.82]	**.76 [.67;.85]**

Bold: significantly higher than corresponding other condition

after CP, whereas it increased from adequate to high for Dari. In sum, we see a positive effect of CP on reliability for SF-12, no effect for RHS-15 and a mixed effect for LS.

2) Study 2

The results for SF-12 showing CP to have a positive effect on reliability found in the first study are replicated, althoutgh the increase found for German was not significant. A significant (but small) increase for RHS-15 is obtained for the Arabic version. For LS, mixed results are obtained again. For Arabic, high reliability before CP significantly decreased to an adequate score after CP. For German, reliability before CP is inadequate and significantly increased to the adequate score after CP. The results obtained in the first sample were mainly supported. For the LS, however, there were no inadequate scores after CP, which was given before CP for German.

Taking the results of both studies together, a positive effect of CP on reliability is given for both health related quality of life instruments. For the LS, the results are mixed. As more results are consistent with Hypothesis H1 that expected a positive effect of CP on reliability, it is largely supported.

3.2 Measurement Invariance

Table 3 summarizes the results of the MG-CFA and Table S4 of Online Supplement shows the results for MGBSEM. Detailed information about the number of model deviations is provided in the Online Source (Tables S5-S8).

3.2.1 Study 1

1) SF-12

The MG-CFA rejected configural invariance before CP (Table 3). The misfit was due to the correlated error terms (between SF-2 and SF-3 as well as SF-11 and SF-12, Table 1).[2] Allowing for these correlated error terms by holding them equal in both language groups significantly improved goodness of fit (Table 3, confm model), making the model acceptable and allowing us to proceed to evaluate metric invariance. Metric invariance was supported, but scalar invariance was slightly violated (due to the corresponding significant change of CMIN and CFI, but not RMSEA, Table 3).

[2] Correlated error terms were introduced between the indicators SF-2 (climbing stars) and SF-3 (other restrictions in physical activities) of the physical health as well as between the items SF-11 (less done mental) and SF-12 (less content mental) of mental health (Table 1). The presence of correlated error terms implies that responses to the items depend on factors other than latent variable, which would be among other things spurious. One example of spurious effect would be that the response to one item conditions the response to the other item. This can be the case for the correlated errors for "stairs" and "hinder" items, as climbing stairs is introduced as an activity in the first item (stairs) and the next item (hinder) provides the definition of this activity as "strenuous" (see Online Source S1 for item labels and wording). A potential non-substantial association between the "less content mental" and "less done mental" can be explained by the complex item wording and potential cognitive difficulty of differentiating between both items. Here and in the following, the author interprets a lower number of correlated error terms as a lower number of measurement errors and does not justify introduced error terms. The search for improvements to the configural model is relevant, because a move to the next step should be based on the configural model with at least just acceptable model fit (cf. e.g., Kline, 2016).

Table 3 MG-CFA results before and after CP

Model	*Before CP*						*After CP*						*Change*
	$\chi^2(df)$	$\Delta\chi^2(df)$	*RMSEA*	*ΔRMSEA*	*CFI*	*ΔCFI*	$\chi^2(df)$	$\Delta\chi^2(df)$	*RMSEA*	*ΔRMSEA*	*CFI*	*ΔCFI*	
SF-12 Study 1													
conf	493.71*** (106)	-	.082		.884		502.01*** (106)		.082		.913		o
confm	294.49*** (104)	-	**.058**	**-**	**.943**		411.7*** (105)	-	**.073**	**-**	**.933**		+
metric	326.40*** (116)	31.55**(12)	.058	**.000**	.937	**-.006**	469.62*** (117)	61.11**(12)	.074	**.001**	.923	**-.010**	o
scalar	394.19*** (128)	70.91***(12)	.062	**.004**	.920	-.017	699.86*** (129)	274.10***(12)	.089	**.015**	.875	-.048	-
SF 12 Study 2													
conf	320.77*** (106)	-	.123		.814		283.61*** (106)		.111		.895		o
confm	173.02*** (104)	-	**.070**	**-**	**.940**		210.79*** (104)	-	.087	**-**	**.937**	-	o
metric	187.77*** (116)	**14.32 (12)**	.068	**-.002**	.938	**-.002**	228.79*** (116)	**18.00 (12)**	.084	**-.003**	.933	**-.004**	o
scalar	288.46*** (128)	117.64*** (12)	.096	.028	.861	-.077	257.77*** (128)	29.53** (12)	.086	**.002**	.923	**.010**	+
RHS-15 Study 1													
conf	765.71*** (180)		.087		.897		500.986*** (180)		**.065**		**.925**		+
confm	497.41*** (177)		.065		.943		392.640*** (179)		**.053**		**.950**		+
metric	547.67*** (192)	53.91*** (15)	.066	**.001**	.937	**-.006**	413.941***(194)	**16.05 (15)**	.052	**-.001**	.949	**-.001**	+
scalar	684.69*** (207)	149.06*** (15)	.074	**.008**	.916	-.021	510.261***(209)	104.55***(15)	.058	**.006**	.930	-.019	o
RHS-15 Study 2													
conf	480.90*** (180)		.113		.790		287.19*** (180)		**.067**		**.903**		+
confm	314.18***(174)		**.079**		**.902**		254.90*** (179)		**.057**		**.931**		+
metric	371.23*** (189)	53.41*** (15)	.086	**.007**	.873	-.029	290.92***(194)	36.02* (15)	.062	**.005**	.912	-.019	+
scalar	555.94*** (204)	177.90*** (15)	.115	.029	.755	-.118	419.12***(209)	122.03*** (15)	.087	.025	.810	-.102	o
LS Study 1													
conf	**6.55 (4)**		**.033**		**.998**		23.84*** (4)		.086		**.978**		o
metric	26.13*** (8)	21.15*** (4)	.061	.028	.985	-.013	36.17*** (8)	10.68* (4)	.072	**-.014**	.968	**-.010**	+
scalar	38.59*** (12)	12.44* (4)	.061	**.000**	.978	**-.007**	53.63*** (12)	17.37** (4)	.072	**.000**	.953	-.015	o
LS Study 2													
conf	12.34* (4)		.122		**.956**		13.05** (4)		.126		**.946**		o

Table 3 (continued)

Model	*Before CP*						*After CP*						*Change*
	$\chi^2(df)$	$\Delta\chi^2(df)$	*RMSEA*	*ΔRMSEA*	*CFI*	*ΔCFI*	$\chi^2(df)$	$\Delta\chi^2(df)$	*RMSEA*	*ΔRMSEA*	*CFI*	*ΔCFI*	
metric	37.59*** (8)	24.19*** (4)	.163	.041	.845	-.111	17.91* (8)	**5.90 (4)**	.093	**-.033**	.940	**-.006**	+
scalar	89.69*** (12)	60.17*** (3)	.215	.052	.594	-.251	76.00*** (12)	65.32*** (4)	.193	.100	.615	-.325	o

***$p < .001$; **$p < .05$; *$p < .05$

conf: configural; confm: configural with correlated error terms

Bold: support for measurement invariance; o: no change; + positive change;—negative change

After CP, configural invariance was slightly improved, as only one correlated error term known from the version before CP (between SF-2 and SF-3, Table S1 of Online Source) needed to be introduced to obtain an acceptable model fit. Metric invariance was supported in this version as well, but scalar invariance was more strongly violated than before CP and therefore worsened. There were three non-invariant intercepts in the items before and six after CP (Online Source, Table S5 and S8).

The results of MGBSEM rejected measurement invariance before and after CPs, if the ppp value is considered (Online Source Table S4). The 95% CI for the ppp value is higher after CP than before CP, which confirms the decrease in scalar invariance for this version shown by the MG-CFA.

2) RHS-15

In the version before CP, configural invariance was strongly rejected by the MG-CFA (Table 3). Three correlated error terms needed to be introduced to obtain a just acceptable model (Online Source Table S6, S8). Metric invariance was slightly violated (as only the CMIN significantly worsened). Accordingly, two items had non-invariant loadings (Online Source Table S8). Scalar invariance was violated according to the significant change in CMIN and CFI and four non-invariant intercepts.

The version after CP shows improved configural invariance, because one correlated error term is introduced to reach acceptable model fit. Metric invariance is slightly improved (as all GOF support it), and scalar invariance is not changed. However, the number of non-invariant intercepts decreased from four to two (Online Source Table S6 and S8).

MGBSEM showed a strong improvement in invariance for the version after CP, as model fit improved and the number of non-invariant loadings and intercepts greatly decreased (Online Source Table S4).

3) LS

According to the MG-CFA, configural invariance held and metric invariance was rejected before CP (Table 3). The loadings of the health satisfaction item differed between language groups (Table 1, Online Source Table S7, S8). After CP, configural, metric and scalar invariance were supported. Therefore, the instrument after CP provided full measurement invariance compared to the previous version. The result is a slight improvement – change of a loading—and MGBSEM provided no difference between the versions (Online Supplement Table S4, S8).

Looking at the three instruments, configural invariance is improved if it was violated, that is in both health-related QoL instruments. Metric invariance was also improved if it was violated, hence the violations were rather minor. The effect on scalar invariance is mixed with negative effect for SF-12 and with positive effect according to MGBSEM for RHS-15.

3.2.2 Study 2

1) SF-12

The configural invariance was violated in the version before CP in the same way as in the first study, and the same correlated error terms had to be included to obtain a just acceptable model fit. Metric invariance was given, but scalar invariance was strongly violated (Table 3). This result was due to two large and numerous small intercept differences between the languages (Online Source Table S5, S8). The configural invariance

was not improved after CP and the same correlated error terms as in the version before CP had to be included. In contrast to the first study, both metric and scalar invariance held after CP, with the latter improved.

2) RHS-15

According to the MG-CFA, the configural invariance of RHS-15 was not given and the factorial structure was very poor (Table 3). Six correlated error terms had to be introduced to achieve a just acceptable fit (Online Source Table S6, S8). Metric invariance is slightly violated (one non-invariant loading), but scalar MI is strongly rejected with the provision of five non-invariant intercepts (Online Source Table S8). In the version after CP, configural invariance is strongly improved, as model fit is just acceptable without any modifications and can be further improved by introducing one correlated error term. Metric invariance is improved as well, but not the scalar invariance. Five non-invariant intercepts produced this result, comparable to the version before CP (Online Source Table S8).

The results of MGBSEM showed a strong improvement of model fit for the version after CP (Online Source Table S4). Both, metric and scalar invariance were improved, with fewer non-invariant loadings and intercepts (Online Source Table S8).

3) LS

Before and after CPs, configural invariance holds according to CFI of the MG-CFA (Table 3). Metric invariance was rejected before CP. In line with the results of the cognitive pretests, the item on living situation had different loadings between Arabic and German (Online Source S7). After CP, metric invariance held and was therefore improved, but scalar invariance was rejected (Table 3). MGBSEM rejected invariance in both versions (Online Source Table S4).

In sum, measurement invariance was positively affected in all cases. For SF-12, full invariance is obtained after CP according to both invariance analysis methods and for RHS-15 invariance is strongly improved according to MGBSEM. MG-CFA also showed an improvement in metric invariance for LS.

3.2.3 Summary for Both Studies

The Hypothesis H2 that expected a positive effect of CP on measurement invariance is partly supported. Thus, the results for cognitive pretests were positive for configural and metric invariance, but mixed for scalar invariance.

4 Discussion

The aim of the present research was to assess whether revisions of instruments based on cross-language (or cross-cultural) cognitive pretests can increase reliability and reduce comparability bias in the QoL measures. We surveyed refugees and host populations, where comparability biases due to language, cultural background and attributions are likely and should be controlled for.

The results regarding reliability were consistent with previous research (Menold et al., 2025; Maitland & Presser, 2018), which showed that it is worth using cognitive pretests when piloting questionnaires to increase reliability, although negative effects are also

possible (as the case of LS shows). In the last case cognitive pretests helped to overcome inadequate scores, however.

Cognitive pretests led to a strong improvement in configural invariance for the RHS-15 and some improvement for SF-12. Metric invariance was almost improved by the cognitive pretests. However, the violations of the metric invariance before CP were rather small. Improvement of scalar invariance was possible for the SF-12 for German and Arabic and for the RHS-15 according to MGBSEM (but not MG-CFA) in all cases. However, as is shown for SF-12 in Study 1, negative effects on measurement invariance analysis results are possible. As for SF-12 new items were developed to balance physical and mental health, new and therefore untested items could be the reason for limited comparability between Arabic and Dari.

Exact MG-CFA and approximate MGBSEM gave similar results for the SF-12, but not for other two instruments. MGBSEM showed a strong violation of metric and scalar invariance before CP and a strong improvement thereafter for the RHS-15, where MG-CFA did not show an improvement in scalar invariance. This result can be explained by the type of error invariance in the data, as MGBSEM provides adequate results in the case of many small and potentially negligible differences (approximate non-invariance), whereas MG-CFA provides valid results for many large and no negligible small differences, referred to as exact measurement invariance (Kim et al., 2017; Pokropek et al., 2019). Therefore, through cognitive pretests, exact invariance could be achieved for the SF-12 in the second sample and approximate measurement invariance could be achieved for the RHS-15 in both samples.

Overall, this study supports the expectation that cognitive pretests are sensible and can help not only to obtain information on comparability bias in the instruments evaluated (Benítez et al., 2022; Meitinger, 2017), but also to improve the quality of measurements. The resolution of conceptual overlap, the clarification of time references and avoidance of DBQs appear to have had a positive effect on reliability and measurement invariance. The improvement for the SF-12 can also be explained by the unification of the rating scales and the use of a higher number of categories for some of the items. In order to be more successful in resolving problems found by cognitive pretests, more research should be conducted on the effect of special features of questionnaire design on comparability bias. For example, little is known about how special types of response categories, time references, or reverse keying affect measurement invariance.

The mixed results obtained point to the limitations associated with conducting cognitive pretests. The first limitation is the selection of the parts of the questionnaire for the explicit pretest. To obtain more conclusive results, selection should either be based on the results of a measurement invariance and reliability analysis, or each item in an instrument should be explicitly tested. As the latter is more laborious and costly, a more purposive approach to item selection would be preferable. Alternatively, web probing can be used and items with probes could be distributed to random groups of respondents for more extensive probing. The next limitation is that cognitive pretests provide information about problems but do not automatically produce improved versions. Researchers, experts and translators are involved in the revisions of questionnaires following cognitive pretests and the quality may be affected by their personal influence and expertise. Provision and testing several alternatives for item formulation would be usable. Further research could also evaluate how to achieve more satisfactory results in terms of questionnaire improvement after cognitive pretests including an additional cognitive test of the revised version.

Overall, our analyses and data are limited to the pairwise comparisons of Arabic with Dari and German and obtaining results on the comparability of three languages is open

to the further research. With these versions, partial invariance approaches can be used, as suggested, for example by Pokropek et al. (2019), to find appropriate models for three-language comparison. Another limitation relates to the different sampling methods, which may decrease comparability between the two studies. In the Facebook sample, respondents may be less motivated than in the probabilistic sample, which may also explain the less satisfactory results for the SF-12 in this sample. Another explanation for the mixed results may be the use of mobile devices, where definitions to clarify terms such as those provided for living arrangements would have been less legible and might therefore be disregarded. As little is known about questionnaire design for mobile devices, more research is needed to evaluate comparability problems in mixed-device cross-cultural surveys.

The present research with the randomized experimental design allows for the conclusion that the differences in results between the questionnaire versions are due to the cognitive pretests used to adapt them. However, further research is needed to replicate the results and to investigate how scalar measurement invariance can be improved by a more purposive selection of indicators to investigate in the cognitive pretests and how rating scales, use of mobile devices and other questionnaire design decisions influence different levels of measurement invariance.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11205-025-03622-w.

Acknowledgements Thank to the Hagen von Hermanni and Jasmin Kadel for their valuable support in conduction of the studies and data collection.

Funding Open Access funding enabled and organized by Projekt DEAL. This study was funded by the German Science Foundation (DFG) in the scope of the ENSURE project as part of the PH-LENS Research Unit (FOR 2928/GZ: ME 3538/10-1). The funder had no influence on the design of the study, analysis or decision to publish.

References

AAPOR (2023). Standard Definitions Final Dispositions of Case Codes and Outcome Rates for Surveys. Retrieved on 5/17/2025 from https://aapor.org/wp-content/uploads/2023/05/Standards-Definitions-10th-edition.pdf

Andersen, H. H., Mühlbacher, A., Nübling, M., Schupp, J., & Wagner, G. G. (2007). Computation of Standard Values for Physical and Mental Health Scale Scores Using the SOEP Version of SF-12v2. *Journal of Contextual Economics – Schmollers Jahrbuch*, *127*(1), 171–182. https://doi.org/10.3790/schm.127.1.171

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

Beauducel, A., & Wittmann, W. W. (2005). Simulation Study on Fit Indexes in CFA Based on Data With Slightly Distorted Simple Structure. *Structural Equation Modeling: A Multidisciplinary Journal, 12*(1), 41–75. https://doi.org/10.1207/s15328007sem1201_3

Benítez, I., van de Vijver, F., & Padilla, J. L. (2022). A Mixed Methods Approach to the Analysis of Bias in Cross-cultural Studies. *Sociological Methods & Research, 51*(1), 237–270. https://doi.org/10.1177/0049124119852390

Borho, A., Morawa, E., & Erim, Y. (2022). Screening der psychischen Gesundheit von syrischen Geflüchteten in Deutschland: Der Refugee Health Screener [Mental Health Screening of Syrian Refugees in Germany: The Refugee Health Screener]. *Zeitschrift fur Psychosomatische Medizin und Psychotherapie*, *68*(3), 269–282. https://doi.org/10.13109/zptm.2022.68.oa1

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

Davidov, E., Schmidt, P., & Schwartz, S. H. (2008). Bringing Values Back In: The Adequacy of the European Social Survey to Measure Values in 20 Countries. *Public Opinion Quarterly*, *72*(3), 420–445. https://doi.org/10.1093/poq/nfn035

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. https://doi.org/10.1007/s11136-012-0232-5

Di Barbiano Belgiojoso, E., Cela, E., & Rimoldi, S. M. L. (2022). The Effect of Migration Experiences on Wellbeing Among Ageing Migrants in Italy. *Social Indicators Research, 161*(2–3), 553–579. https://doi.org/10.1007/s11205-020-02335-6

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method* (4th edition), Wiley. Retrieved from https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=1762797

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

Feliciano, C. (2020). Immigrant Selectivity Effects on Health, Labor Market, and Educational Outcomes. *Annual Review of Sociology, 46*(1), 315–334. https://doi.org/10.1146/annurev-soc-121919-054639

Fleishman, J. A., & Lawrence, W. F. (2003). Demographic variation in SF-12 scores: True differences or differential item functioning? *Medical Care*, *41*(7 Suppl), III75-III86. https://doi.org/10.1097/01.MLR.0000076052.42628.CF

Fong, D. Y. T., Lam, C. L. K., Mak, K. K., Lo, W. S., Lai, Y. K., Ho, S. Y., & Lam, T. H. (2010). The Short Form-12 Health Survey was a valid instrument in Chinese adolescents. *Journal of Clinical Epidemiology, 63*(9), 1020–1029. https://doi.org/10.1016/j.jclinepi.2009.11.011

Giovanis, E. (2022). The effects of international migration on well-being of natives and immigrants: evidence from Germany, Switzerland and the UK. *SN Business & Economics*, *2*(6). https://doi.org/10.1007/s43546-022-00230-5

Hadler, P., Nießen, D., Lenzner, T., Steins, P., Quint, F., & Neuert, C. [Cornelia.] (2021). *Translation of established public health measurement instruments into Arabic and Dari (ENSURE) (English Version). Cognitive pretest.* GESIS – Pretest Lab. https://doi.org/10.17173/pretest103

Harkness, J. A. (2003). Questionnaire Translation. In J. A. Harkness, F. J. R. de van Vijver, & P. P. Mohler (Eds.), *Wiley series in survey methodology. Cross-cultural survey methods* (pp. 35–56). Wiley-Interscience.

Hoffmeyer-Zlotnik, J. H., & Warner, U. (2014). Harmonising Demographic and Socio-Economic Variables for Cross-National Comparative Survey Research. *SpringerLink Bücher. Dordrecht: Springer.* https://doi.org/10.1007/978-94-007-7238-0

Höhne, J. K., Krebs, D., & Kühnel, S.-M. (2021). Measurement properties of completely and end labeled unipolar and bipolar scales in Likert-type questions on income (in)equality. *Social Science Research, 97*, Article 102544. https://doi.org/10.1016/j.ssresearch.2021.102544

Hollifield, M., Toolson, E. C., Verbillis-Kolp, S., Farmer, B., Yamazaki, J., Woldehaimanot, T., & Holland, A. (2016). Effective Screening for Emotional Distress in Refugees: The Refugee Health Screener. *The Journal of Nervous and Mental Disease, 204*(4), 247–253. https://doi.org/10.1097/NMD.0000000000000469

Hu, L., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Iacobucci, D. (2010). Structural equations modeling: Fit Indices, sample size, and advanced topics. *Journal of Consumer Psychology, 20*(1), 90–98. https://doi.org/10.1016/j.jcps.2009.09.003

Jacobsen, J., Klikar, J., & Schupp, J. (2017). Scales Manual IAB-BAMFSOEP Survey of Refugees in Germany – revised version.

Kevin, K. (2015). *Using* Mplus *for Structural Equation Modeling: A Researcher's Guide*. 2455 Teller Road, Thousand Oaks California 91320: Sage Publications, Inc. https://doi.org/10.4135/9781483381664

Kiesel, D. (Ed.) (1995). Arnoldshainer Texte: Vol. 88. Bittersüße Herkunft: Zur Bedeutung ethnisch-kultureller Aspekte bei Erkrankungen von Migrantinnen und Migranten. Frankfurt am Main: Haag und Herchen.
Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement Invariance Testing with Many Groups: A Comparison of Five Approaches. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(4), 524–544. https://doi.org/10.1080/10705511.2017.1304822
Kline, R. B. (2016). *Principles and practice of structural equation modeling* (Fourth edition). *Methodology in the social sciences*. New York, London: The Guilford Press.
Lenzner, T., & Neuert, C. [Cornelia] (2017). Pretesting Survey Questions Via Web Probing – Does it Produce Similar Results to Face-to-Face Cognitive Interviewing? *Survey Practice*, *10*(4), 1–11. https://doi.org/10.29115/SP-2017-0020
Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading: Addison-Wesley.
Maitland, A., & Presser, S. (2018). How Do Question Evaluation Methods Compare in Predicting Problems Observed in Typical Survey Conditions? *Journal of Survey Statistics and Methodology, 6*(4), 465–490. https://doi.org/10.1093/jssam/smx036
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. https://doi.org/10.1207/S15328007SEM1101_5
Meitinger, K. (2017). Necessary but Insufficient: Why Measurement Invariance Tests Need Online Probing as a Complementary Tool. *The Public Opinion Quarterly, 81*(2), 447–472. https://doi.org/10.1093/poq/nfx009
Meitinger, K., & Behr, D. (2016). Comparing Cognitive Interviewing and Online Probing. *Field Methods, 28*(4), 363–380. https://doi.org/10.1177/1525822X15625866
Menold, N., Hadler, P., & Neuert, C. (2025). Improving Cross-Cultural Comparability of Measures on Gender and Age Stereotypes by Means of Piloting Methods. *Sociological Methods & Research, 0*(0). https://doi.org/10.1177/00491241241307600
Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. https://doi.org/10.1007/BF02294825
Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.
Muthén, B. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*(1), 81–117. https://doi.org/10.2333/bhmk.29.81
Muthén, B. O., & Asparouhov, T. (2013). BSEM measurement invariance analysis. Retrieved from http://www.statmodel.com/examples/webnote.shtml
Muthén, L., & Muthén, B. (Eds.). (2024). *Mplus User's Guide*. Muthén & Muthén.
Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313–335. https://doi.org/10.1037/a0026802
Pokropek, A., Davidov, E., & Schmidt, P. (2019). A Monte Carlo Simulation Study to Assess The Appropriateness of Traditional and Newer Approaches to Test for Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(5), 724–744. https://doi.org/10.1080/10705511.2018.1561293
Rammstedt, B., Beierlein, C., Brähler, E., Eid, M., Hartig, J., Kersting, M., ... & Weichselgartner, E. (2015). *Quality Standards for the Development, Application, and Evaluation of Measurement Instruments in Social Science Survey Research* (RatSWD Working Papers No. 245). Retrieved on 5/17/2025 from RatSWD website: http://www.ratswd.de/dl/RatSWD_WP_245.pdf
Raykov, T. (2023). Scale Construction and Development Using Structural Equation Modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 472–492). The Guilford Press.
Raykov, T., & Marcoulides, G. A. (2011). *Introduction to Psychometric Theory*. Taylor & Francis.
Richter, D., Rohrer, J., Metzing, M., Nestler, Wiebke, Weinhardt, M., & Schupp, J. (2017). SOEP Scales Manual. *SOEP Survey Papers*. (423: Series C)
Rothgeb, J., Willis, G., & Forsyth, B. (2007). Questionnaire Pretesting Methods: Do Different Techniques and Different Organizations Produce Similar Results? *Bulletin of Sociological Methodology/bulletin De Méthodologie Sociologique, 96*(1), 5–31. https://doi.org/10.1177/075910630709600103
Scheuch, E. K. (1993). The cross-cultural use of sample surveys: Problems of comparability. Advance online publication. https://doi.org/10.12759/hsr.18.1993.2.104-138
Schimmack, U., Krause, P., Wagner, G., & Schupp, J. (2009). Stability and change of Well Being: An experimentally enhanced latent State-Trait-Error Analysis. *Social Indicators Research., 95*, 19–31.
Tibubos, A. N., & Kröger, H. (2020). A cross-cultural comparison of the ultrabrief mental health screeners PHQ-4 and SF-12 in Germany. *Psychological Assessment, 32*(7), 690–697. https://doi.org/10.1037/pas0000814
Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.
Van de Vijver, F. J. R. (2018). Capturing Bias in Structural Equation Modeling. In *Cross-Cultural Analysis* (pp. 3–43). Routledge. https://doi.org/10.4324/9781315537078-1

Ware, J., Kosinski, M., & Keller, S. D. (1996). A 12-Item Short-Form Health Survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care, 34*(3), 220–233. https://doi.org/10.1097/00005650-199603000-00003

Willis, G. B. (2015). The Practice of Cross-Cultural Cognitive Interviewing. *The Public Opinion Quarterly, 79*(S1), 359–395. https://doi.org/10.1093/poq/nfu092

Willis, G. B., & Miller, K. (2011). Cross-Cultural Cognitive Interviewing. *Field Methods, 23*(4), 331–341. https://doi.org/10.1177/1525822X11416092

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the Meaning of Factorial Invariance and Updating the Practice of Multi-group Confirmatory Factor Analysis: A Demonstration With TIMSS Data. *Practical Assessment, Research & Evaluation*, *12*(3). https://doi.org/10.7275/mhqa-cd89

Yan, T., Kreuter, F., & Tourangeau, R. (2012). Evaluating survey questions: A comparison of methods. *Journal of Official Statistics, 28*(4), 503–529.

Zercher, F., Schmidt, P., Cieciuch, J., & Davidov, E. (2015). The comparability of the universalism value over time and across countries in the European Social Survey: Exact vs. *Approximate Measurement Invariance. Frontiers in Psychology, 6*, 733. https://doi.org/10.3389/fpsyg.2015.00733