Cross-national validation of the WHO–5 well-being index within adolescent populations: Findings from 43 countries.

Philipp E. Sischka[1], Gina Martin[2], Caroline Residori[3], Nour Hammami[4], Nicholas Page[5], Christina Schnohr[6], Alina Cosma[7,8]

[1] Department of Behavioural and Cognitive Sciences, University of Luxembourg, Luxembourg

[2] Faculty of Health Disciplines, Athabasca University, Athabasca, Canada

[3] Department of Social Sciences, University of Luxembourg, Luxembourg

[4] Trent University Durham, Ontario, Canada

[5] DECIPHer, School of Social Sciences, Cardiff University, Wales, UK

[6] Department of Public Health, University of Copenhagen, Copenhagen, Denmark

[7] Trinity Centre for Global Health, School of Psychology, Trinity College Dublin, Dublin, Ireland

[8] Olomouc University Social Health Institute, Palacky University Olomouc, Olomouc, Czechia

Author Note

Philipp E. Sischka https://orcid.org/0000-0001-8414-0817

Gina Martin http://orcid.org/0000-0001-5295-316X

Caroline Residori https://orcid.org/0000-0002-6437-6002

Nour Hammami https://orcid.org/0000-0001-5816-5949

Nicholas Page https://orcid.org/0000-0002-4671-2797

Christina Schnohr https://orcid.org/0000-0002-3068-9879

Alina Cosma https://orcid.org/0000-0002-0603-5226

We have no conflicts of interest to disclose. Analyses were done with *R Statistics* and *Mplus*. This study was not preregistered. The R and Mplus scripts used in this article are stored on Open Science Framework (https://osf.io/pbexq).

Correspondence concerning this article should be addressed to Philipp Sischka, University of Luxembourg, Department of Behavioural and Cognitive Sciences, Institute for Health and Behavior, Maison des Sciences Humaines, 11, Porte des Sciences, L-4366 Esch-sur-Alzette, Luxembourg. E-mail: philipp.sischka@uni.lu

Author Contributions

Conceptualization: Philipp E. Sischka, Alina Cosma

Data curation: Philipp E. Sischka.

Formal analysis: Philipp E. Sischka, Review and proofreading of syntax: Caroline Residori

Methodology: Philipp E. Sischka.

Writing – original draft: Philipp E. Sischka, Alina Cosma, Nicholas Page, Christina Schnohr

Writing – review & editing: Philipp E. Sischka, Alina Cosma, Gina Martin, Caroline Residori, Nour Hammami, Nicholas Page, Christina Schnohr

Funding

Alina Cosma has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 101028678, Project GenerationZ. The study was supported by the project „Research of Excellence on Digital Technologies and Wellbeing CZ.02.01.01/00/22_008/0004583“ which is co-financed by the European Union. Nicholas Page is supported by the Centre for

Development, Evaluation, Complexity and Implementation in Public Health Improvement (DECIPHer), funded by Health Care Research Wales.

Acknowledgments

Health Behaviour in School-aged Children is an international study carried out in collaboration with WHO/EURO. Jo Inchley (University of Glasgow) was the International Coordinator for the 2021/22 survey. The Data Bank Manager was Professor Oddrun Samdal (University of Bergen). The 2021/22 survey included in this study was conducted by the following principal investigators in the 43 countries and regions: Albania (Gentiana Qirjako), Armenia (Sergey G. Sargsyan and Marina Melkumova), Austria (Rosemarie Felder-Puig), Flemish Belgium (Bart De Clercq), French Belgium (Katia Castetbon), Bulgaria (Lidiya Vasileva), Bulgaria (Anna Alexandrova-Karamanova, Elitsa Dimitrova), Canada (William Pickett and Wendy Craig), Croatia (Ivana Pavic Simetin), Czech Republic (Michal Kalman), Denmark (Katrine Rich Madsen), England (Sabina Hulbert, Sally Kendal), Estonia (Leila Oja), Finland (Jorma Tynjälä), France (Emmanuelle Godeau), Germany (Matthias Richter), Georgia (Lela Shengelia), Greece (Anna Kokkevi, Anastasios Fotiou), Greenland (Birgit Niclasen), Hungary (Ágnes Németh), Iceland (Arsaell M. Arnarsson), Ireland (Saoirse Nic Gabhainn), Italy (Lorena Charrier, Paola Nardone), Kazakhstan (Shynar Abdrakhmanova and Valikhan Akhmetov), Kyrgyzstan (Gulat Maemerova), Latvia (Iveta Pudule), Lithuania (Kastytis Šmigelskas), Luxembourg (Carolina Catunda, Maud Moinard), Malta (Charmaine Gauci), Moldova (Galina Lesco), the Netherlands (Gonneke Stevens and Saskia van Dorsselaer), North Macedonia (Lina Kostarova Unkovska), Norway (Oddrun Samdal), Poland (Anna Dzielka, Agnieszka Malkowska-Szkutnik), Portugal (Margarida Gaspar de Matos), Romania (Adriana Baban), Scotland (Jo Inchley), Slovakia (Andrea Madarasova Geckova), Slovenia (Helena Jericek), Spain (Carmen Moreno), Sweden (Petra Lofstedt),

Switzerland (Marina Delgrande-Jordan, Hervé Kuendig), Tajikistan (Sabir Kurbanov, Zohir Nabiev), Wales (Chris Roberts). For details, see http://www.hbsc.org

Cross-national validation of the WHO–5 well-being index within adolescent populations: Findings from 43 countries.

Abstract

The five-item World Health Organization Well-Being Index (WHO-5) is among the most frequently used brief standard measures to assess hedonic well-being. Numerous studies have investigated different facets of its psychometric properties in adult populations. However, whether these results apply to adolescents is uncertain and only few psychometric studies employed adolescent populations. Thus, the current study aimed to conduct an in-depth psychometric item response theory analysis of the WHO-5 among adolescents from 43 countries using the Health Behaviour in School-aged Children (HBSC) 2022 data set and investigated its (1) dimensionality and measurement structure, (2) test information values and marginal reliability, (3) cross-country measurement invariance and differential item/test functioning, and (4) convergent validity with other mental health and well-being related measures across countries. The WHO-5 showed a unidimensional measurement structure and overall high test information values and marginal reliability. Furthermore, although a large proportion of parameters were flagged as non-invariant, differential test functioning of the WHO-5 was only modest. Moreover, the WHO-5 mainly showed a concurring nomological network with the other mental health and well-being related measures across countries, although with some differences in effect sizes. The WHO–5 Well-Being Index is a psychometrically sound measure that has shown promise for cross-cultural research among adolescents in the included European, Central Asia and North American countries. The translated versions of the WHO-5 are available at https://osf.io/pbexq.

Keywords: WHO–5 well-being index, item response theory, measurement invariance, differential item functioning, cross-cultural research, adolescents, short scale, Well-being, Health Behaviour in School-aged Children,

Introduction

Subjective well-being refers to an individual's overall evaluation of their life. It encompasses both cognitive assessments of life satisfaction and affective experiences of positive and negative emotions (Diener et al., 2000). In the context of adolescents, studying subjective well-being becomes paramount, as it provides a nuanced understanding of their mental health and overall life quality during a crucial developmental stage. The study of subjective well-being among diverse adolescent samples sheds light not only on immediate well-being but also on the foundations for long-term mental health outcomes. Research has consistently shown that positive subjective well-being in adolescence is associated with numerous benefits, including better academic performance, enhanced social relationships, and a lower risk of mental health issues later in life (Rees et al., 2016; Shaffer-Hudkins et al., 2010; Tomas et al., 2020). Given that this decade has seen the largest population of adolescents in human history with most living in low- and middle-income countries (Patton et al., 2016), it becomes imperative that researchers have access to well-validated measures that capture subjective well-being of this population and are suitable for cross-national comparisons (Boer et al., 2018). Such measurement instruments could then support the development of targeted interventions and support systems that positively impact the mental health trajectories of adolescents, promoting resilience and well-being into adulthood (Tejada-Gallardo et al., 2020). They could also inform national, regional, and cross-national health and social policy development. Ideally, these measures are not overly long as many research contexts demand short scales, such as multi-topic (cross-cultural) large scale surveys, longitudinal studies, and prescreening surveys.

The WHO–5 Well-being Index (WHO–5; World Health Organization, 1998) represents a promising measure that potentially could meet these requirements (for a discussion of other self-reported subjective well-being measures see McDowell, 2010; Jacobs

et al., 2023). However, most previous psychometric research on the WHO-5 has been conducted among adults. Thus, it remains unclear whether the WHO-5 is also a sound psychometric measure for adolescents across different countries. Therefore, the aim of this study was to investigate the psychometric properties of the WHO–5 in representative samples of adolescents across 43 European, Central Asia and North American countries.

The WHO–5 Well-Being Index

The WHO–5 is a five-item composite scale used to assess subjective hedonic well-being (Bech, 2012; Kusier & Folker, 2020). The items represent non-invasive questions that are positively worded and capture both feelings (e.g. 'I have felt calm and relaxed') and functioning (e.g. 'My daily life has been filled with things that interest me') in the past two weeks, with response options ranging from 0 (= *at no time*) to 5 (= *all of the time*). The WHO–5 is not specific to any disease or condition making it a useful tool for assessing well-being in both clinical and non-clinical samples. Nonetheless, it displays strong diagnostic accuracy when screening for clinical depression in adults (Topp et al., 2015; Low et al., 2023; Krieger et al., 2014; McDowell, 2010). Furthermore, the WHO–5 has also been found to be a sound measure for cross-cultural comparisons in adults (e.g., Jami & Kemmelmeier, 2020; Sischka et al., 2020).

It is important to note that most studies investigating the psychometric properties of the WHO–5 have been conducted among adults (e.g., Jami & Kemmelmeier, 2020; Lara-Cabrera et al., 2022; Sischka et al., 2020; Topp et al., 2015), and thus, cannot be unambiguously generalized to adolescent populations (e.g., Borgers et al., 2000; Conijn et al., 2020; Taber, 2010). Research needs to establish that survey measures are developmentally appropriate (Krause et al., 2022; Rose et al., 2017), and that measures that were originally developed for application in adult populations can be used within adolescent populations.

Given the WHO-5 has a Flesch Reading Ease Score of 90.0 (de Wit et al., 2007)[1], which is equivalent to a U.S. 5th grade reading level (Peter et al., 2018) and comparable to other quality of life measures typically employed within adolescence populations (Krause et al., 2022), it is possible that it could be applied well in this context.

Psychometric properties of the WHO–5 in adolescent populations

The few psychometric studies in adolescent populations have investigated the dimensionality, the measurement structure, reliability (usually investigated by means of Cronbach's α), measurement invariance or differential item/test functioning across various groups, and convergent validity of the WHO–5. These results are summarized below.

Dimensionality and measurement structure

Previous studies using parallel analysis (Quansah et al., 2022), principal component analysis with eigenvalue-greater-than-one criterion (deWit et al., 2007), or confirmatory factor analysis (Cosma et al., 2022; deWit et al., 2007)[2] mainly indicated that the WHO–5 exhibits an essentially unidimensional measurement structure within adolescent samples. In samples of school-aged children across fifteen countries, Cosma et al. (2022) reported factor loadings across all items and all countries between .50 and .89. One study (Quansah et al., 2022) also employed an IRT framework and specified a graded response model to the WHO-5 items within a Ghanaian sample. The discrimination parameter (a) ranged between 1.22 (item 1) and 2.80 (item 3). Moreover, they found only small threshold steps between the response

[1] This score focuses on sentence length and syllables per word as an indicator of overall readability (Peter et al., 2018).

[2] Cosma et al. (2022) interpreted some of their results critically and judged the model fit as "poor" in 12 out of 15 countries based on different fit indices and the WLSMV estimator. However, all countries showed *CFI* values above .95, *TLI* values above .90 and *SRMR* values below .03. Only the *RMSEA* was above .10 in five countries. Notably, Shi and Maydeu-Olivares (2020) showed in a simulation study that the *RMSEA* and *CFI* are substantially affected by the estimation method (i.e., comparing maximum likelihood, unweighted least squares and diagonally weighted least squares). They recommend using the *SRMR* that was robust to the estimation method. Thus, while Cosma et al. (2022) used the WLSMV estimator, the *SRMR* seems to be a more reliable fit index that showed no substantial model misfit.

categories ‘Less than half of the time’ and ‘More than half of the time’, between ‘More than half of the time’ and ‘Most of the time’, and between ‘Most of the time’ and ‘All of the time’.

Reliability

Previous research indicated that the WHO–5 showed sufficient marginal reliability (i.e., internal consistency) within adolescent samples but reliability estimates varied considerably. Adjorlolo and Anum (2021) reported a Cronbach’s α value of .70 within a sample of senior high school students in Ghana. Cosma et al. (2022) found moderate to high internal consistency in terms of Cronbach’s α (ranged between .75 and .92) in samples of school-aged children across fifteen countries. DeWit et al. (2007) reported a Cronbach’s α value of .82 in a sample of adolescents with type 1 diabetes in the Netherlands. Finally, employing a graded response model, Quansah et al. (2022) determined a marginal reliability estimate of .86 and test information values of 8 and above over a wide range of theta within a sample of senior high school students in Ghana.

Measurement invariance

Measurement invariance (MI) testing is a statistical approach used to assess the extent that scale items operate equivalently across contexts (e.g., gender, age group, country). This is critical to understanding the wider transferability of a scale and the prospective accuracy of cross-group comparisons of latent mean scores. Thus, demonstrating measurement invariance is paramount for survey measures that are intended to be used for comparative (and often cross-cultural) research (Millsap, 2011). To the best of our knowledge, only Cosma et al. (2022) investigated cross-country MI of the WHO–5 among adolescent populations. They included representative samples of adolescents across 15 European and Central Asian countries and conducted a multigroup confirmatory factor analysis (with WLSMV estimator). MI was assessed according to the change in global model fit statistics (i.e., *RMSEA, CFI, TLI*). Cosma et al. (2022) found that a one-factor model of the WHO–5 did not exhibit

configural invariance. In a next step, based on modification indices they excluded the first item ('I have felt cheerful and in good spirits') as its error term correlated with the error term of the second item. They deemed this four-item version to have acceptable model fit and subsequently tested it for more stringent MI models. Cosma et al. (2022) found that this four-item version exhibited configural and metric invariance, but not scalar invariance across countries. By freeing the intercepts (or thresholds) of items 2 and 3 they established a partial invariance model.

Convergent validity and nomological network

Strong convergent validity (i.e., the extent to which the WHO–5 corresponds to measures of related constructs) has also been reported in several studies using adolescent samples. For example, the WHO–5 has been found to be negatively correlated with well-validated measures of depression and anxiety such as the Patient Health Questionnaire-9 (r = – .36), the Center for Epidemiologic Studies Depression Scale (r = – .67), the Reynolds Adolescent Depression Scale-Short Form (r = –.59; Lambert et al 2014) and the General Anxiety Disorder-7 (r = – .35; Adjorlolo & Anum, 2021; de Wit et al., 2007). Positive correlation with the Warwick-Edinburgh Mental Well-being Scale (r = .57) has been shown among a UK-based sample of 13–16-year-olds (Clarke et al., 2011). Negative correlations with psychosomatic complaints (r = – .45; Cosma et al., 2022), health rated as poor (r = -.34; Cosma et al., 2022)[3] and negative affect (r = – .75; Quansah et al., 2022), and positive correlations with life satisfaction (r = .43; Cosma et al., 2022), general mood (r = .54; Lambert et al., 2014), positive affect (r = .58; Quansah et al., 2022), have also been observed. In samples of adolescents with type 1 diabetes, the WHO-5 correlated positively with the total generic and subscale-specific scores on the Paediatric Quality of Life Inventory (de Wit et al.,

[3] The coding of self-rated health was incorrectly described in Cosma et al. (2022). Higher values actually represented worse self-rated health.

2012), and negatively with depressive symptoms and diabetes burden (Steinberg et al., 2017). The WHO–5 also displayed good criterion validity in screening for depression in pediatric and adolescent psychiatric settings (Allgaier et al., 2012; Blom et al., 2012; Tittel et al., 2023).

Item response theory analysis

As this overview of the literature reveals, previous studies mainly applied psychometric methods within a "classical test theory" (CTT) framework (De Champlain, 2010) and treated the WHO-5 items as continuous indicators. Compared to CCT approaches, item response theory (IRT) analysis offers some advantages. First, IRT models account for the ordinal nature of items rather than assuming that they are continuous.[4] Second, IRT analysis focuses more on understanding the performance of each individual item (Houts et al., 2022), incorporating both discrimination parameters (similar to factor loadings in CFA) and item difficulty (threshold parameters). Therefore, IRT is often called an item-level theory, whereas CTT is referred to as test-level theory (DeMars, 2018). Third, unlike Cronbach's alpha, which assumes uniform measurement error across the latent variable continuum, IRT allows for varying reliability across the continuum depending on item characteristics (Houts et al., 2022). Fourth, IRT models typically utilize all available data, providing full-information estimates, whereas CFA with WLSMV estimation uses summary statistics such as polychoric correlations (Wirth & Edwards, 2007), which is why IRT is often referred to as a full-information item factor analysis (e.g., Cai et al., 2011).

Lack of research

[4] Notably, with the introduction of specific estimation methods (e.g., weighted least square mean and variance adjusted [WLSMV]), CFA can now accommodate the ordinal nature of items, similar to how IRT models handle such data. In fact, CFA using the WLSMV estimator provides estimates comparable to those obtained with the (normal ogive) graded response model. The main distinction is that WLSMV uses limited information (i.e., only the first and second moments of the data), whereas IRT models typically utilize full information from all available data. Nevertheless, maximum likelihood estimation, which assumes continuous indicators, still appears to be the most commonly used estimator in the CFA context (Sellbom & Tellegen, 2019).

The discussion of the psychometric properties showed that the WHO–5 seems to be a promising candidate to assess general hedonic well-being among adolescents. However, a few things should be noted. First, as discussed, previous studies mainly employed a "classical test theory" (CTT) framework and did not exploit the potential of an in-depth IRT analysis. Second, previous studies have only included a limited set of countries often within convenience samples. Thus, we have only limited information regarding the psychometric properties of the WHO-5 in adolescent populations across a wider range of countries. Third, as noted earlier, in order to be able to make valid cross-country comparisons it is essential to test whether the measurement structure across countries is the same, i.e., that MI holds (e.g., Maassen et al., 2023). With the exception of Cosma et al. (2022) who included 15 countries, previous studies have only investigated single-country data. However, as Cosma et al. (2022) investigated MI only based on the change of global model fit, no information about which countries deviated from each other is available. Moreover, they only tested a configural model on the full five-item set and omitted more stringent MI models because of global model fit indices that might have also been judged as "acceptable" (i.e., *RMSEA* = .073, *CFI* = .982; *TLI* = .964; see e.g., Little, 2013). Thus, we have only limited evidence whether the WHO–5 shows the same measurement structure across a wide range of countries. Fourth, no study so far has investigated whether the nomological net of the WHO–5 is similar across countries. Fifth, we have only limited knowledge about normative data of the WHO-5 within the adolescent population across countries.

Aim of the present study

Given its growing application in adolescent cross-national research (e.g., Adjorlolo & Anum, 2021; Cosma et al., 2023; Cosma et al., 2022; Winzer et al., 2021), the present study provides an in-depth psychometric analysis of the WHO–5 in large representative samples of adolescents across 43 countries in Europe, Central Asia and Canada. First, the dimensionality

of the scale was investigated. Second, an IRT analysis was conducted. Since the WHO-5 items have ordered-response categories and previous studies have indicated that the items differ in terms of discrimination (e.g., Cosma et al., 2022), the graded response model (GRM) might be an adequate IRT model. The GRM is also among the most frequently applied models (Depaoli et al., 2018; Houts et al., 2022; Toland, 2014). Due to its relaxed assumptions, the GRM is an IRT model that is particularly relevant to evaluate survey items with ordinal response categories (Depaoli et al., 2018; Houts et al., 2022; Toland, 2014). However, a reduced version of this model (R-GRM) is also investigated, that estimates one common slope parameter across the ordinal response categories for all items (Toland, 2014) and is, thus, more parsimonious. Third, cross-national measurement invariance and differential item functioning were investigated to test whether the measurement structure of the WHO–5 is comparable across countries. Fourth, item properties, test information values, and marginal reliability were inspected. Fifth, the associations between the WHO–5 and other well-established measures of subjective well-being were investigated. Finally, norm values (mean, standard deviation, percentile norms) were provided.

Method

This study's design and its analysis were not preregistered. The data and materials used in this study will become publicly available in October 2027. The code used for this analysis and the Electronical supplement [ESM] is available at https://osf.io/pbexq. All analyses were conducted in R (version 4.3.1; R Core Team, 2023) and Mplus Version 8.8 (Muthén & Muthén, 1998-2017). See section *Software information* for the used packages.

Data collection and survey design

The Health Behaviour in School-aged Children (HBSC) study is a large cross-national research study undertaken in collaboration with the WHO Regional Office for Europe. Since 1983, an increasing number of countries from across Europe, Central Asia and North America

have joined the HBSC study, which conducts a school-based survey every 4 years; most recently in 2022 with 44 participating countries. Representative samples of adolescents aged 11, 13, and 15 years are invited to complete the survey and asked about their health behaviors, well-being, and social context. All participating countries follow a standard protocol, with stratified sampling used in each country to represent the regional, economic, and public–private distribution of schools (Inchley et al., 2023). Ethical approval to conduct the survey is sought by the principal investigator of each participating country from their respective ethics review board (or equivalent regulatory institution). Participation in the survey was voluntary with pupils informed that responses were anonymous. Informed consent was obtained from schools, parents, and children, prior to completion. The survey was administered in a classroom setting, with countries able to utilize either paper-based or electronic questionnaires. Detailed information on study methods is available elsewhere (Inchley et al., 2020).

Participants

The initial sample consisted of N = 279,117 respondents from 44 countries in the 2021/2022 survey cycle. Serbia accidentally omitted one response category of the WHO–5, thus, was excluded (n = 3,713). Due to incomplete data (i.e., one or more missing values on the WHO–5 items), 5.8% (n = 16,015) of respondents were also excluded from the analyses (see Figure A1 in the ESM [https://osf.io/pbexq] for the missing data pattern).[5] The analytical sample therefore consisted of N = 259,389 respondents from 43 countries (50.9% girls; M_{age} = 13.6 years SD_{age} = 1.64 years). The number of respondents per country ranged between 1,229 (Greenland) and 34,427 (Wales). See Table A1 in the ESM for further sample details.

Measures

[5] The higher missing rates in Denmark are due to the fact that the WHO–5 items were not presented to participants aged 11.

WHO–5 Well-Being Index

The WHO–5 Well-Being Index starts with the instructions “Please indicate for each of the five statements which is the closest to how you have been feeling over the last two weeks.”. The five items are as follows: “Over the last two weeks…” (1) “... I have felt cheerful and in good spirits”, (2) “... I have felt calm and relaxed”, (3) “... I have felt active and vigorous”, (4) “... I woke up feeling fresh and rested”, (5) “... my daily life has been filled with things that interest me”. The items have the following response options: *At no time* (0), *Some of the time* (1), *Less than half of the time* (2), *More than half of the time* (3), *Most of the time* (4), and *All of the time* (5). All language versions are available at https://osf.io/pbexq.

Other well-being variables

Life satisfaction was measured with Cantril Ladder, a single item, it reads: “Here is a picture of a ladder. The top of the ladder “10” is the best possible life for you and the bottom “0” is the worst possible life for you. In general, where on the ladder do you feel you stand at the moment? Tick the box next to the number that best describes where you stand.”. The response options present a ladder from 0 (= *worst possible life*) to 10 (= *best possible life*)

Self-rated health (SRH) was also measured with a single item, it reads “Would you say your health is…?” with the response options ranging from 1 (= *excellent*) to 4 (= *poor*). For our analysis, SRH was recoded into two response options, 0 (= *poor*) and 1 (= *fair/good/excellent*), according to recommendations of Schnohr et al. (2016).

Psychosomatic complaints were assessed with the HBSC Symptom Checklist (Heinz et al., 2022). This item included the question “In the last 6 months: how often have you had the following….?”, and included eight symptoms, (headache, stomachache, backache, feeling low, irritability/bad temper, feeling nervous, difficulties in getting to sleep, feeling dizzy). Adolescents had five response options to choose from: 1 (= *about every day*), 2 (= *more than*

once a week), 3 (= *about every week*), 4 (= *about every month*), and 5 (= *rarely or never*). For this analysis, these responses were recoded into 0 (= *rarely or never*) to 4 (= *about every day*).

Loneliness was measured with a single item "During the past 12 months, how often have you felt lonely?", with response options 1 (= *never*), 2(= *rarely*), 3(= *sometimes*), 4 (= *often*), and 5 (= *always*).

Demographic variables

To identify their *gender*, respondents were asked to indicate whether they are a boy or a girl. Additionally, respondents were asked to indicate the month and year of their birth to determine their *age* at the time of data collection.

The socioeconomic status of the families of the respondents was measured using the Family Affluence Scale (FAS-III) scale (Torsheim et al., 2016; Boer et al., 2023). FAS III is a self-report measure of family wealth that includes the following six items: (1) "Does your family own a car, van or truck?"; (2) "Do you have your own bedroom for yourself?"; (3) "How many computers do your family own (including laptops and tablets, not including game consoles and smartphones)?"; (4) "How many bathrooms (room with a bath/shower or both) are in your home?"; (5) "Does your family have a dishwasher at home?" and (6) "How many times did you and your family travel out of [country] for a holiday/vacation last year?" Answer categories are item specific with 0 (= *no*), 1 (= *yes, one*), and 2 (= *yes, two or more*) for item (1), 0 (= *no*) and 1 (= *yes*) for items (2) and (5), 0 (= *none*), 1 (= *one*), 2 (= *two*), 3 (= *more than two*) for items (3) and (4), and 0 (= *not at all*), 1 (= *once*), 2 (= *twice*), and 3 (= *more than twice*) for item (6). The FAS score is based on a ridit-scaled variable derived from the sum of the responses, which classifies adolescents into three affluence categories in each country: lowest 20%, medium 60%, and highest 20% (Elgar et al., 2017).

Statistical analysis

Preliminary data analysis

In a first step, the distributions of the items were evaluated as an initial check of item properties, and to determine whether each response category within each item was sufficiently used to obtain stable parameter estimates (Toland, 2014). In a next step, polychoric correlations between the items were calculated to get a first impression of the dimensionality of the WHO–5 items and to identify potential problematic items within each country (Watkins, 2018)

Dimensionality assessment

The *dimensionality* was tested by submitting a polychoric correlation matrix to exploratory graph analysis (EGA) with the *glasso* algorithm, a recently proposed network psychometric method for dimensionality assessment (Golino et al., 2017; Golino et al., 2020). This procedure has been shown to outperform many other dimensionality assessment approaches and assesses the number of dimensions and the relation between the indicators and the dimension in a single step (Golino et al., 2020).

IRT model comparison and model-data fit

The appropriateness of the GRM and the R-GRM was evaluated with goodness of fit statistics (*RMSEA*, *SRMSR*, *CFI*, *TLI*) relying on the limited-information test statistic C_2 (Cai & Monroe, 2014; Monroe & Cai, 2015). It should however be noted that, just like in a CFA context (e.g., McNeish & Wolf, 2023), judging model fit against fixed cutoffs might be problematic as many factors other than model (mis-)fit have an impact on them.[6] The two IRT models were thus further compared with the likelihood ratio test and three information criteria (i.e., AIC, BIC, and SABIC). Moreover, to assess the relative improvement in the proportion of variability accounted for by one model over the other, we calculated the change in R^2 based

[6] In addition, it should be noted that goodness of fit statistics based on the C_2 test statistic are relatively new, and therefore caution should be exercised in interpreting these statistics based on simple rules of thumb. Moreover, it has been shown that the *RMSEA* based on C_2 test statistic is influenced by the number of answer categories of the items with more answer categories tend to increase *RMSEA* values (Monroe & Cai, 2015).

on the likelihood ratio G^2 test statistic (De Ayala, 2009). As these fit statistics represent global model fit statistics that do not tell us where the model misfit stems from, they always should be interpreted in conjunction with local model fit statistics. The *local* (or conditional) *independence* assumption was evaluated with the Jackknife Slope Index (JSI) and its respective cutoff value (mean of the JSI values plus twice the standard deviation) as proposed by Edwards et al. (2018). Positive JSI values indicate that the removal of a specific item causes the slope of another item to decrease, while negative JSI values indicate that the removal of a specific item leads to an increased slope of another item. *Item fit* was investigated with the generalized $S\text{-}X^2$ item fit index (Kang & Chen, 2011) and corresponding item-level *RMSEA* values as measure of effect size. Raw residual plots were created to assess the *functional form* (or monotonicity) assumption (Wells & Hambleton, 2016).

Measurement invariance and differential item/test functioning analysis

Measurement invariance of the WHO–5 across countries was first evaluated by means of a multigroup IRT analysis with increasingly restrictive nested models, starting with the *configural invariance* model (i.e., factor variances fixed to one, factor means fixed to zero, discrimination and threshold parameters freely estimated in all countries), then the *metric invariance* model (i.e., factor variance fixed to one only in the first country, factor means fixed to zero in all countries, discrimination parameters constrained to be equal across all countries, threshold parameters freely estimated in all countries), and finally the *scalar invariance* model (i.e., factor variance fixed to one only in the first country, factor means fixed to zero only in the first country, discrimination and threshold parameters constrained to be equal across all countries).

Potential *differential item/test functioning* across countries was investigated with the alignment optimization method (Asparouhov & Muthén, 2014; DeMars, 2020; Marsh et al., 2018; Muthén & Asparouhov, 2014) with robust maximum likelihood estimator (MLR).

Applying a simplicity function, this procedure starts with the configural invariance model (assuming that a reasonable configural invariance model exists) and searches for a set of (discrimination and threshold) parameters that can be constrained across countries without loss in model fit. The alignment optimization is a linking approach under a partial invariance scenario and the aligned model has the same model fit as the configural invariance model. The aligned model is determined in two steps: First, the configural invariance model is estimated. Second, the factor means and variances are freely estimated, and their values are chosen based on the simplicity function to minimize the total size of non-invariance for every pair of groups and every discrimination and threshold parameter (e.g., Kim et al., 2017). The alignment procedure also provides an R^2-like measure (i.e., ranging from 0 to 1) for every parameter that represents variation in this parameter across groups in the configural model that can be explained by variation in factor mean and factor variance across groups and not non-invariance. Thus, higher values indicate higher levels of invariance. Asparouhov and Muthén (2014) recommend to start with the FREE approach (factor variance of the reference group is set to 1 and freely estimated in all other groups, factor mean is freely estimated in all groups) and to switch to the FIXED approach (factor mean and factor variance of the reference group is set to 0 and 1 respectively, and freely estimated in all other groups) if the FREE approach is poorly identified. For methodological and technical details of the alignment procedure (such as the computation of the simplicity function), see Asparouhov and Muthén (2014), Kim et al. (2017) and Marsh et al. (2018). For examples of applied studies see Sischka et al. (2024) and Heinz et al. (2022).

The aligned IRT parameters were used to further analyze country pairwise differential test functioning with the compensatory (*sDRF*) and non-compensatory differential response functioning (*uDRF*) statistics (Chalmers, 2018) and respective 99% bootstrapped confidence intervals (with n = 10.000 bootstrap samples).

Item properties, information functions and marginal reliability

The psychometric properties of the WHO–5 were investigated using the aligned multigroup IRT model, as this approach places the group-specific item parameters onto a common metric (e.g., Muthén & Asparouhov, 2014). Thus, the item parameters across countries can be directly compared. Based on this model, item and test characteristic curves (ICC, TCC) as well as item and test information functions (IIF, TIF) were derived, with empirical marginal reliability (ρ) as summary measure of score precision (Brown, 2018).

Relationship with external criteria

The association of the WHO–5 with other variables was investigated by means of correlational analysis. Confidence intervals were derived with bootstrapping (n = 1.000 bootstrap samples).

Results

Preliminary data analysis

The items of the WHO–5 showed some amount of skewness and kurtosis ($M_{skewness}$ = -0.48, $SD_{skewness}$ = 0.36, $Min_{skewness}$ = -1.60, $Max_{skewness}$ = 0.39, $M_{kurtosis}$ = -0.61, $SD_{kurtosis}$ = 0.57, $Min_{kurtosis}$ = -1.40, $Max_{kurtosis}$ = 2.20; see Table A2 and Figure A2 in the ESM [https://osf.io/pbexq]) but these values were not pronounced enough to indicate any issues with item performance. Almost all response categories within each item across countries were sufficiently used (i.e., $n \geq 30$).[7] The items' polychoric correlations over all countries and item pairs ranged between .42 and .80 ($M_{polycor}$ = .59, $SD_{polycor}$ = .07). Kyrgyzstan showed the lowest average item intercorrelations ($M_{polycor}$ = .48, $SD_{polycor}$ = .03, $Min_{polycor}$ = .42, $Max_{polycor}$ = .53) whereas Bulgaria showed the highest average item intercorrelations ($M_{polycor}$ = 0.72,

[7] The only exception was the first category of the first item in Denmark, with only 19 respondents selecting this category. Nevertheless, for the sake of consistency, we have refrained from collapsing response categories.

$SD_{polycor}$ = 0.04, $Min_{polycor}$ = .68, $Max_{polycor}$ = .81, see Figure A3 in the ESM). The polychoric correlation matrices showed no abnormalities.

Dimensionality assessment

The EGA indicated a one-factor solution for all countries (see Figure A4 ESM). Moreover, the EGA showed that items 1 and 2 were more strongly connected than any other item pair for most countries.

IRT model comparison and model-data fit

According to commonly used thresholds (e.g., Kline, 2016; Little, 2013), the goodness of fit statistics based on the test statistic C_2 for the GRM, the *SRMSR*, the *CFI*, and the *TLI* indicated a good/very good model fit, whereas the *RMSEA* indicated a poor fit (i.e., values above .1) for some countries (i.e., Bulgaria, Switzerland, Denmark, Malta; see Table 1). For the R-GRM most fit indices got slightly worse but were still in a good range (see Table A4 in the ESM). For the R-GRM, three countries showed *RMSEA* values above .1 (i.e., Bulgaria, Denmark, Slovenia). The likelihood ratio test (Table A5) and the information criteria (Figure A5 in the ESM) favored the GRM over the R-GRM in almost all countries (except for Greenland). However, the ΔR^2 between the R-GRM and the GRM for every country ranged between .000 (Greenland; Kyrgyzstan) and .006 (Slovenia), indicating overall only small model improvements. Nevertheless, as the main aim of the present study was to find a configural model that fits in every country and to accommodate the heterogeneity across countries, the items were further analyzed based on the GRM (see Table A6-10 in the ESM for item parameters).

(Table 1 about here)

The JSI flagged local dependence between item 1 and 2 for Austria, Estonia, Scotland, Wales, Croatia, Hungary, Moldova, North Macedonia, Norway, Slovakia, and Sweden (see Figure A6 in the ESM). However, the values were only slightly above the threshold; thus,

local dependency might be considered still in an acceptable range. The generalized S-X^2 item fit index flagged most of the items as deviating from the GRM curves. However, the item-level *RMSEA* ranged between .009 and .039 for item 1, between .000 and .049 for item 2, between .008 and .032 for item 3, between .012 and .032 for item 4, and between .006 and .032 for item 5, indicating low to medium deviation of the items from the GRM (see Figure A7 in the ESM). Finally, the raw residual plots indicated no strong deviation from monotonicity (see Figure A8-A50 in the ESM). See Figure A51-A55 for the item parameter, ICC, IIF, TCC, and TIF for the different countries. These analyses indicated that the GRM can be used as a starting point for the configural invariance model.

Measurement invariance and differential item/test functioning analysis

The multigroup analysis revealed a very good fit for the configural invariance model (C_2 = 8,246.855, *df* = 215, $p < .001$, *RMSEA* [90% CI] = .012 [.012; .012], *SRMSR* for each country ranged between .027 and .063, *TLI* = .979, *CFI* = .989; see also Table A11 in the ESM), indicating that the model structure is the same across countries. Constraining the discrimination parameters to be equal (metric invariance model) across countries had almost no effect on model fit (C_2 = 11,846.283, *df* = 383, $p < .001$, *RMSEA* [90% CI] = .011 [.011; .011], *SRMSR* for each country ranged between .029 and .078, *TLI* = .983, *CFI* = .985), indicating the same metric of the WHO-5 in the countries. However, constraining the item thresholds to be equal (scalar invariance model) across countries lead to a substantial loss in model fit according to some goodness of fit statistics (C_2 = 68,062.554, *df* = 1,391, $p < .001$, *RMSEA* [90% CI] = .014 [.014; .014], *SRMSR* for each country ranged between .029 and .189, *TLI* = .973, *CFI* = .912), indicating non-invariance for at least some threshold parameters.

We started the alignment method with the FREE approach. However, Mplus provided a warning that the model might be poorly identified, thus, we switched to the FIXED

approach with Poland as reference group as indicated by the Mplus warning. Table 2 shows the fit statistics of the alignment analysis with the FIXED approach and with Poland as reference group (with mean fixed to 0 and variance fixed to 1). The average invariance index (mean over all R^2 values) equaled .564 and 47.2% of the parameters were flagged as being non-invariant. The R^2 values for the item discrimination ranged between .350 and .831 (M_{R2} = .626; SD_{R2} = .180) and the percentage of approximate invariant countries between 62.8% and 88.4%. Interestingly, the discrimination parameter of item 1 showed the lowest R^2 value, but the second highest number of invariant countries compared to the item discrimination parameter of the other items. This indicates that the non-invariance in this parameter came from a few 'outlier' countries.

(Table 2 about here)

The R^2 values for the item thresholds ranged between .000 and. 877 (M_{R2} = .552; SD_{R2} = .288) and the percentage of approximate invariant countries between 30.2% and 62.8%.[8] Figure 1 shows the item parameters of the GRM after the alignment procedure (see also Table A13-17 in the ESM for the exact values). These parameters can be directly compared because of the scale linking via alignment. Figure 1 gives a more nuanced picture of the (non-)Invariance of the parameter and shows which countries were deviating most. For instance, it can be seen that the item discrimination parameters for item 1 in Greenland, Kyrgyzstan, and Tajikistan were clearly non-invariant compared to the other countries (see also Figure A56 in the ESM for a quick overview on (non-)invariant parameters).

(Figure 1 about here)

[8] A simulation study (n_{sim} = 500) revealed a very high factor mean country ranking stability. The correlation between the population factor means (i.e., the means estimated from the original alignment analysis) and the estimated factor means for each replication averaged over replications equaled r = .997. This indicated reliable alignment results (Muthén & Asparouhov, 2014) even though almost half of the parameters were flagged as non-invariant. The proportion of replications for which the 95% confidence interval contains the mean ranged between 93.0% and 96.6% (M = 94.8; SD = 0.01; see Table A12 in the ESM for detailed information).

Figure 2 shows the test characteristic curves of the unidimensional GRM after alignment. Exemplary, it can be seen that at lower levels on the latent variable the expected test scores were especially low for Albania, Kyrgyzstan, and Tajikistan, whereas at higher levels on the latent variable the expected test scores were especially high for North Macedonia.

(Figure 2 about here)

Figure 3 gives a more fine-grained insight in the differential test functioning across countries. It shows the difference in expected test scores dependent on the level of the latent variable together with the *sDRF* and *uDRF* statistics with England as reference group. Negative values indicate that students in England had higher expected test scores, whereas positive values indicate that the other group had higher expected test scores. The *sDRF* and *uDRF* statistics summarize the differential test functioning across the full range of the latent variable. For example, when comparing England and Switzerland only minor differential test functioning effects occurred, whereas differential test functioning is larger between England and Kyrgyzstan. Taking England as reference group, the *sDRF* statistics ranged between -0.34 and 0.79 (M_{sDRF} = 0.12, SD_{sDRF} = 0.25) and the *uDRF* statistics between 0.03 and 1.40 (M_{uDRF} = 0.42, SD_{uDRF} = 0.36).

(Figure 3 about here)

Item properties, information functions and marginal reliability of the aligned IRT model

All items showed high or very high discrimination parameters (Baker, 2001) across all countries (see Figure 1; $M_{\text{item discrimination}}$ = 2.48, $SD_{\text{item discrimination}}$ = .40, $Min_{\text{item discrimination}}$ = 1.68, $Max_{\text{item discrimination}}$ = 3.67). Moreover, all items showed smaller distances between the thresholds b2 and b3, and between b3 and b4. Figure 4 displays the test- and item information

functions of the aligned GRM across countries. The marginal reliability ranged between .80 and .91. In most countries, item 1 provided the most amount of information.

(Figure 4 about here)

The correlations between factor scores and manifest sum scores ranged between .94 and .99 within each country (see Figure A57 in the ESM). Figure 5 shows the association between the WHO–5 country means when the scoring of the WHO–5 is performed via manifest sum scores or via (EAP) factor scores of the aligned GRM. The correlation is very high (r = .97), indicating mostly negligible differences in country ranking between the two scoring methods (see also Figure A58 in the ESM for the order of the countries).

(Figure 5 about here)

Relationship with external criteria

The correlations between the WHO–5 (for manifest sum scores and for factor scores of the aligned GRM) and the other variables are shown in Figure 6. The correlations between the WHO–5 manifest sum scores and gender ranged between –.32 and –.02 (M_{cor} = -.23, SD_{cor} = .06), between the WHO–5 sum scores and age between –.34 and –.04 (M_{cor} = -.20, SD_{cor} = .06), between the WHO–5 sum scores and family affluence between -.02 and .14 (M_{cor} = .07, SD_{cor} = .04), between the WHO–5 sum scores and life satisfaction between .24 and .69 (M_{cor} = .55, SD_{cor} = 0.10), between the WHO–5 sum scores and self-rated health between .05 and .27 (M_{cor} = .18, SD_{cor} = 0.05), between the WHO–5 sum scores and Symptom-Checklist between –.65 and –.25 (M_{cor} = -.53, SD_{cor} = 0.09), and between the WHO–5 sum scores and loneliness between –.60 and –.26 (M_{cor} = -.50, SD_{cor} = .07) across countries. Overall, the differences between the correlations of the manifest sum scores and the criterion variables and the correlations of the factor scores from the aligned GRM and the criterion variables were not substantial. Across all variables and countries, the differences ranged between –.05 and .04, indicating that the correlations derived from the manifest sum scores and those from the factor

scores are nearly equivalent. This suggests that using either scoring method yields similar results when assessing relationships with external criteria.

(Figure 6 about here)

Norm values

Table 3 presents the mean sum scores, standard deviations, and percentiles of the WHO-5 across countries. The country means ranged between 12.7 (Poland) and 19.8 (Tajikistan).[9] See Table A21 in the ESM for WHO-5 norm values stratified by age groups and gender.[10]

(Table 3 about here)

Discussion

In the present study using national representative samples of adolescents from 43 European and Central Asia countries as well as Canada, we provided an in-depth psychometric analysis of the WHO–5 Well-being Index by investigating its dimensionality and measurement structure, item properties, reliability, cross-national measurement invariance, and its nomological network in adolescent samples.

[9] Figure 5 indicates that Tajikistan is an outlier in terms of the WHO-5's sample mean. We can only speculate whether this difference in mean level is due to cultural factors, response styles, social desirability, translation issues, or data quality concerns (e.g., Chen, 2008; Javeline, 1999). Furthermore, considering that the WHO-5 in Tajikistan showed weaker associations with external criteria, we urge caution when interpreting and applying these norm values.

[10] We conducted country-wise analyses of measurement invariance and differential item/test functioning across gender and age groups. Regarding gender, comparing the configural and metric invariance models, Δ*CFI* ranged between .000 and .002 across countries. Comparing the metric and scalar invariance models, Δ*CFI* ranged between .004 and .020 ($M_{country}$ = .009; $SD_{country}$ = .003). These findings suggest that discrimination parameters were generally invariant, while at least some threshold parameters showed non-invariance across gender in certain countries (see Table A18 in the ESM). A subsequent alignment analysis (FIXED approach with boys as reference group) revealed that the percentage of non-invariant parameters across gender ranged from 0% to 40% across countries. Despite this, *uDRF* values ranged between 0.09 and 0.41 (M_{uDRF} = 0.18, SD_{uDRF} = 0.072), indicating negligible differential test functioning across gender within each country. Regarding age groups, comparing the configural and metric invariance models, Δ*CFI* ranged between .000 and .003 across countries. Comparing the metric and scalar invariance models, Δ*CFI* ranged between .002 and .042 ($M_{country}$ = .018; $SD_{country}$ = .009). These results suggest that discrimination parameters were generally invariant, while at least some threshold parameters showed non-invariance across age groups in certain countries (see Table A19 in the ESM). A subsequent alignment analysis (FIXED approach with children aged 15 and above as reference group) revealed that the percentage of non-invariant parameters across age ranged from 0% to 42.2% across countries. However, *uDRF* values ranged between 0.07 and 0.78 (M_{uDRF} = 0.38, SD_{uDRF} = 0.22), again indicating negligible differential test functioning across age groups within each country.

Dimensionality, measurement structure, item properties, and reliability

In line with previous research (e.g., Cosma et al., 2022; deWit et al., 2007), EGA dimensionality analysis revealed that the WHO–5 items map onto one latent (hedonic well-being) dimension in all countries. The GRM showed a good model-data fit in most countries, and its assumptions (i.e., local independence, monotonicity, item fit) were mostly met. Regarding the local independence assumption, items 1 and 2 showed the highest amount of local dependence within most countries. This finding is in line with recent research (Cosma et al., 2022). However, contrary to Cosma et al. (2022), our results indicate that the violation of the local independence assumption seems to be negligible and suggests that all five items of the WHO-5 should be retained in order to maintain construct depth.

All items exhibited substantial discriminatory power across all countries (i.e., $a \geq 1.68$; e.g., Bakker, 2001). This indicates that all items are relevant indicators of hedonic well-being in all countries. Our results mirror previous research on the WHO-5 measurement structure within adolescence (Quansah et al., 2022) and adults (Sischka et al., 2020) and showed somewhat smaller distances between the thresholds b2 (*some of the time* vs. *less than half the time*) and b3 (*less than half the time* vs. *more than half the time*), and between b3 and b4 (*more than half the time* vs. *most of the time*). The marginal reliability was high for all countries (≥ 80). Moreover, test information values of the WHO-5 were also sufficiently high (≥ 4, O'Connor, 2018) over different levels of the latent variable in all countries.

Cross-country measurement invariance and differential item/test functioning

Overall, the multigroup IRT analysis indicated that the WHO-5 exhibits configural and metric, but not scalar cross-country invariance. However, a subsequent alignment procedure revealed some amount of non-invariance at the item discrimination and threshold level (47.2% of all parameters were non-invariant). Especially item 1 showed some cross-country item discrimination variability (e.g., Greenland, Tajikistan). Moreover, all items

showed a higher amount of threshold non-invariance for the first (b1) and last (b5) threshold compared to the other thresholds. Consequently, differential test functioning was particularly pronounced at the extreme levels of the latent continuum. A finding that has also been shown in adult samples (Sischka et al., 2020). Overall, the results indicated small to moderate differential test functioning across countries. For instance, data from England and Albania showed an average deviation of 0.79 (99% CI 0.63; 0.96]) points from the WHO–5 expected test scores (ranging between 0 and 25) for the same level on the latent variable.[11] Especially Armenia, Kyrgyzstan, and Tajikistan showed some amount of non-invariance compared to the other countries.

Relationship with external criteria

Overall, the WHO-5 showed the expected associations with the external criteria (i.e., moderate to strong positive correlations with life satisfaction and self-rated health; and negative correlations with Symptom Checklist and loneliness). These associations have in part already been established in both adolescent (e.g., Cosma et al., 2022) and adult (e.g., Aliyev et al., 2024; Sischka, Schmidt et al., 2020) populations. In line with expectations, on average, boys (Salk et al., 2017), younger students (González-Carrasco et al., 2017; Michel et al., 2009), and students with a higher socioeconomic status (Sweeting & Hunt, 2014) scored higher on the WHO-5.

Correlational analysis revealed a similar nomological net of the WHO–5 across countries, with the exceptions of Greenland and Tajikistan that showed substantial lower correlations with some other measures (i.e., Symptom-Checklist and loneliness). The scoring

[11] See also Figure 3. As a reading example: Respondents from Albania yielded expected test scores that were up to 1.06 (99% CI [0.57; 1.55]) lower at lower levels of the latent variable, whereas expected test scores that were up to 1.27 (99% CI [0.97; 1.58]) points higher at higher levels of the latent variable compared with respondents from England. This means that respondents from Albania and England who have the same level on the latent variable will have different WHO–5 sum scores (e.g., 3.64 vs. 4.69 at the lower level and 22.00 vs. 20.72) at the upper level.

method of the WHO-5 (GRM alignment factor scores versus manifest sum scores) had almost no effects on the associations between the WHO-5 and the external criteria.

Study strengths, limitations, constraints to generality, and future research

One strength of the current study is the alignment in research protocol across countries during data collection, which ensures functional equivalence (Schnohr et al., 2015) and makes cross-national data more comparable. Another strength of the HBSC-study is the nationally representative datasets across 43 countries with large sample size for all included countries (i.e., *n* ranged between 1,229 and 34,427). Therefore, the analysis has obtained reasonable item parameter recovery (Ostini et al., 2015) and the statistical power was large enough to detect even small differential item/test functioning across countries (Nguyen et al., 2014).

One limitation concerns the included countries as many of them represent WEIRD societies (Western, Educated, Industrialized, Rich, and Democratic; Henrich et al., 2010). While our sample included a few lower-middle-income and middle-income countries, there was no representation from low-income countries or regions in the Global South. Most of the participating countries were located in Europe and Central Asia, resulting in a somewhat limited cultural and economic diversity. Thus, future research might test the WHO-5's psychometric properties and measurement invariance in adolescent populations across a wider range of countries (e.g., in South America, Asia, and Africa). This would further enhance our understanding of the measure's applicability and validity across diverse cultural and economic contexts. Moreover, the current study only used cross-sectional data, which limits the conclusions regarding the causal direction of the associations between the WHO-5 and the external criteria. In addition, test-retest reliability, i.e., temporal stability and the longitudinal factor structure, remains unknown. Thus, future studies might investigate the predictive evidence of validity (Cooper, 2019) and temporal stability as well as temporal invariance (Widaman et al., 2010) of the WHO-5 in adolescent populations. Finally, although the WHO–

5 was originally developed as a generic, global measure of subjective well-being, it has frequently been applied and tested as a screening tool for depression, especially in adult populations (e.g., Topp et al., 2015). Since the HBSC data lacks a gold standard measure to assess depression (e.g., structured clinical interviews), the sensitivity and specificity of the WHO–5 for detecting this disorder in adolescents remain unknown. Therefore, the present study is unable to derive or recommend cutoff values for the WHO–5 to identify depression in this age group (but see Allgaier et al., 2012; Blom et al., 2012; Tittel et al., 2023 for adolescent-specific applications). Nevertheless, previous research in adults has demonstrated that the WHO–5 shows adequate sensitivity and specificity in identifying depressive symptoms (Topp et al., 2015), comparable to widely used tools like the PHQ-9 (He et al., 2020) and the Beck Depression Inventory Revised (von Glischinski et al., 2019).

Implications: The WHO-5 in applied research

The current study indicates that the psychometric properties of the WHO-5 are robust against different cultural/language contexts and thus it can be applied to cross-country research on adolescent mental health/well-being. It is also worth noting that sum/mean scoring is a suitable scoring method for the WHO-5 across all countries that introduces negligible bias. This makes it a suitable measure for applied researchers and practitioners with limited psychometric knowledge. To reiterate, the WHO-5 is especially useful when the aim is to assess hedonic well-being. Compared to previous research (Cosma et al., 2022), our results do not indicate the need to remove items of the WHO-5. Instead, it can be used in its current form. In order to facilitate meaningful assessments and comparisons of individual or group scores, general and age/gender-specific norm values for the WHO-5 were provided. Even if cross-county comparisons in subjective well-being may be less of interest to each participating country, the development of standards - as used for adults in the WHO–5 - and hence as a screening tool for depression too, is an important next step.

Conclusion

Due to its brevity and use of non-invasive questions, the WHO–5 Well-Being Index seems to be an optimal measure to assess hedonic well-being in adolescence. The current study revealed that it exhibits a unidimensional factor structure, a high degree of cross-country measurement invariance, high reliability, and a similar nomological network across countries. The WHO–5 Well-Being Index is a psychometrically sound measure that has shown promise for cross-cultural research among adolescents in the included European and Central Asian countries. However, given the limited representation of lower-income countries and regions outside of Europe and Central Asia, further research is needed to establish its applicability and validity in more diverse cultural, geographic, and economic contexts.

Software Information

Data analysis was done in R (Version 4.3.1; R Core Team, 2023 and Mplus (v8.8; Muthén & Muthén, 1998-2017). Data transformations were done with the *tidyverse (Wickham et al., 2019)*, *car* (Fox & Weisberg, 2019), *labelled* (Larmarange, 2021), and *sjlabelled* (Lüdecke, 2021) packages. Descriptive statistics were calculated with the *weights (Pasek et al., 2021)* and the *Weighted.Desc.Stat* (Parchami, 2016) packages. Dimensionality assessment was done with the *EGAnet* (Golino & Christensen, 2022) package. Item response analyses were done with the *mirt* (Chalmers, 2012) and *irtQ* (Lim et al., 2023) packages. The graphs were created with the *ggplot2 (Wickham, 2016)* and *ggpubr* (Kassambara, 2020) packages. The alignment analysis was done in Mplus and read in R with the package *MplusAutomation* (Hallquist & Wiley, 2018).

References

Adjorlolo, S., & Anum, A. (2021). Positive and negative psychosis risk symptoms among adolescents in Ghana. *International Journal of Adolescence and Youth*, *26*(1), 307–320. https://doi.org/10.1080/02673843.2021.1933110

Aliyev, B., Rustamov, E., Satici, S. A., & Zalova Nuriyeva, U. (2024). Azerbaijani adaptation of the WHO-5 wellbeing index: investigating its relationship with psychological distress, resilience, and life satisfaction. *BMC Psychology*, *12*. Article 100. https://doi.org/10.1186/s40359-024-01593-0

Allgaier, A.-K., Pietsch, K., Frühe, B., Prast, E., Sigl-Glöckner, J., & Schulte-Körne, G. (2012). Depression in pediatric care: Is the WHO-Five Well-Being Index a valid screening instrument for children and adolescents? *General Hospital Psychiatry*, *34*(3), 234–241. https://doi.org/10.1016/j.genhosppsych.2012.01.007

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling*, *21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

Baker, F. B. (2001). *The basics of item response theory* (2nd ed). ERIC Clearinghouse on Assessment and Evaluation.

Bech, P. (2012). *Clinical Psychometrics.* John Wiley and Sons.

Blom, E. H., Bech, P., Högberg, G., Larsson, J. O., & Serlachius, E. (2012). Screening for depressed mood in an adolescent psychiatric context by brief self-assessment scales – testing psychometric validity of WHO–5 and BDI-6 indices by latent trait analyses. *Health and Quality of Life Outcomes*, *10*(1), 149. https://doi.org/10.1186/1477-7525-10-149

Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology*, *49*(5), 713-734. https://doi.org/10.1177/0022022117749042

Boer, M., Moreno-Maldonado, C., Dierckens, M., Lenzi, M., Currie, C., Residori, C., Bosáková, L., Berchialla, P., Eida, T., & Stevens, G. (2023). The Implications of the COVID-19 Pandemic for the Construction of the Family Affluence Scale: Findings

from 16 Countries. *Child Indicators Research, 17*, 395–418. https://doi.org/10.1007/s12187-023-10082-6

Borgers, N., De Leeuw, E., & Hox, J. (2000). Children as respondents in survey research: Cognitive development and response quality 1. *Bulletin of Sociological Methodology/Bulletin de méthodologie sociologique*, *66*(1), 60-75. https://doi.org/10.1177/075910630006600l0

Brown, A. (2018). Item response theory approaches to test scoring and evaluating the score accuracy. In F. P. Irwing, T. Booth, & D. J. Hughes (eds..), *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale, and Test Development* (pp. 607–638). Wiley Blackwell.

Cai, L., & Monroe, S. (2014). *A new statistic for evaluating item response theory models for ordinal data*. https://files.eric.ed.gov/fulltext/ED555726.pdf

Chalmers, R. P. (2018). Model-based measures for detecting and quantifying response bias. *Psychometrika*, *83*(3), 696–732. https://doi.org/10.1007/s11336-018-9626-9

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*(5), 1005–1018. https://doi.org/10.1037/a0013193

Clarke, A., Friede, T., Putz, R., Ashdown, J., Martin, S., Blake, A., Adi, Y., Parkinson, J., Flynn, P., Platt, S., & Stewart-Brown, S. (2011). Warwick-Edinburgh Mental Well-being Scale (WEMWBS): Validated for teenage school students in England and Scotland. A mixed methods assessment. *BMC Public Health, 11*, Article 487. https://doi.org/10.1186/1471-2458-11-487

Conijn, J. M., Smits, N., & Hartman, E. E. (2020). Determining at what age children provide sound self-reports: An illustration of the validity-index approach. *Assessment*, *27*(7), 1604-1618. https://doi.org/10.1177/1073191111983265

Cooper, C. (2019). *Psychological testing: Theory and practice*. Routledge.

Cosma, A., Költő, A., Chzhen, Y., Kleszczewska, D., Kalman, M., & Martin, G. (2022). Measurement invariance of the WHO–5 Well-Being Index: Evidence from 15 European countries. *International Journal of Environmental Research and Public Health*, *19*(16), 9798. https://doi.org/10.3390/ijerph19169798

De Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education, 44*(1), 109–117. https://doi.org/10.1111/j.1365-2923.2009.03425.x

De Wit, M., Pouwer, F., Gemke, R. J. B. J., Delemarre-van De Waal, H. A., & Snoek, F. J. (2007). Validation of the WHO–5 Well-Being Index in adolescents with Type 1 diabetes. *Diabetes Care*, *30*(8), 2003–2006. https://doi.org/10.2337/dc07-0447

De Wit, M., Winterdijk, P., Aanstoot, H. J., Anderson, B., Danne, T., Deeb, L., ... & DAWN Youth Advisory Board. (2012). Assessing diabetes-related quality of life of youth with type 1 diabetes in routine clinical care: the MIND Youth Questionnaire (MY-Q). *Pediatric Diabetes*, *13*(8), 638-646. https://doi.org/10.1111/j.1399-5448.2012.00872.x

DeMars, C. E. (2020). Alignment as an alternative to anchor purification in DIF analyses. *Structural Equation Modeling*, *27*(1), 56–72. https://doi.org/10.1080/10705511.2019.1617151

Depaoli, S., Tiemensma, J., & Felt, J. M. (2018). Assessment of health surveys: Fitting a multidimensional graded response model. *Psychology, Health & Medicine*, *23*(sup1), 13–31. https://doi.org/10.1080/13548506.2018.1447136

Diener, E. (2000). Subjective well-being: The science of happiness and a proposal for a national index. *American Psychologist, 55*(1), 34–43. https://doi.org/10.1037/0003-066X.55.1.34

Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods*, *23*(1), 138–149. https://doi.org/10.1037/met0000121

Elgar, F., Xie, A., Pförtner, T., White, J., & Pickett, K., (2017). Assessing the view from bottom: How to measure socioeconomic position and relative deprivation in adolescents. In *Sage Research Methods Cases Part 2*. SAGE. https://doi.org/10.4135/9781526406347

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PLOS ONE*, *12*(6), e0174035. https://doi.org/10.1371/journal.pone.0174035

Golino, H., Shi, D., Christensen, A. P., Garrido, L. E., Nieto, M. D., Sadana, R., Thiyagarajan, J. A., & Martinez-Molina, A. (2020). Investigating the performance of exploratory graph analysis and traditional techniques to identify the number of latent factors: A simulation and tutorial. *Psychological Methods*, *25*(3), 292–320. https://doi.org/10.1037/met0000255

González-Carrasco, M., Casas, F., Malo, S., Viñas, F., & Dinisman, T. (2017). Changes with Age in Subjective Well-Being Through the Adolescent Years: Differences by Gender. *Journal of Happiness Studies*, *18*, 63–88. https://doi.org/10.1007/s10902-016-9717-1

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling*, *25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

He, C., Levis, B., Riehm, K. E., Saadat, N., Levis, A. W., Azar, M., ... & Benedetti, A. (2020). The accuracy of the Patient Health Questionnaire-9 algorithm for screening to detect major depression: an individual participant data meta-analysis. *Psychotherapy and Psychosomatics*, *89*(1), 25-37. https://doi.org/10.1159/000502294

Heinz, A., Sischka, P. E., Catunda, C., Cosma, A., García-Moya, I., Lyyra, N., Ravens-Sieberer, U., & Pickett, W. (2022). Item response theory and differential test functioning analysis of the HBSC-Symptom-Checklist across 46 countries. *BMC Medical Research Methodology*, *22*(253). https://doi.org/10.1186/s12874-022-01698-3

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

Houts, C. R., Savord, A., & Wirth, R. J. (2022). Overview of modern measurement theory and examples of its use to measure execution function in children. *Journal of Pediatric Neuropsychology*, *8*, 1–14. https://doi.org/10.1007/s40817-021-00117-7

Jacobs, P., Power, L., Davidson, G., Devaney, J., McCartan, C., McCusker, P., & Jenkins, R. (2023). A Scoping Review of Mental Health and Wellbeing Outcome Measures for Children and Young People: Implications for Children in Out-of-home Care. *Journal of Child & Adolescent Trauma*. Advance online publication. https://doi.org/10.1007/s40653-023-00566-6

Jami, W. A., & Kemmelmeier, M. (2020). Assessing well-being across space and time: Measurement equivalence of the WHO–5 in 36 European countries and over 8 years. *Journal of Well-Being Assessment*, *4*(3), 419–445. https://doi.org/10.1007/s41543-021-00042-8

Javeline, D. (1999). Response Effects in Polite Cultures: A Test of Acquiescence in Kazakhstan. *Public Opinion Quarterly*, *63*(1), 1-28. https://doi.org/10.1086/297701

Kang, T., & Chen, T. T. (2011). Performance of the generalized S-X2 item fit index for the graded response model. *Asia Pacific Education Review*, *12*(1), 89–96. https://doi.org/10.1007/s12564-010-9082-4

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*, *24*(4), 524–544. https://doi.org/10.1080/10705511.2017.1304822

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Krause, K. R., Jacob, J., Szatmari, P., & Hayes, D. (2022). Readability of Commonly Used Quality of Life Outcome Measures for Youth Self-Report. *International Journal of Environmental Research and Public Health*, *19*(15), Article 9555. https://doi.org/10.3390/ijerph19159555

Krieger, T., Zimmermann, J., Huffziger, S., Ubl, B., Diener, C., Kuehner, C., & Grosse Holtforth, M. (2014). Measuring depression with a well-being index: Further evidence for the validity of the WHO Well-Being Index (WHO–5) as a measure of the severity of depression. *Journal of Affective Disorders*, *156*, 240–244. https://doi.org/10.1016/j.jad.2013.12.015

Kusier, A. O., & Folker, A. P. (2020). The Well-Being Index WHO–5: hedonistic foundation and practical limitations. Medical Humanities, 46(3), 333–339. https://doi.org/10.1136/medhum-2018-011636

Lambert, M., Fleming, T., Ameratunga, S., Robinson, E., Crengle, S., Sheridan, J., Denny, S., Clark, T., & Merry, S. (2014). Looking on the bright side: An assessment of factors associated with adolescents' happiness. *Advances in Mental Health*, *12*(2), 101–109. https://doi.org/10.1080/18374905.2014.11081888

Lara-Cabrera, M. L., Betancort, M., Muñoz-Rubilar, A., Rodríguez-Novo, N., Bjerkeset, O., & Cuevas, C. D. L. (2022). Psychometric properties of the WHO–5 well-being index among nurses during the COVID-19 pandemic: a cross-sectional study in three

countries. *International Journal of Environmental Research and Public Health*, *19*(16), 10106. https://doi.org/10.3390/ijerph191610106

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.

Low, K.-Y., Pheh, K.-S., & Tan, C.-S. (2023). Validation of the WHO–5 as a screening tool for depression among young adults in Malaysia. *Current Psychology*, *42*(10), 7841–7844. https://doi.org/10.1007/s12144-021-02152-1

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2018). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods*, *23*(3), 524–545. https://doi.org/10.1037/met0000113

McDowell, I. (2010). Measures of self-perceived well-being. *Journal of Psychosomatic Research*, *69*(1), 69-79. https://doi.org/10.1016/j.jpsychores.2009.07.002

McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*, *28*(1), 61–88. https://doi.org/10.1037/met0000425

Michel, G., Bisegger, C., Fuhr, D. C., & Abel, T. (2009). Age and gender differences in health-related quality of life of children and adolescents in Europe: a multilevel analysis. *Quality of Life Research*, *18*(9), 1147–1157. https://doi.org/10.1007/s11136-009-9538-3

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

Monroe, S., & Cai, L. (2015). Evaluating structural equation models for categorical outcomes: A new test statistic and a practical challenge of interpretation. *Multivariate Behavioral Research*, *50*(6), 569–583. https://doi.org/10.1080/00273171.2015.1032398

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology*, *5*, *978*. https://doi.org/10.3389/fpsyg.2014.00978

Muthén, L.K., Muthén, B.O. (2017). *Mplus user's guide*, 8th ed. CA, Muthén and Muthén.

Nguyen, T. H., Han, H.-R., Kim, M. T., & Chan, K. S. (2014). An introduction to item response theory for patient-reported outcome measurement. *The Patient - Patient-Centered Outcomes Research*, *7*, 23–35. https://doi.org/10.1007/s40271-013-0041-0

O'Connor, B. P. (2018). An illustration of the effects of fluctuations in test information on measurement error, the attenuation of effect sizes, and diagnostic reliability. *Psychological Assessment*, *30*(8), 991–1003. https://doi.org/10.1037/pas0000471

Ostini, R., Finkelman, M., & Nering, M. (2015). Selecting among polytomous IRT models. In S. P. Reise & D. A. Revicki (eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 285–304). Routledge, Taylor & Francis Group.

Patton, G. C., Sawyer, S. M., Santelli, J. S., Ross, D. A., Afifi, R., Allen, N. B., ... & Viner, R. M. (2016). Our future: a Lancet commission on adolescent health and wellbeing. *The Lancet*, *387*(10036), 2423-2478. https://doi.org/10.1016/S0140-6736(16)00579-1

Peter, S. C., Whelan, J. P., Pfund, R. A., & Meyers, A. W. (2018). A text comprehension approach to questionnaire readability: An example using gambling disorder measures. *Psychological Assessment*, *30*(12), 1567–1580. https://doi.org/10.1037/pas0000610

Rees, G., & Main, G. (2016). Subjective well-being and mental health. In J. Bradshaw (ed.), *The well-being of children in the UK* (pp. 123-148). Policy Press.

Revicki, D. A., Chen, W. H., & Tucker, C. (2015). Developing item banks for patient-reported health outcomes. In S. P. Reise & D. A. Revicki (eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 335–363). Routledge, Taylor & Francis Group.

Rose, T., Joe, S., Williams, A., Harris, R., Betz, G., & Stewart-Brown, S. (2017). Measuring mental wellbeing among adolescents: A systematic review of instruments. *Journal of*

Child and Family Studies, *26*(9), 2349–2362. https://doi.org/10.1007/s10826-017-0754-0

Schnohr, C. W., Gobina, I., Santos, T., Mazur, J., Alikasifuglu, M., Välimaa, R., Corell, M., Hagquist, C., Dalmasso, P., Movseyan, Y., Cavallo, F., van Dorsselaer, S., & Torsheim, T. (2016). Semantics bias in cross-national comparative analyses: is it good or bad to have "fair" health? *Health and Quality of Life Outcomes*, *14*, 70. https://doi.org/10.1186/s12955-016-0469-8

Schnohr, C. W., Kreiner, S., Due, E. P., Currie, C., Boyce, W., & Diderichsen, F. (2008). Differential item functioning of a family affluence scale: Validation study on data from HBSC 2001/02. *Social Indicators Research*, *89*(1), 79–95. https://doi.org/10.1007/s11205-007-9221-4

Schnohr, C. W., Molcho, M., Rasmussen, M., Samdal, O., de Looze, M., Levin, K., Roberts, C. J., Ehlinger, V., Krolner, R., Dalmasso, P., & Torsheim, T. (2015). Trend analyses in the health behaviour in school-aged children study: methodological considerations and recommendations. *The European Journal of Public Health*, *25*(suppl 2), 7–12. https://doi.org/10.1093/eurpub/ckv010

Sellbom, M., & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological Assessment, 31*(12), 1428–1441. https://doi.org/10.1037/pas0000623

Shaffer-Hudkins, E., Suldo, S., Loker, T., & March, A. (2010). How adolescents' mental health predicts their physical health: Unique contributions of indicators of subjective well-being and psychopathology. *Applied Research in Quality of Life, 5*(3), 203-217. https://doi.org/10.1007/s11482-010-9105-7

Shi, D., & Maydeu-Olivares, A. (2020). The effect of estimation methods on SEM fit indices. *Educational and Psychological Measurement*, *80*(3), 421-445. https://doi.org/10.1177/0013164419885164

Sischka, P. E., Costa, A. P., Steffgen, G., & Schmidt, A. F. (2020). The WHO–5 well-being index – validation based on item response theory and the analysis of measurement invariance across 35 countries. *Journal of Affective Disorders Reports*, *1*, 100020. https://doi.org/10.1016/j.jadr.2020.100020

Sischka, P. E., Schmidt, A. F., & Steffgen, G. (2020). Further Evidence for Criterion Validity and Measurement Invariance of the Luxembourg Workplace Mobbing Scale. *European Journal of Psychological Assessment*, *36*(1), 32–43. https://doi.org/10.1027/1015-5759/a000483

Sischka, P. E., Grübbel, L., Reisinger, C., Neufang, K.M., & Schmidt, A.F. (2024, in press). On the dimensionality, suitability of sum/mean scores, and cross-country measurement invariance of the Perceived Stress Scale 10 (PSS-10) – Evidence from 41 countries. *International Journal of Stress Management*.

Sweeting, H., & Hunt, K. (2014). Adolescent socio-economic and school-based social status, health and well-being. *Social Science and Medicine*, *121*, 39–47. https://doi.org/10.1016/j.socscimed.2014.09.037

Taber, S. M. (2010). The veridicality of children's reports of parenting: A review of factors contributing to parent–child discrepancies. *Clinical Psychology Review*, *30*(8), 999-1010. https://doi.org/10.1016/j.cpr.2010.06.014

Tejada-Gallardo, C., Blasco-Belled, A., Torrelles-Nadal, C., & Alsinet, C. (2020). Effects of school-based multicomponent positive psychology interventions on well-being and distress in adolescents: A systematic review and meta-analysis. *Journal of Youth and Adolescence, 49*(10), 1943-1960. https://doi.org/10.1007/s10964-020-01289-9

Tittel, S. R., Kulzer, B., Warschburger, P., Merz, U., Galler, A., Wagner, C., ... & Holl, R. W. (2023). The WHO-5 well-being questionnaire in type 1 diabetes: screening for depression in pediatric and young adult subjects. *Journal of Pediatric Endocrinology and Metabolism*, *36*(4), 384-392. https://doi.org/10.1515/jpem-2023-0013

Toland, M. D. (2014). Practical guide to conducting an item response theory analysis. *The Journal of Early Adolescence*, *34*(1), 120–151. https://doi.org/10.1177/0272431613511332

Tomás, J. M., Gutiérrez, M., Pastor, A. M., & Sancho, P. (2020). Perceived social support, school adaptation and adolescents' subjective well-being. *Child Indicators Research, 13*, 1597-1617. https://doi.org/10.1007/s12187-020-09717-9

Topp, C. W., Østergaard, S. D., Søndergaard, S., & Bech, P. (2015). The WHO–5 Well-Being Index: A systematic review of the literature. *Psychotherapy and Psychosomatics*, *84*(3), 167–176. https://doi.org/10.1159/000376585

Torsheim, T., Cavallo, F., Levin, K. A., Schnohr, C., Mazur, J., Niclasen, B., Currie, C., & the FAS Development Study Group. (2016). Psychometric validation of the Revised Family Affluence Scale: A latent variable approach. *Child Indicators Research*, *9*(3), 771–784. https://doi.org/10.1007/s12187-015-9339-x

von Glischinski, M., von Brachel, R., & Hirschfeld, G. (2019). How depressed is "depressed"? A systematic review and diagnostic meta-analysis of optimal cut points for the Beck Depression Inventory revised (BDI-II). *Quality of Life Research*, *28*, 1111-1118. https://doi.org/10.1007/s11136-018-2050-x

Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology*, *44*(3), 219-246. https://doi.org/10.1177/0095798418771807

Wells, C. S., & Hambleton, R. K. (2016). Model fit with residual analyses. In W. J. van der Linden (Ed.), *Handbook of item response theory, Volume two: Statistical tools,* (pp. 395-413). CRC Press.

Wickham, H. (2016). *ggplot2. Elegant graphics for data analysis*. Springer.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10-18. https://doi.org/10.1111/j.1750-8606.2009.00110.x

Winzer, R., Vaez, M., Lindberg, L., & Sorjonen, K. (2021). Exploring associations between subjective well-being and personality over a time span of 15–18 months: a cohort study of adolescents in Sweden. *BMC Psychology*, *9*, Article 173. https://doi.org/10.1186/s40359-021-00673-9

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*(1), 58–79. https://doi.org/10.1037/1082-989x.12.1.58

World Health Organization. (1998). *Wellbeing measures in primary health care/the DepCare Project: report on a WHO meeting: Stockholm, Sweden, 12–13 February 1998*. World Health Organization. Regional Office for Europe. https://iris.who.int/handle/10665/349766

Tables

Table 1. Goodness of fit statistics for the graded response model.

Country	C_2	*p*	*RMSEA* [90% CI]	*SRMSR*	*TLI*	*CFI*
ALB - Albania (*n* = 5,122)	122.604	< .001	.068 [.058; .078]	.045	.981	.990
ARM - Armenia (*n* = 3,975)	37.835	< .001	.041 [.029; .053]	.047	.992	.996
AUT - Austria (*n* = 5,048)	149.677	< .001	.076 [.066; .086]	.031	.980	.990
BEL_Fl - Belgium (Flemish) (*n* = 9,156)	444.629	< .001	.098 [.090; .106]	.037	.965	.983
BEL_Fr - Belgium (French) (*n* = 5,625)	165.164	< .001	.075 [.066; .086]	.030	.974	.987
BGR - Bulgaria (*n* = 2,318)	188.000	< .001	.126 [.111; .141]	.063	.963	.981
CAN - Canada (*n* = 11,510)	365.132	< .001	.079 [.072; .086]	.028	.983	.992
CHE - Switzerland (*n* = 6,607)	364.700	< .001	.104 [.095; .114]	.036	.968	.984
CYP - Cyprus (*n* = 4,535)	138.042	< .001	.077 [.066; .088]	.034	.980	.990
CZE - Czechia (*n* = 12,176)	472.966	< .001	.088 [.081; .094]	.033	.971	.986
DEU - Germany (*n* = 6,282)	127.861	< .001	.063 [.053; .072]	.027	.985	.993
DNK - Denmark (*n* = 3,299)	183.380	< .001	.104 [.091; .117]	.039	.957	.978
ESP - Spain (*n* = 4,787)	172.310	< .001	.096 [.084; .109]	.045	.973	.986
EST - Estonia (*n* = 3,624)	177.511	< .001	.085 [.074; .096]	.028	.980	.990
FIN - Finland (*n* = 3,309)	65.024	< .001	.060 [.048; .074]	.050	.988	.994
FRA - France (*n* = 4,778)	233.072	< .001	.098 [.087; .109]	.038	.961	.981
GB_ENG - England (*n* = 4,001)	180.488	< .001	.094 [.082; .106]	.037	.975	.987
GB_SCT - Scotland (*n* = 4,093)	86.242	< .001	.063 [.052; .075]	.034	.988	.994
GB_WLS - Wales (*n* = 34,427)	1025.586	< .001	.077 [.073; .081]	.038	.981	.991
GRC - Greece (*n* = 1,229)	256.563	< .001	.091 [.082; .101]	.037	.973	.986
GRL - Greenland (*n* = 6,023)	38.259	< .001	.074 [.053; .096]	.058	.981	.990
HRV - Croatia (*n* = 5,175)	181.241	< .001	.083 [.072; .093]	.031	.979	.989
HUN - Hungary (*n* = 3,852)	153.860	< .001	.088 [.076; .100]	.049	.978	.989
IRL - Ireland (*n* = 3,361)	95.487	< .001	.073 [.061; .087]	.032	.985	.992
ISL - Iceland (*n* = 9,300)	164.819	< .001	.059 [.051; .066]	.040	.988	.994
ITA - Italy (*n* = 4,453)	201.581	< .001	.094 [.083; .105]	.038	.967	.984
KAZ - Kazakhstan (*n* = 9,647)	101.749	< .001	.052 [.044; .061]	.057	.989	.994
KGZ - Kyrgyzstan (*n* = 7,119)	123.692	< .001	.050 [.042; .057]	.037	.986	.993
LTU - Lithuania (*n* = 4,851)	147.318	< .001	.077 [.066; .087]	.034	.977	.989
LUX - Luxembourg (*n* = 4,078)	111.140	< .001	.072 [.061; .084]	.028	.980	.990
LVA - Latvia (*n* = 5,656)	269.314	< .001	.097 [.087; .107]	.055	.978	.989
MDA - Moldova (*n* = 5,491)	92.835	< .001	.057 [.047; .067]	.038	.984	.992
MKD - North Macedonia (*n* = 4,017)	58.525	< .001	.052 [.040; .064]	.033	.989	.995
MLT - Malta (*n* = 3,188)	170.531	< .001	.102 [.089; .115]	.053	.971	.986
NLD - Netherlands (*n* = 4,197)	98.534	< .001	.067 [.056; .079]	.035	.984	.992
NOR - Norway (*n* = 3,114)	141.001	< .001	.093 [.081; .107]	.043	.964	.982
POL - Poland (*n* = 5,173)	211.082	< .001	.089 [.079; .100]	.039	.976	.988
PRT - Portugal (*n* = 5,182)	164.522	< .001	.078 [.068; .089]	.031	.979	.989
ROU - Romania (*n* = 7,839)	236.105	< .001	.077 [.069; .085]	.038	.978	.989
SVK - Slovakia (*n* = 4,109)	171.245	< .001	.079 [.069; .090]	.030	.978	.989
SVN - Slovenia (*n* = 6,141)	166.514	< .001	.073 [.063; .082]	.029	.984	.992
SWE - Sweden (*n* = 5,301)	118.495	< .001	.074 [.063; .086]	.036	.981	.990
TJK - Tajikistan (*n* = 6,221)	78.040	< .001	.048 [.039; .058]	.055	.991	.995

Notes. *df* = 5; *RMSEA* = root mean squared error of approximation; *SRMR* = standardized root mean square residual; *TLI* = Tucker-Lewis index; *CFI* = comparative fit index.

Table 2. Alignment fit statistics.

Item	Parameter	R^2	Weighted mean across invariant groups	Weighted standard deviation across invariant groups	Weighted mean across all groups	Weighted standard deviation across all groups	Number (percentage) of approx. invariant groups
Item 1	Discrimination	.350	3.16	0.15	3.07	0.28	33 (76.7%)
	Threshold 1	.000	-2.04	0.16	-1.90	0.24	27 (62.8%)
	Threshold 2	.000	-0.88	0.08	-0.85	0.19	20 (46.5%)
	Threshold 3	.109	-0.47	0.04	-0.42	0.11	13 (30.2%)
	Threshold 4	.726	0.18	0.04	0.25	0.13	19 (44.2%)
	Threshold 5	.547	1.79	0.11	1.55	0.26	15 (34.9%)
Item 2	Discrimination	.672	2.54	0.15	2.50	0.26	30 (69.8%)
	Threshold 1	.047	-1.64	0.12	-1.59	0.22	24 (55.8%)
	Threshold 2	.355	-0.61	0.08	-0.60	0.16	21 (48.8%)
	Threshold 3	.872	-0.04	0.03	-0.06	0.07	26 (60.5%)
	Threshold 4	.877	0.64	0.08	0.62	0.10	26 (60.5%)
	Threshold 5	.600	1.92	0.16	1.80	0.29	20 (46.5%)
Item 3	Discrimination	.572	2.58	0.13	2.45	0.26	27 (62.8%)
	Threshold 1	.330	-1.67	0.13	-1.62	0.20	21 (48.8%)
	Threshold 2	.463	-0.73	0.06	-0.69	0.17	21 (48.8%)
	Threshold 3	.814	-0.17	0.03	-0.16	0.08	25 (58.1%)
	Threshold 4	.771	0.46	0.05	0.49	0.12	25 (58.1%)
	Threshold 5	.425	1.60	0.22	1.47	0.27	18 (41.9%)
Item 4	Discrimination	.831	2.26	0.08	2.23	0.12	38 (88.4%)
	Threshold 1	.820	-0.99	0.07	-0.93	0.11	23 (53.5%)
	Threshold 2	.839	-0.10	0.03	-0.11	0.10	20 (46.5%)
	Threshold 3	.819	0.39	0.04	0.37	0.13	21 (48.8%)
	Threshold 4	.713	0.97	0.07	0.90	0.22	21 (48.8%)
	Threshold 5	.549	1.83	0.11	1.80	0.34	19 (44.2%)
Item 5	Discrimination	.707	2.21	0.09	2.12	0.18	30 (69.8%)
	Threshold 1	.330	-1.71	0.11	-1.67	0.19	22 (51.2%)
	Threshold 2	.524	-0.59	0.06	-0.60	0.15	18 (41.9%)
	Threshold 3	.845	-0.09	0.03	-0.10	0.09	23 (53.5%)
	Threshold 4	.822	0.51	0.05	0.54	0.13	19 (44.2%)
	Threshold 5	.595	1.70	0.10	1.57	0.23	16 (37.2%)

Notes. *MLR* estimator; FIXED approach; POL as reference group. Items (1) *cheerful*, (2) *calm and relaxed*, (3) *active and vigorous*, (4) *fresh and rested*, (5) *interest*.

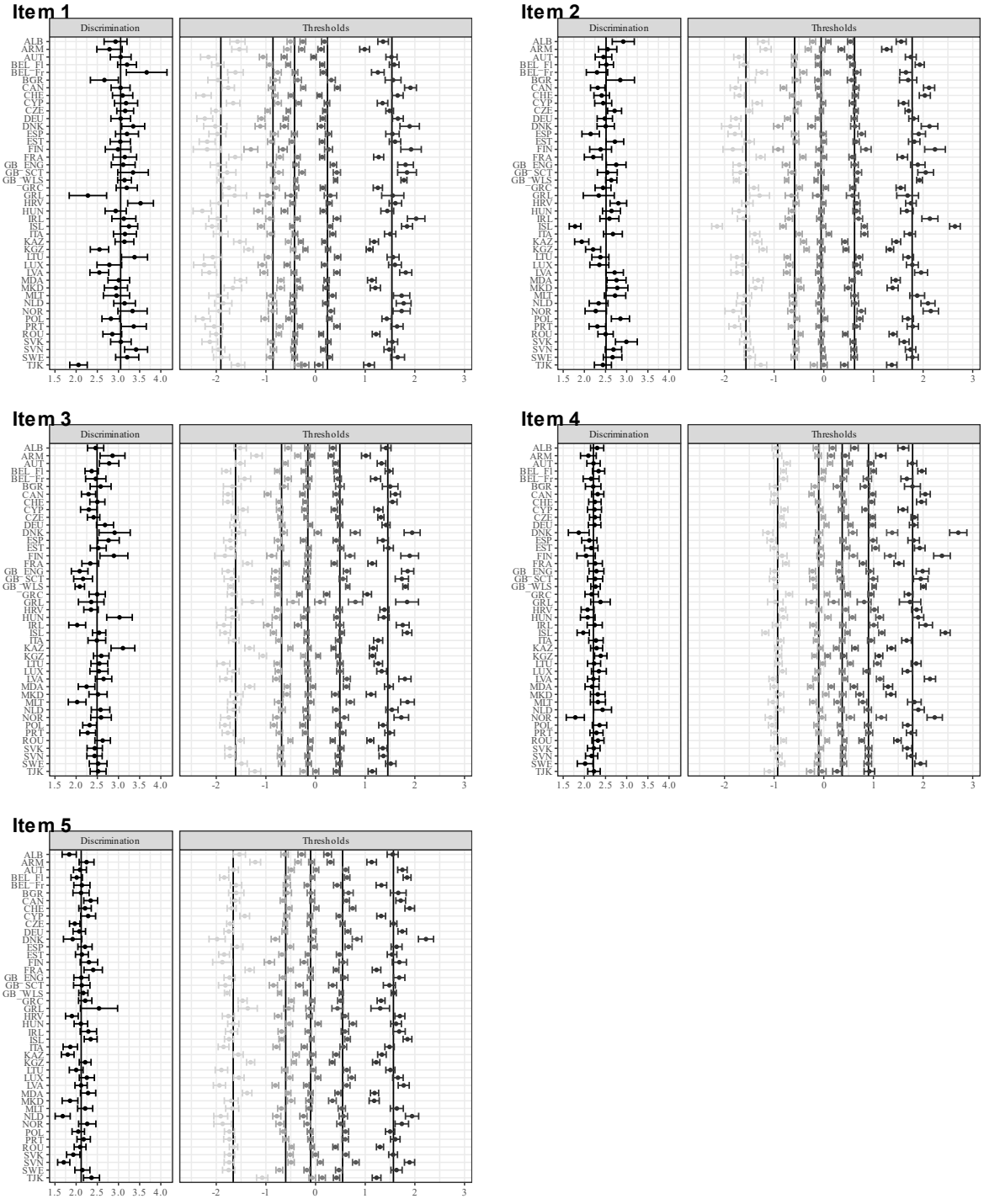
Table 3. Mean, standard deviation, and percentile norms of the WHO-5 Well-Being Index across countries.

Country	*M* (*SD*)	10%	20%	30%	40%	50%	60%	70%	80%	90%
ALB	17.9 (5.2)	10	14	16	18	19	20	21	22	24
ARM	17.8 (5.7)	10	13	15	17	19	20	22	23	25
AUT	14.0 (5.7)	6	9	11	13	14	16	18	19	21
BEL_Fl	15.3 (5.3)	8	11	13	14	16	17	19	20	22
BEL_Fr	14.7 (5.4)	7	10	12	14	15	17	18	19	21
BGR	15.6 (6.6)	6	10	12	15	16	18	20	22	25
CAN	14.9 (5.8)	7	10	12	14	15	17	19	20	22
CHE	14.7 (5.7)	6	9	12	14	15	17	18	20	22
CYP	15.5 (5.9)	7	10	13	14	16	18	19	21	23
CZE	13.8 (5.5)	6	9	11	13	14	16	17	19	21
DEU	14.4 (5.3)	7	10	12	13	15	16	18	19	21
DNK	15.6 (4.5)	9	12	14	15	16	17	18	19	21
ESP	14.7 (6.0)	6	9	11	13	15	17	18	20	22
EST	14.2 (6.0)	6	8	11	13	15	16	18	20	22
FIN	15.2 (5.0)	8	11	13	14	15	17	18	20	21
FRA	13.9 (6.0)	6	8	11	13	14	16	18	19	21
GB_ENG	13.3 (5.9)	5	8	10	12	14	15	17	19	20
GB_SCT	14.5 (5.7)	6	9	12	13	15	17	18	20	21
GB_WLS	14.4 (5.6)	7	9	11	13	15	16	18	20	21
GRC	14.9 (6.1)	6	9	11	14	15	17	19	21	23
GRL	17.1 (5.5)	10	13	15	16	18	19	20	22	24
HRV	16.3 (5.5)	8	11	14	15	17	18	20	21	23
HUN	14.4 (5.9)	6	9	11	13	15	16	18	20	22
IRL	14.9 (5.7)	7	10	12	14	15	17	18	20	22
ISL	15.2 (5.2)	8	11	13	14	16	17	18	20	21
ITA	13.8 (5.6)	6	9	11	13	14	16	17	19	21
KAZ	17.9 (5.6)	10	13	15	17	19	20	22	23	25
KGZ	17.7 (5.5)	10	13	15	17	19	20	21	23	24
LTU	14.8 (5.3)	8	10	12	14	15	16	18	20	22
LUX	14.7 (5.4)	7	10	12	14	15	17	18	20	21
LVA	14.9 (6.3)	6	9	12	14	15	17	19	20	24
MDA	16.5 (5.3)	9	12	14	16	17	19	20	21	23
MKD	17.5 (5.6)	9	13	15	17	18	20	21	23	25
MLT	14.8 (6.1)	6	9	12	14	15	17	19	20	23
NLD	15.6 (5.4)	8	11	13	15	16	18	19	20	22
NOR	15.7 (4.9)	9	12	14	15	16	17	18	20	21
POL	12.7 (6.0)	5	7	9	11	13	14	16	18	21
PRT	15.5 (5.5)	8	10	13	15	16	18	19	20	22
ROU	15.1 (6.1)	6	9	12	14	15	17	19	21	23
SVK	14.8 (5.6)	7	10	12	14	15	17	18	20	22
SVN	13.3 (6.1)	5	7	10	12	14	15	17	19	21
SWE	15.2 (5.5)	7	10	12	14	16	17	19	20	22
TJK	19.8 (5.6)	12	15	18	20	21	23	24	25	25

Notes. *M* = Mean; *SD* = Standard deviation.

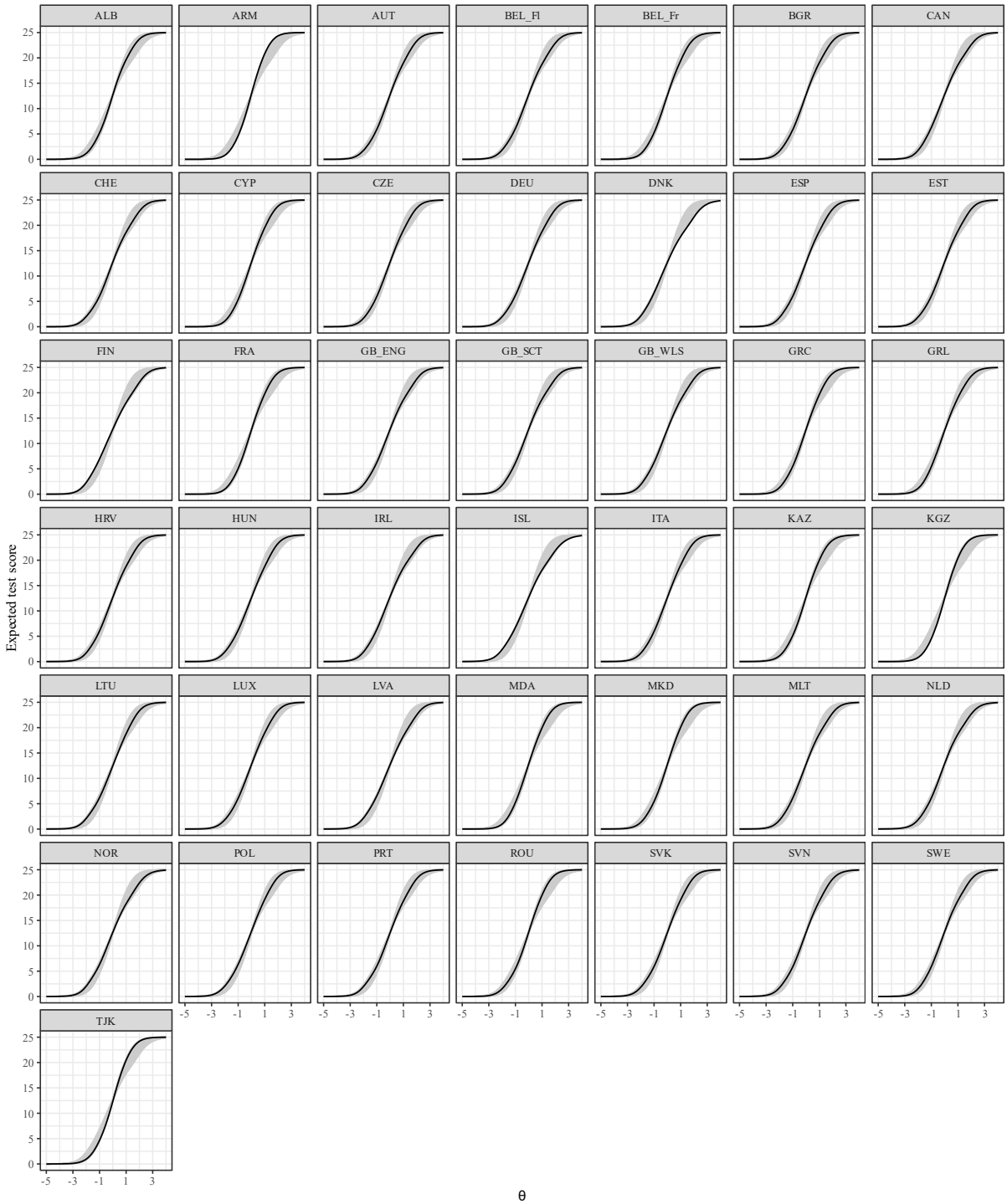
Figures

Figure 1. Item parameter for the GRM after alignment.



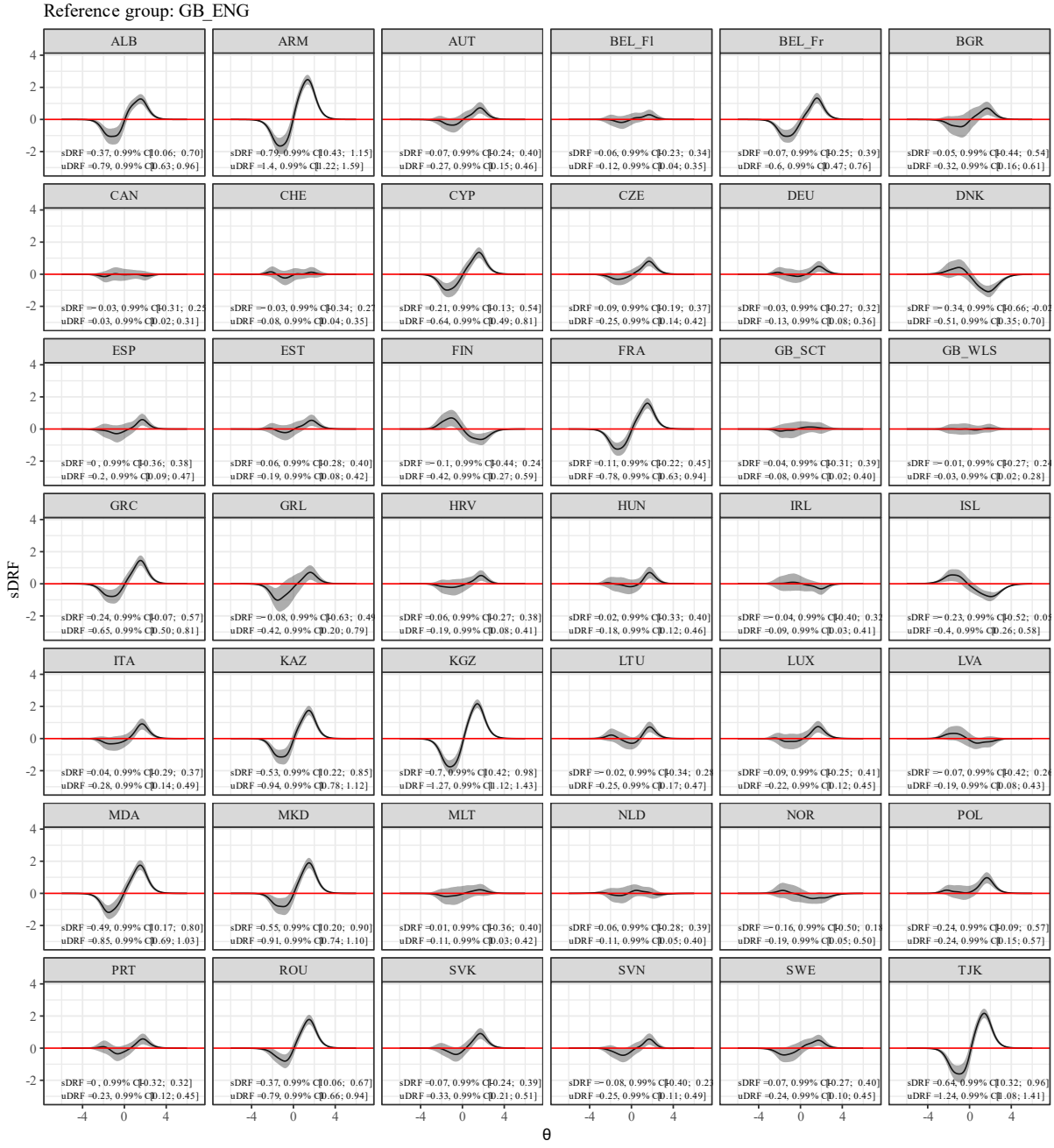
Notes. Item discrimination and threshold parameters with 99% CI. Vertical lines represent the weighted average across all groups. Items (1) *cheerful*, (2) *calm and relaxed*, (3) *active and vigorous*, (4) *fresh and rested*, (5) *interest*. The different grey colors represent the different thresholds. From light grey to dark grey: "At no time" vs. "Some of the time", "Some of the time" vs "Less than half of the time ", "Less than half of the time" vs. "More than half of the time", "More than half of the time" vs. "Most of the time ", and "Most of the time " vs. "All of the time ".

Figure 2. Test characteristic curves for the GRM after alignment.



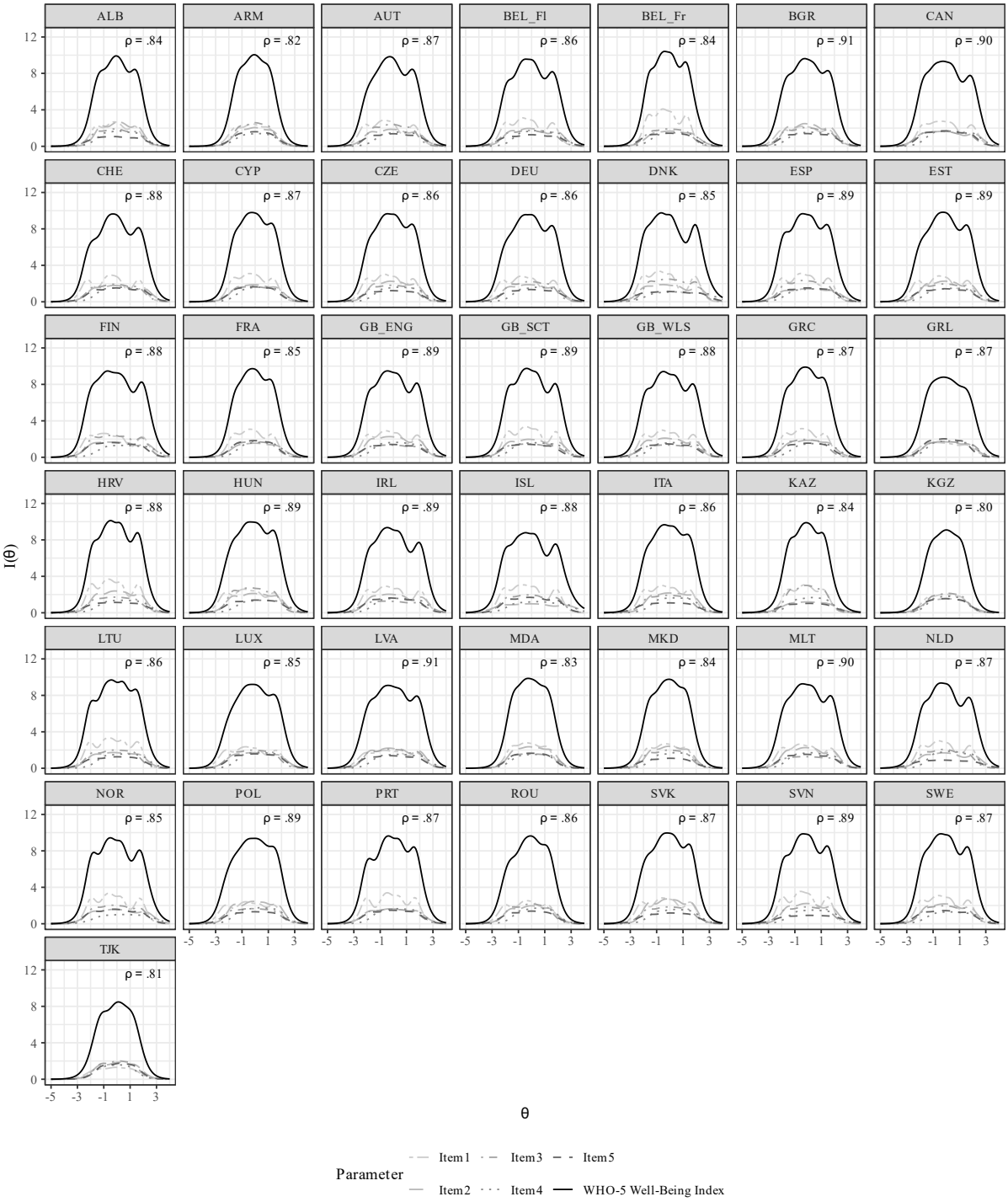
Notes. The black line represents the test characteristic curve in the respective country whereas the grey lines indicate the test characteristic curves in the remaining countries.

Figure 3. Differential test functioning for the GRM after alignment.



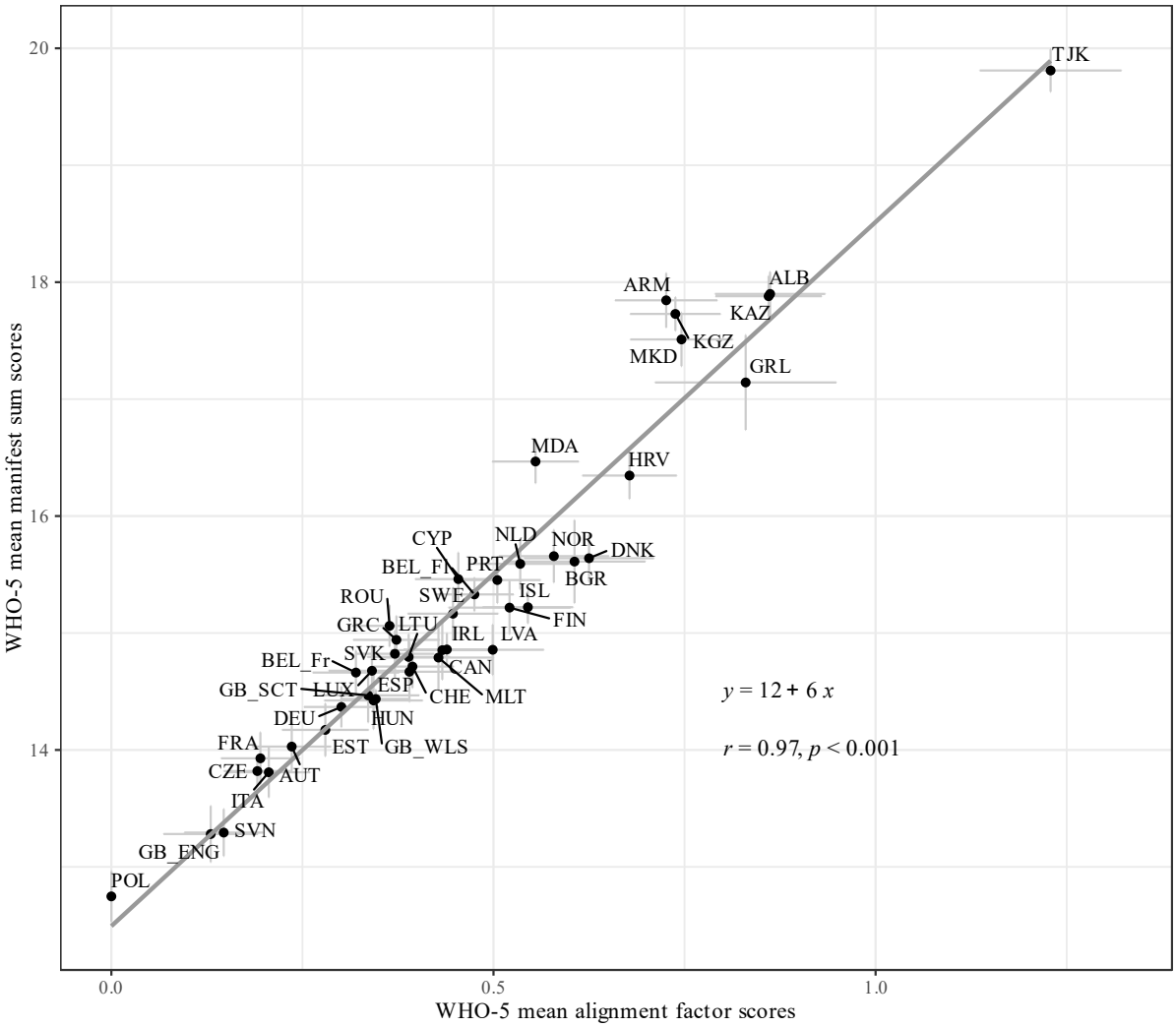
Notes. The curves show differences in expected test scores (with 99% CI) dependent on the level of the latent variable (GB_Eng [England] as reference group), *sDRF* = compensatory differential response functioning statistic with 99% CI, *uDRF* = non-compensatory differential response functioning statistic with 99% CI.

Figure 4. Item and test information functions for the GRM after alignment.



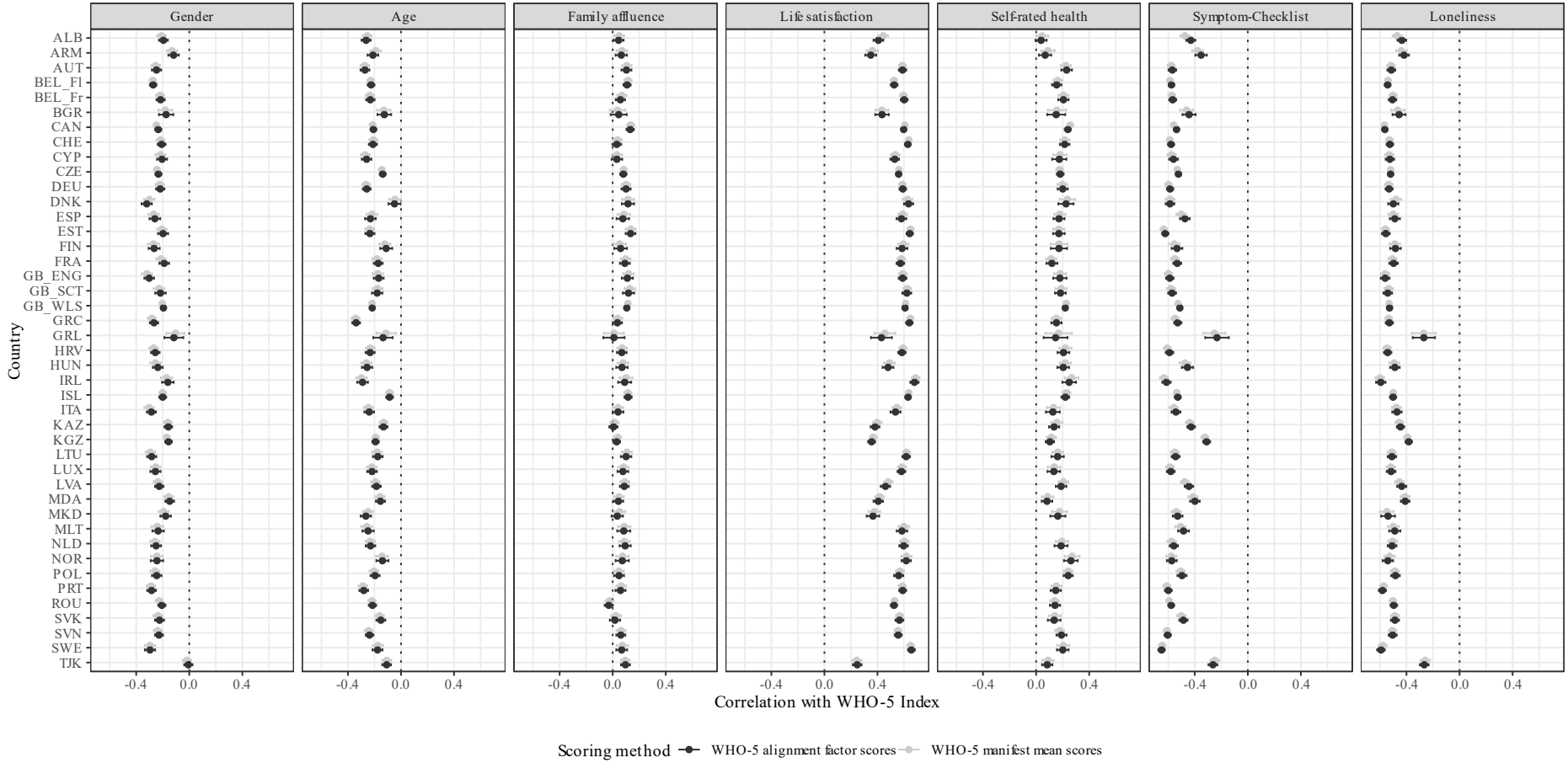
Notes. ρ represents the empirical marginal reliability. Items (1) *cheerful*, (2) *calm and relaxed*, (3) *active and vigorous*, (4) *fresh and rested*, (5) *interest*.

Figure 5. Scatterplot with country means of factor scores and manifest sum scores.



Notes. Factor scores were estimated via expected a-posterior (EAP) method. The regression equation and correlation coefficient are shown.

Figure 6. Correlations between the WHO–5 Index and the other variables.



Notes. Correlations with 99% bootstrapped confidence intervals. *Gender* was coded as 0 (= *boy*) and 1 (= *girl*). Self-rated health was coded as 0 (= *poor*) and 1 (= *fair/good/excellent*).